TOPOLOGY
A First Course

JAMES R. MUNKRES
Professor of Mathematics
Massachusetts Institute of Technology

PRENTICE-HALL, INC., Englewood Cliffs, New Jersey

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," he will be missing much of the motivation for the concepts introduced in the first part of the book. Things will go more smoothly if he 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 Chapter 8, 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) which 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 his 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 him to begin the study of topology.

How the book is organized. When this book is used for a one-semester course, some choices will have to be made concerning what material to cover. I have attempted to organize the book as flexibly as possible, so as to enable the instructor to follow his own preferences in this matter.

Part I of the book, consisting of the first four chapters, deals with that body of material which 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 topological spaces, connectedness, compactness (through compactness of finite products), and the countability and separation axioms (through the Urysohn metrization theorem). Certain sections are marked with an asterisk; these do not form part of the basic core and may be omitted or postponed with no loss of continuity.

Part II of the book consists of four chapters which are entirely independent of one another. They depend only on the material of Part I; the instructor may take them up in any order he chooses. Furthermore, if he wishes to cover only a portion of one of these later chapters, he can consult the introduction to that chapter, where there appears a diagram showing the relations of dependence among the sections of the chapter. The instructor who wishes, for instance, to conclude his course with a proof of the Jordan curve theorem can determine from this diagram which of the earlier sections of Chapter 8 are essential, and which peripheral, to his purpose.

Some of the material of the later chapters depends on one or more of the asterisked sections in Part I. Each such dependence is indicated in a footnote at the beginning of the asterisked section, and again in the introduction to the chapter in question. Some of the exercises also depend on earlier asterisked sections, but in such cases the dependence is obvious.

Possible course outlines. Most instructors who use this text for a one-semester course will wish to cover the "core" material of Part I, along with the Tychonoff theorem (§5-1). Many will cover additional topics as well. One might, for instance, treat some of the asterisked sections of Part I. (I

usually do local compactness, at least.) Or he may choose one or more topics from Part II. Possibilities include: the Stone-Čech compactification (§5-3), metrization theorems (Chapter 6), the Peano curve (§7-2), one or both versions of Ascoli's theorem (§7-3 and §7-6), dimension theory (§7-9), the fundamental group and applications (§8-1–§8-10), or the Jordan curve theorem (§8-13). I have in different semesters followed each of these options.

For the instructor who wishes to emphasize algebraic topology, one possible course outline would consist of Chapters 1 to 3 followed by Chapter 8 in its entirety. Omitting Chapter 4 will cause no difficulty, provided one skips Exercise 5 of §8-12, which involves the concept of normality.

Still another possible outline is the one suggested by the Committee on the Undergraduate Program in Mathematics (of the Mathematical Association of America) for a one-semester course in topology at the first-year graduate level. It would consist of Chapters 2, 3, and 4, followed by §5-1; §6-1, §6-3, §6-4; §7-1; §8-1 through §8-5, §8-8 through §8-11, and §8-14. This program assumes that the student has already had an introduction to set theory equivalent to our Chapter 1.

In a two-semester course, one can reasonably expect to cover the entire book.

Acknowledgements. 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 Robert Mosher and John Hemperly, and to my colleagues George Whitehead and Kenneth Hoffman. My appreciation goes to Miss Viola Wiley, who deciphered my handwriting and converted it into neat copy, and to the employees of Bertrick Associate Artists, Inc., who drew the illustrations.

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.

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.

1-1 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 $A, B, \ldots$ to denote sets, and lowercase letters $a, b, \ldots$ 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

$$A \subsetneq B.$$

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

§ 1-1

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 which 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 writ-

ing 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:

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

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:

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

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.

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:

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

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

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

is the statement

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

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.

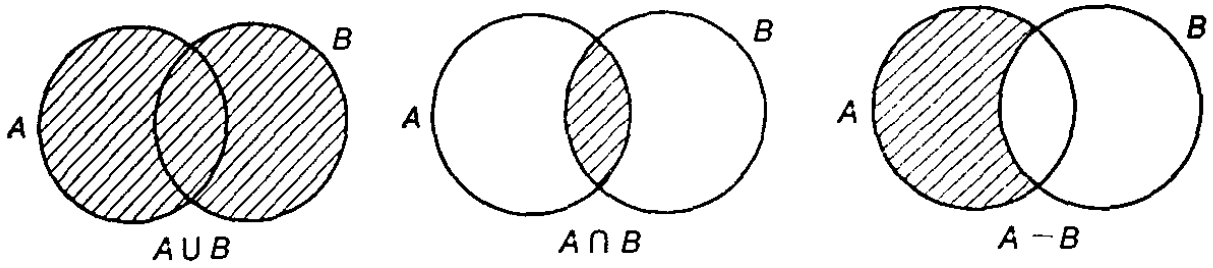


Figure 1

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 2 shows.

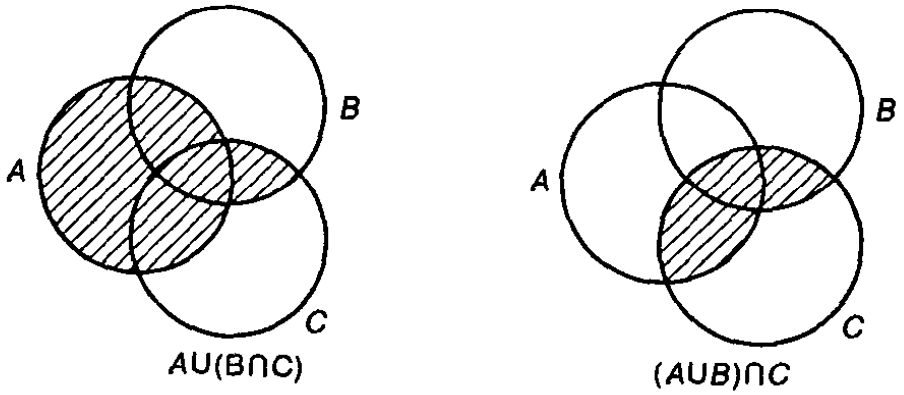


Figure 2

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

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

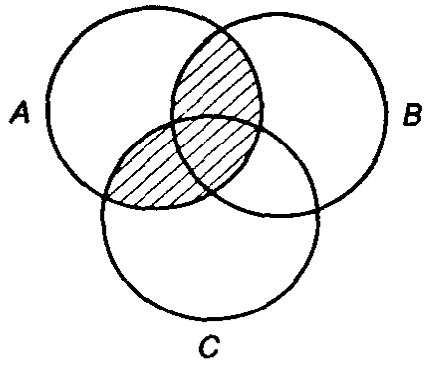


Figure 3

holds. The equation is illustrated in Figure 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 $\cup$.

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 lower-case 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

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

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 geome-

§ 1-1

try, 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 pos-

sible implications holds. If an equality fails, determine whether one or the other of the possible inclusions holds.

(a) $A \supset C$ and $B \supset C \Longleftrightarrow (A \cup B) \supset C$.
(b) $A \supset C$ or $B \supset C \Longleftrightarrow (A \cup B) \supset C$.
(c) $A \supset C$ and $B \supset C \Longleftrightarrow (A \cap B) \supset C$.
(d) $A \supset C$ or $B \supset C \Longleftrightarrow (A \cap B) \supset C$.
(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 $B \subset D \Longrightarrow (A \times B) \subset (C \times D)$.
(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) $x \in \bigcup_{A \in \mathcal{A}} A \Longrightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (b) $x \in \bigcup_{A \in \mathcal{A}} A \Longrightarrow x \in A$ for every $A \in \mathcal{A}$.
   (c) $x \in \bigcap_{A \in \mathcal{A}} A \Longrightarrow x \in A$ for at least one $A \in \mathcal{A}$.
   (d) $x \in \bigcap_{A \in \mathcal{A}} A \Longrightarrow x \in A$ 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 $-$.

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

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?

§1-2

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

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

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

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,

$$\text{domain } r = \{c | \text{there exists } d \in D \text{ such that } (c, d) \in r\},$$

$$\text{image } r = \{d | \text{there exists } c \in C \text{ such that } (c, d) \in r\}.$$

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 $f: A \longrightarrow B$ and if a is an element of A, we denote by $f(a)$ the unique element of B which 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 R denote the real numbers)

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

or one can equally well write

> "Let $f: R \longrightarrow R$ 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.

Definition. If $f: A \longrightarrow B$ 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 $f \mid A_0$, which is read "f restricted to A_0."

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

$$f: R \longrightarrow R \quad \text{defined by} \quad f(x) = x^2,$$
$$g: \bar{R}_+ \longrightarrow R \quad \text{defined by} \quad g(x) = x^2,$$

†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.

§ 1-2

$$h: R \longrightarrow \bar{R}_+ \quad \text{defined by} \quad h(x) = x^2,$$
$$k: \bar{R}_+ \longrightarrow \bar{R}_+ \quad \text{defined by} \quad k(x) = x^2.$$

The function g is different from the function f, because their rules are different subsets of $R \times R$; it is the restriction of f to the set $\bar{R}_+$. 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 4.

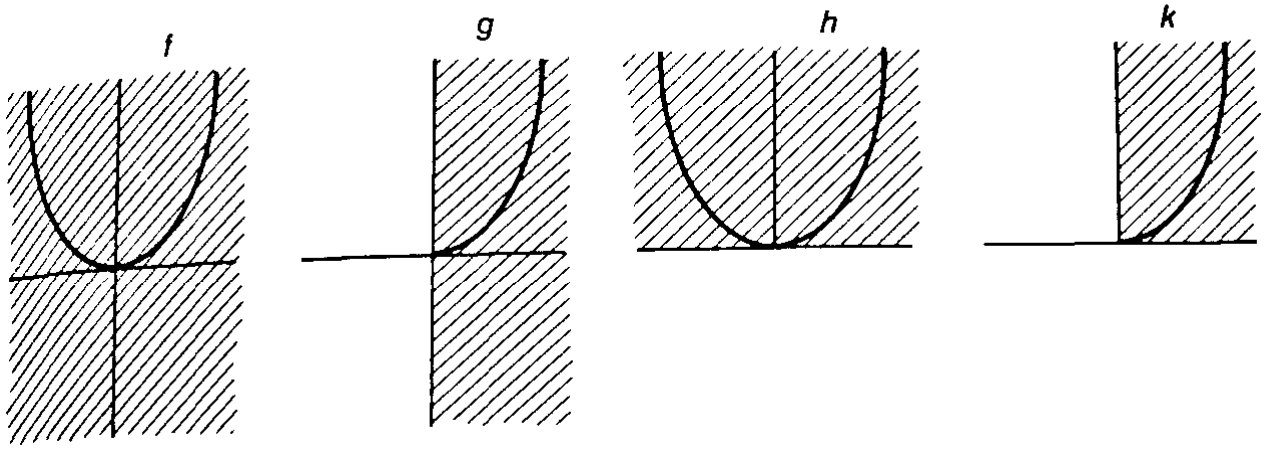


Figure 4

Let $f: A \longrightarrow B$. 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,

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

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.

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. But the operation f, when applied to subsets of A, does not preserve all these set operations. See Exercises 2 and 3.

As another example, it is not true in general that $f^{-1}(f(A_0)) = A_0$ and $f(f^{-1}(B_0)) = B_0$, as the following example shows. The relevant rules, which we leave to you to check, are the following: If $f: A \longrightarrow B$, then

$$f^{-1}(f(A_0)) \supset A_0 \quad \text{for } A_0 \subset A,$$
$$f(f^{-1}(B_0)) \subset B_0 \quad \text{for } B_0 \subset B.$$

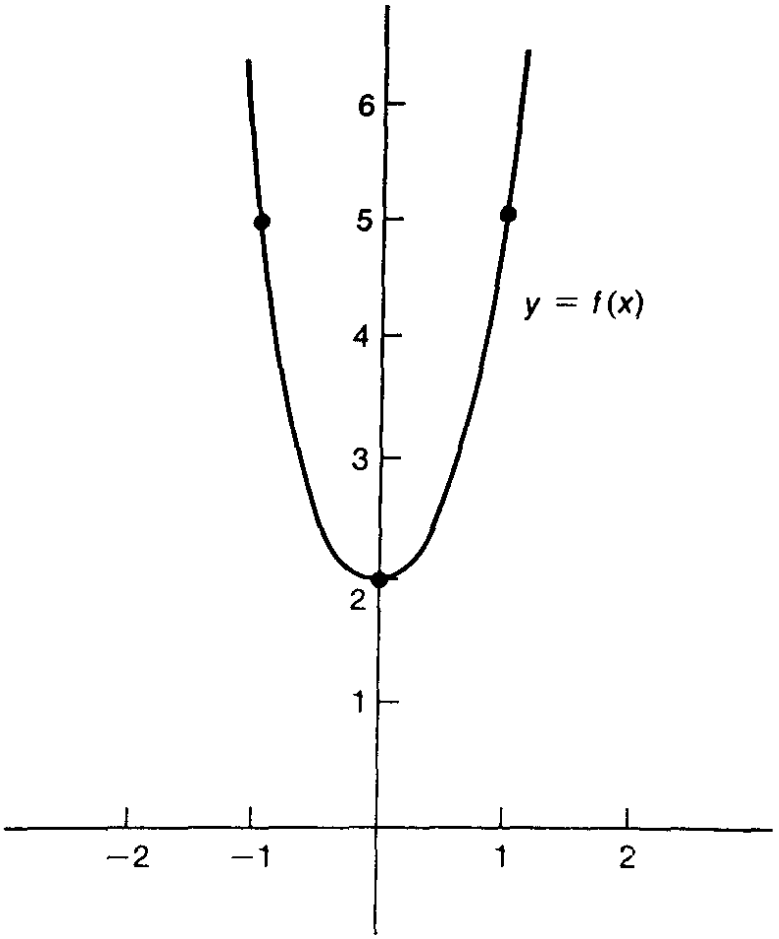


Figure 5

EXAMPLE 2. Consider the function $f: R \longrightarrow R$ given by $f(x) = 3x^2 + 2$ (Figure 5). Let $[a, b]$ denote the closed interval $a \leq x \leq b$. Then

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

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

Formally, $g \circ f: A \longrightarrow C$ 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 6.

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

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

$g \circ f: R \longrightarrow R$ 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 $f \circ g : R \longrightarrow R$ given by

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

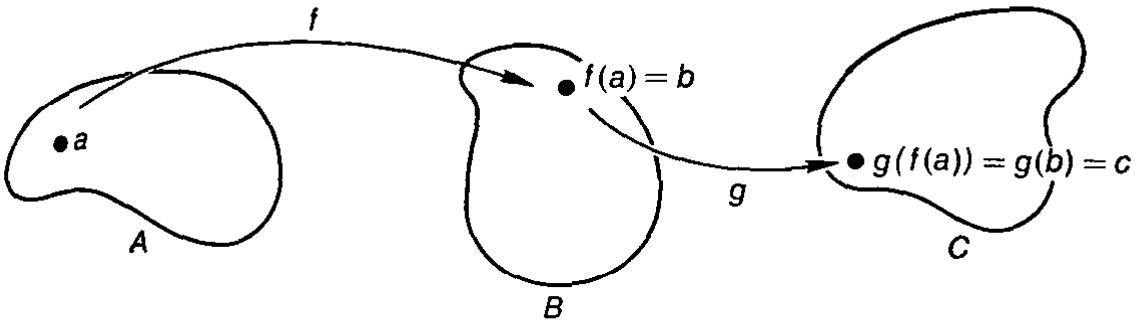


Figure 6

Definition. A function $f: A \longrightarrow B$ 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

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

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 4. Consider again the functions f, g, h, and k of Figure 4. The function $f: R \longrightarrow R$ 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 $h: R \longrightarrow \bar{R}_+$ obtained from f by changing the range is surjective but not injective. The function $k: \bar{R}_+ \longrightarrow \bar{R}_+$ 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* $g = h = f^{-1}$.

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 no ambiguity.

Exercises

1. Let $f: A \to B$. Let $A_0 \subset A$ and $B_0 \subset B$.
   (a) Show that $f^{-1}(f(A_0)) \supset A_0$ 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) $f(A_0 \cap A_1) \subset f(A_0) \cap f(A_1)$; give an example where equality fails.
   (h) $f(A_0 - A_1) \supset f(A_0) - f(A_1)$; give an example where equality fails.
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 $f: R \to R$ 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.)

1-3 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 *simple order relations*.

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 7.) We show that $E = E'$.

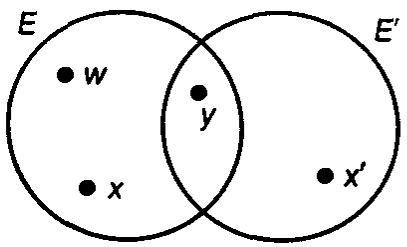


Figure 7

§1-3

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 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 $\mathfrak{L}$ be the collection of all straight lines in the plane parallel to the line $y = -x$. Then $\mathfrak{L}$ is a partition of the plane, since every point lies on some such line and each pair of distinct lines are disjoint. The partition $\mathfrak{L}$ comes from the equivalence relation on the plane that declares the points (x_1, y_1) and (x_2, y_2) to be equivalent if $x_1 + y_1 = x_2 + y_2$.

EXAMPLE 6. Let $\mathfrak{L}'$ be the collection of *all* straight lines in the plane. Then $\mathfrak{L}'$ is not a partition of the plane, for distinct elements of $\mathfrak{L}'$ 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 relation S satisfies only (3). The descendant relation D does somewhat better, for it satisfies both (2) and (3); but 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 §1-11).

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

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

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

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

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

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

is an order-preserving bijective correspondence, as you can check. It is pictured in Figure 8.

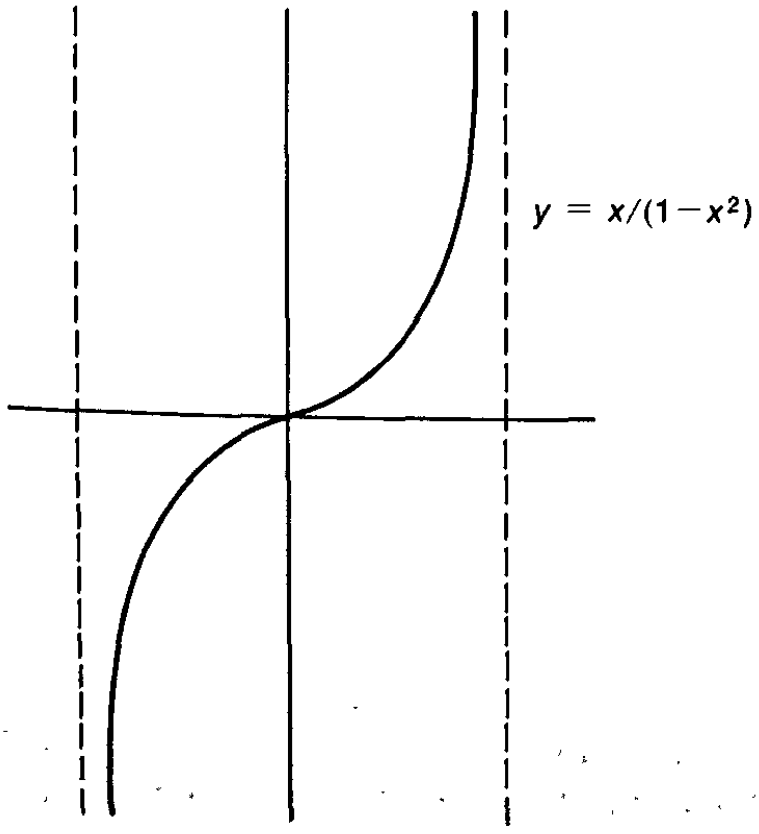


Figure 8

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

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

of 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:

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 $R \times R$. 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 Z_+ of positive integers, both in their usual orders; give $Z_+ \times [0, 1)$ 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 $[0, 1) \times Z_+$ 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 9.

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.

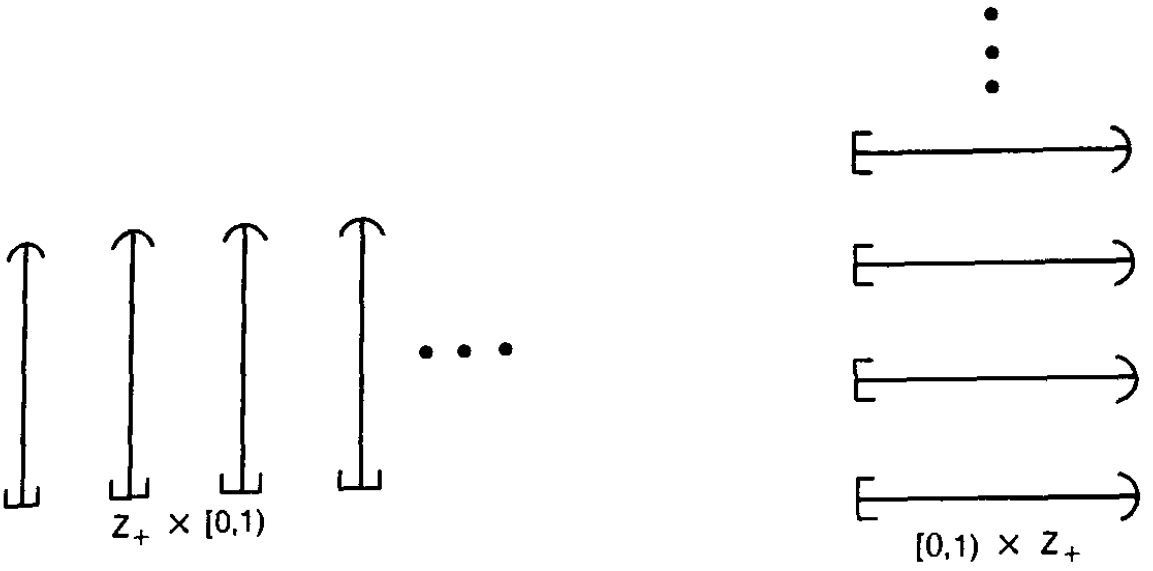


Figure 9

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** of A_0. It is denoted by lub 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** of A_0. It is denoted by glb 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 this set 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 $\{-1/2n \mid n \in Z_+\}$ 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 $\{-1/2n \mid n \in Z_+\}$, for instance, 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 $f: A \longrightarrow B$ 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 $f: (-1, 1) \longrightarrow R$ of Example 9 is order-preserving.
(b) Show that the equation $g(y) = 2y/[1 + (1 + 4y^2)^{1/2}]$ defines a function $g: R \longrightarrow (-1, 1)$ 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 Z_+ denote the set of positive integers. Consider the following order relations on $Z_+ \times Z_+$:
(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

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

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 $[0, 1) \times [0, 1]$?

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 §1-2. 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 $*$; but there are many others.

Assumption

We assume there exists a set R, called the set of **real numbers,** two binary operations $+$ and $\cdot$ on R, called the **addition** and **multiplication** operations, respectively, and an order relation $<$ on 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 R.

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

(3) There exists a unique element of R called **zero**, denoted by **0**, such that $x + 0 = x$ for all $x \in R$.

There exists a unique element of R called **one**, different from 0 and denoted by 1, such that $x \cdot 1 = x$ for all $x \in R$.

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

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

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 R_+ and the nonnegative reals (for reasons to be explained later) by $\bar{R}_+$.

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). Nevertheless, we in-

clude (8) in our list of basic properties of the real numbers to emphasize the fact that it and the least upper bound property are the two crucial properties of the order relation for R. From these two properties many of the topological properties of 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). First, we make the following definition: A subset A of the real numbers is said to be **inductive** if for every x in A, the number $x + 1$ is also in A.

Definition. Let $\mathcal{A}$ be the collection of all inductive subsets of R that contain 1. Then the set Z_+ of **positive integers** is defined by the equation

$$Z_+ = \bigcap_{A \in \mathcal{A}} A.$$

Note that the set R_+ of positive real numbers contains 1 and is inductive (if $x > 0$, then $x + 1 > 0$), so that R_+ belongs to $\mathcal{A}$. Therefore, $Z_+ \subset R_+$, so the elements of Z_+ are indeed positive, as the choice of terminology suggests. Indeed, one sees readily that 1 is the smallest element of Z_+, because the set of all real numbers x for which $x \geq 1$ is inductive and contains 1.

The basic properties of Z_+, which follow readily from the definition, are the following:

(1) $1 \in Z_+$.

(2) Z_+ is inductive.

(3) (Principle of induction). If Z_0 is an inductive set of positive integers that contains 1, then $Z_0 = Z_+$.

We define the set Z of **integers** to be the set consisting of the positive integers Z_+, the number 0, and the negatives of the elements of Z_+. One proves that the sum, difference, and product of two integers are integers, but the quotient is not necessarily an integer. The set 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 $n < a < n + 1$.

All these are familiar facts. One property of the positive integers that may not be quite so familiar, but will be very useful, is the following:

Theorem 4.1 (Well-ordering property). *Every nonempty subset of Z_+ has a smallest element.*

Proof. If $n \in Z_+$, we use $\{1, \ldots, n\}$ to denote the set $\{x \mid x \in Z_+ \text{ and } 1 \leq x \leq n\}$. Because there is no integer a between n and $n + 1$,

$$\{1, \ldots, n + 1\} = \{1, \ldots, n\} \cup \{n + 1\}.$$

We first prove "by induction" that, for each $n \in Z_+$, the following statement holds: *Every nonempty subset of $\{1, \ldots, n\}$ has a smallest element.*

Let Z_0 be the set of all positive integers n for which this statement holds.

§1-4

Then Z_0 contains 1, since if $n = 1$, the only nonempty subset of $\{1, \ldots, n\}$ is the set $\{1\}$ itself. Then, supposing Z_0 contains n, we show that it contains $n + 1$. So let C be a nonempty subset of the set $\{1, \ldots, n + 1\}$. If C consists of the single element $n + 1$, then that element is the smallest element of C. Otherwise, consider the set $C \cap \{1, \ldots, n\}$, which is nonempty. Because $n \in Z_0$, this set has a smallest element, which will automatically be the smallest element of C also. Thus Z_0 is inductive and contains 1, so we conclude that $Z_0 = Z_+$, whence the statement is true for all $n \in Z_+$.

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

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

The least upper bound axiom is also used to prove a number of other things about R. It is used for instance 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 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 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 $\mathfrak{A}$ is a collection of inductive sets, then the intersection of the elements of $\mathfrak{A}$ is an inductive set.
(b) Prove the basic properties (1), (2), (3) of Z_+.

4. (a) Show that $a \in Z_+ \Rightarrow a - 1 \in Z_+ \cup \{0\}$. [*Hint:* Let $X = \{x | x \in R$ and $x - 1 \in Z_+ \cup \{0\}\}$; show that X is inductive and contains 1.]
(b) Show that if $n \in Z$, there is no $a \in Z$ such that $n < a < n + 1$. [*Hint:* Prove it first for $n \in Z_+$.]

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

6. Prove the following properties of Z and Z_+:
(a) $a, b \in Z_+ \Rightarrow a + b \in Z_+$. [*Hint:* Show that given $a \in Z_+$, the set $X = \{x | x \in R$ and $a + x \in Z_+\}$ is inductive and contains 1.]
(b) $a, b \in Z_+ \Rightarrow a \cdot b \in Z_+$.
(c) $c, d \in Z \Rightarrow c + d \in Z$. [*Hint:* Prove it first for $d = 1$.]

§1-4

(d) $c, d \in Z \Rightarrow c - d \in Z$.
(e) $c, d \in Z \Rightarrow c \cdot d \in Z$.
(f) $x > 1 \Rightarrow 1/x \notin Z$.

7. Let $a \in R$. Define inductively

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

for $n \in Z_+$. (See §1-7 for a discussion of the process of inductive definition.) Show that for $n, m \in Z_+$ and $a, b \in R$,

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

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

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

9. Assume the least upper bound axiom.
(a) Show that $\text{glb}\{1/n \mid n \in Z_+\} = 0$.
(b) Show that given a with $0 < a < 1$, $\text{glb}\{a^n \mid n \in Z_+\} = 0$. [*Hint:* Let $h = (1 - a)/a$, and show that $a^n \leq 1/(1 + nh)$.]

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

11. Assuming the least upper bound axiom, 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 $b = \text{lub } B$; show that $b^2 = a$.
(d) Show that if b and c are positive and $b^2 = c^2$, then $b = c$.

12. Given $m \in Z$, we say that m is **even** if $m/2 \in Z$, and m is **odd** otherwise.
(a) Show that if m is odd, $m = 2n + 1$ for some $n \in Z$. [*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 $n \in Z_+$.

(c) Show that if $a > 0$ is rational, then $a = m/n$ for some $m, n \in Z_+$ where not both m and n are even. [*Hint:* Let n be the smallest element of the set $\{x \mid x \in Z_+ \text{ and } x \cdot a \in Z_+\}$.]

(d) *Theorem.* $\sqrt{2}$ *is irrational.*

1-5 Arbitrary Cartesian Products

We have already defined what we mean by the cartesian product $A \times B$ of two sets. Now we introduce cartesian products of arbitrarily many sets. We consider first some special cases, and then we give the general definition.

As usual, let $\{1, \ldots, m\}$ denote the set of all positive integers a such that $1 \leq a \leq m$. Given a set X, we define an m-tuple of elements of X to be a function

$$\mathbf{x} : \{1, \ldots, m\} \longrightarrow X.$$

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 we often denote the function $\mathbf{x}$ itself by the symbol

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

We let X^m denote the set of all m-tuples of elements of X.

Now we can define the cartesian product.

Definition. Suppose that $\{A_1, \ldots, A_m\}$ is a collection of sets, indexed with the positive integers from 1 to m. Let $X = A_1 \cup \cdots \cup A_m$. The **cartesian product** of this indexed collection of sets, denoted by

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

is defined to be the set of all m-tuples $(x_1, \ldots, x_m)$ 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 $\mathbf{x} : \{1, 2\} \longrightarrow A \cup B$ 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:

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

To generalize the preceding definition, we proceed by analogy. Given a set X, we define an **ω-tuple of elements of X** to be a function

$$\mathbf{x} : Z_+ \longrightarrow X;$$

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 we denote $\mathbf{x}$ itself by the symbol

$$(x_1, x_2, \ldots) \qquad \text{or} \qquad (x_n)_{n \in Z_+}.$$

We let X^ω denote the set of all ω-tuples of elements of X.

Definition. Suppose that $\{A_1, \ldots, A_n, \ldots\}$ is a collection of sets, indexed with the positive integers. Let X be the union of the sets in this collection. The **cartesian product** of this indexed collection of sets, denoted by

$$\prod_{i \in Z_+} A_i \qquad \text{or} \qquad A_1 \times A_2 \times \cdots,$$

is defined to be the set of all ω-tuples $(x_1, x_2, \ldots)$ 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 X^m of *all* m-tuples of elements of X; and the product $A_1 \times A_2 \times \cdots$ is just the set X^ω of all ω-tuples of elements of X.

EXAMPLE 3. If R is the set of real numbers, then 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, R^ω is sometimes called "infinite-dimensional euclidean space"; it is the set of all ω-tuples $(x_1, x_2, \ldots)$ of real numbers, that is, the set of all functions $\mathbf{x} : Z_+ \longrightarrow R$.

Now we turn to the general definition of cartesian product, which will include these as special cases. First we must make more precise what we mean by an "indexed collection of sets," a concept tacitly assumed above.

Definition. Let $\mathcal{A}$ be a 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 the element α of 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 $\mathcal{A}$.

The main use of indexing functions comes when we define arbitrary cartesian products:

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 $\mathbf{x}(\alpha)$. 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 $\alpha \in J$.

Said differently, this cartesian product is just 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. In dealing with the set X^J, we sometimes use "tuple notation" for its elements, and sometimes we use functional notation, depending on which is more convenient.

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 a 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) Let J be an index set; write $J = K \cup L$, where K and L are disjoint and nonempty. Show there is a bijective correspondence of
$$\prod_{\alpha \in J} A_\alpha \quad \text{with} \quad \left(\prod_{\alpha \in K} A_\alpha\right) \times \left(\prod_{\alpha \in L} A_\alpha\right).$$
3. Let J be a (nonempty) index set. Consider two indexed families $\{A_\alpha\}_{\alpha \in J}$ and $\{B_\alpha\}_{\alpha \in J}$. Prove the following:
(a) If $A'_\alpha \subset A_\alpha$ for all $\alpha \in J$, then $\prod_{\alpha \in J} A'_\alpha \subset \prod_{\alpha \in J} A_\alpha$.†
(b) The converse of (a) holds if $\prod A'_\alpha$ is nonempty.
(c) $\left(\prod_{\alpha \in J} A_\alpha\right) \cap \left(\prod_{\alpha \in J} B_\alpha\right) = \prod_{\alpha \in J} (A_\alpha \cap B_\alpha)$.
(d) $\left(\prod_{\alpha \in J} A_\alpha\right) \cup \left(\prod_{\alpha \in J} B_\alpha\right) \subset \prod_{\alpha \in J} (A_\alpha \cup B_\alpha)$.
(e) If at least one A_α is empty, then $\prod_{\alpha \in J} A_\alpha$ is empty. What can you say if each A_α is nonempty? (See Exercise 4 of §1-9.)
4. Let $m, n \in Z_+$. Let $X \neq \varnothing$.
(a) If $m \leq n$, find an injective map $f: X^m \longrightarrow X^n$.
(b) Find a bijective map $g: X^m \times X^n \longrightarrow X^{m+n}$.
(c) Find an injective map $h: X^n \longrightarrow X^\omega$.
(d) Find a bijective map $k: X^n \times X^\omega \longrightarrow X^\omega$.
(e) Find a bijective map $l: X^\omega \times X^\omega \longrightarrow X^\omega$.
(f) If $A \subset B$, find an injective map $m: X^A \longrightarrow X^B$.
5. Which of the following subsets of R^ω can be expressed as the cartesian product of subsets of R?
(a) $\{\mathbf{x} | x_i \text{ is an integer for all } i\}$.
(b) $\{\mathbf{x} | x_i \geq i \text{ for all } i\}$.
(c) $\{\mathbf{x} | x_i \text{ is an integer for all } i \geq 100\}$.
(d) $\{\mathbf{x} | x_2 = x_3\}$.

1-6 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.

†Strictly speaking, if we are given a function $\mathbf{x}$ mapping J into $\bigcup A'_\alpha$, we must change its range before it can be considered as a function mapping J into $\bigcup A_\alpha$. We shall ignore this technicality when dealing with cartesian products.

Recall that if n is a positive integer, we use $\{1, \ldots, n\}$ to denote the set

$$\{x \mid x \in Z_+ \text{ and } 1 \leq x \leq n\};$$

it is called a **section** of the positive integers. The sets $\{1, \ldots, n\}$ are the prototypes for what we call the finite sets.

Definition. A set A is said to be **finite** if it is empty or if there is a bijective correspondence

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

for some $n \in Z_+$. In the former case, we say that A **has 0 elements**; in the latter case, we say that A **has n elements.**

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

Now note carefully: *We have not yet shown that, given the finite set A, the number of elements "which A has" is uniquely determined by A.* As far as we know, there might exist bijective correspondences of a given 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* $\{1, \ldots, n\}$.

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

$$f(x) = g(x) \quad \text{for } x \in A - \{a_0\},$$
$$f(a_0) = n + 1.$$

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 $f \mid A - \{a_0\}$ 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

$$h(a_0) = n + 1,$$
$$h(a_1) = m,$$
$$h(x) = f(x) \quad \text{for } x \in A - \{a_0\} - \{a_1\}.$$

(See Figure 10.) It is easy to check that h is a bijection.

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

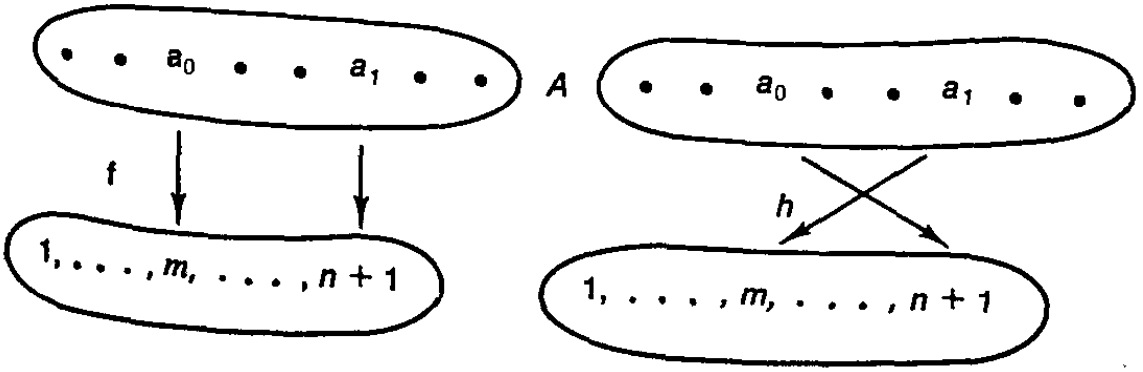


Figure 10

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* $n \in Z_+$. *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 Z_0 be the subset of Z_+ consisting of those integers n for which the theorem holds. We shall show that Z_0 contains the number 1 and is inductive. From this we conclude that $Z_0 = Z_+$, 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 $f: A \to \{1, \ldots, n + 1\}$ is a bijection, and B is a nonempty proper subset of A. Choose a point a_0 in B and a point a_1 in $A - B$. We apply the preceding lemma to conclude there is a bijection

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

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 $h: B - \{a_0\} \to \{1, \ldots, n\}$.

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

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

The preceding lemma, combined with (1), implies that there is no bijection of B with $\{1, \ldots, n + 1\}$. This is the first half of what we wanted to prove. 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 $\{1, \ldots, p + 1\}$. In either case, there is a bijection of B with $\{1, \ldots, m\}$ for some $m < n + 1$, as desired.

The induction principle now shows that the theorem is true for all $n \in Z_+$. □

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 $g: A \to \{1, \ldots, n\}$ for some n. The composite $g \circ f^{-1}$ is then a bijection of B with $\{1, \ldots, n\}$. This contradicts the preceding theorem. □

§ 1-6

Corollary 6.4. The number of elements in a finite set A *is uniquely determined by* A.

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

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

Then the composite

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

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

Corollary 6.5. If B *is a subset of the finite set* A, *then* B *is finite. If* B *is a proper subset of* A, *then the number of elements in* B *is less than the number of elements in* A.

Corollary 6.6. Z_+ *is not finite.*

Proof. The function $f: Z_+ \longrightarrow Z_+ - \{1\}$ defined by $f(n) = n + 1$ is a bijection of Z_+ with a proper subset of itself. □

Theorem 6.7. Let B *be a nonempty set; let* n *be a positive integer. Then the following are equivalent:*

(1) *There is a surjective function* $f: \{1, \ldots, n\} \longrightarrow B$.

(2) *There is an injective function* $g: B \longrightarrow \{1, \ldots, n\}$.

(3) B *is finite and has at most* n *elements.*

Proof. We prove that (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1); this will suffice to show the statements are equivalent. Let $A = \{1, \ldots, n\}$.

(1) $\Rightarrow$ (2). Given a surjective function $f: A \longrightarrow B$, define $g: B \longrightarrow A$ by the equation

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

Because f is surjective, the set $f^{-1}(\{b\})$ is nonempty. We know by the well-ordering property that every nonempty subset of Z_+ has a unique smallest element, so g is well 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.

(2) $\Rightarrow$ (3). Let $g: B \longrightarrow A$ be injective. Let C be the image set $g(B)$; then the function $g': B \longrightarrow C$ obtained from g by changing its range is bijective. Because C is a subset of $A = \{1, \ldots, n\}$, there is a bijection $h: C \longrightarrow \{1, \ldots, p\}$ for some $p \leq n$. The composite $h \circ g'$ is a bijection of B with $\{1, \ldots, p\}$.

(3) $\Rightarrow$ (1). Let B have m elements, where $1 \leq m \leq n$. Then there is a bijection

$$h : \{1, \ldots, m\} \longrightarrow B.$$

If $m = n$, then h is a surjection of A onto B, as desired. If $m < n$, we extend h to the desired surjection $f : \{1, \ldots, n\} \longrightarrow B$ by defining $f(i) = h(1)$ for $m < i \leq n$. □

Corollary 6.8. Let $\{A_\alpha\}_{\alpha \in J}$ be an indexed family of sets. If each set A_α is finite and if the index set J is finite, then the sets

$$\bigcup_{\alpha \in J} A_\alpha \quad \text{and} \quad \prod_{\alpha \in J} A_\alpha$$

are finite.

To put it succinctly, this theorem states that *finite unions and finite products of finite sets are finite*. The proof is left to the exercises.

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 having three elements does not have four elements.)

(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 number of elements in A.]
   (b) Show that A has the order type of a section of the positive integers.

5. (a) Show that $A \cup B$ is finite if and only if A and B are finite.
   (b) Let $J \neq \emptyset$. Show that if J is finite and each set A_α for $\alpha \in J$ is finite, then $\bigcup_{\alpha \in J} A_\alpha$ is finite.
   (c) To what extent does the converse of (b) hold?

6. (a) Show that $A \times B$ is finite if A and B are finite.
   (b) Let $J \neq \emptyset$. Show that if J is finite and each set A_α is finite, then $\prod_{\alpha \in J} A_\alpha$ is finite.
   (c) To what extent does the converse of (b) hold?

§1-7

7. Show that if A is finite, then the set $\mathcal{P}(A)$ of all subsets of A is finite. [*Hint:* Let X be the two element set $\{0, 1\}$. Define a bijection of $\mathcal{P}(A)$ with the cartesian product X^A.]

8. Let A and B be sets; let $\mathcal{F}$ be the set of all functions $f: A \longrightarrow B$. Show that if A and B are finite, then $\mathcal{F}$ is finite.

1-7 Countable and Uncountable Sets

Just as the sets $\{1, \ldots, n\}$ are the prototypes for the finite sets, the set Z_+ 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

$$f: Z_+ \longrightarrow A.$$

EXAMPLE 1. The set Z of all integers is countably infinite. One checks easily that the function $f: Z \longrightarrow Z_+$ 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 $Z_+ \times Z_+$ is countably infinite. If we represent the elements of the product $Z_+ \times Z_+$ by the integer points in the first quadrant, then the left-hand portion of Figure 11 suggests how to "count" the points, that is, how to put them in bijective correspondence with the positive integers.

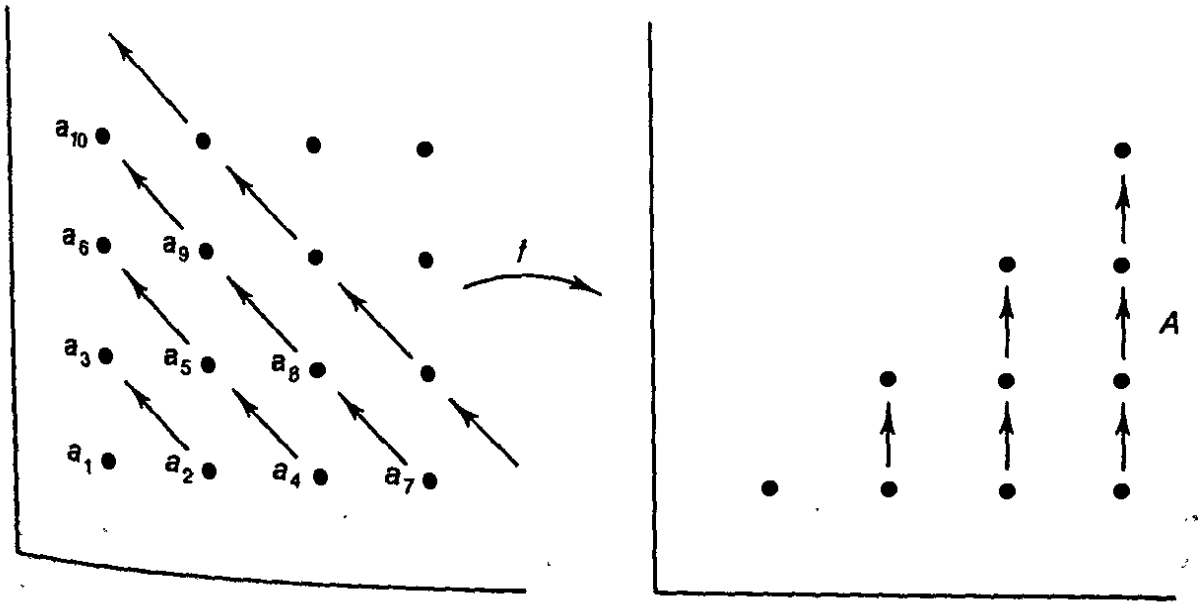


Figure 11

A picture is not a proof, of course, but this picture suggests a proof. First, we construct a bijection $f: Z_+ \times Z_+ \to A$, where A is the subset of $Z_+ \times Z_+$ 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 $g: A \to Z_+$ by the formula

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

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

Another proof that $Z_+ \times Z_+$ is countably infinite will be given later.

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 theorem concerning countable sets that is similar to Theorem 6.7 about finite sets. It is the following:

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

(1) *There is a surjective function* $f: Z_+ \to B$.

(2) *There is an injective function* $g: B \to Z_+$.

(3) B *is countable.*

Proof. We prove the implications (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (1).

(1) $\Rightarrow$ (2). Let $f: Z_+ \to B$ be a surjection. Define $g: B \to Z_+$ 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.

(2) $\Rightarrow$ (3). Let $g: B \to Z_+$ 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 Z_+. Thus to prove our result, it suffices to show that every subset of Z_+ is countable. So let C be a subset of Z_+.

If C is finite, it is countable by definition. So suppose that C is infinite. We have to prove C is countably infinite; that is, we have to construct a bijection $h: Z_+ \to C$.

We define h "by induction." Define $h(1)$ to be the smallest element of C; it exists because every nonempty subset C of Z_+ has a smallest element. Then assuming that $h(1), \ldots, h(n-1)$ are defined, define

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

The set $C - h(\{1, \ldots, n-1\})$ is not empty; for if it were empty, then $h: \{1, \ldots, n-1\} \to C$ would be surjective, whence C would be finite (by Theorem 6.7). Thus $h(n)$ is well defined. By induction, we have defined $h(n)$ for all $n \in Z_+$.

§1-7

To show that h is injective is easy. Given $m < n$, note that $h(m)$ belongs to the set $h(\{1, \ldots, n-1\})$, 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 $h(Z_+)$ cannot be contained in the finite set $\{1, \ldots, c\}$, because $h(Z_+)$ is infinite (since h is injective). Therefore, there is an n in Z_+ such that $h(n) > c$. Let m be the *smallest* element of 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 $h(\{1, \ldots, m-1\})$. Since $h(m)$ is defined as the smallest element of the set $C - h(\{1, \ldots, m-1\})$, we must have $h(m) \leq c$. Putting the two inequalities together, we have $h(m) = c$, as desired.

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

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

This completes the proof of the theorem. □

Corollary 7.2. A subset of a countable set is countable.

Corollary 7.3. The set $Z_+ \times Z_+$ is countably infinite.

Proof. In view of the preceding theorem, it suffices to construct an injective map $f: Z_+ \times Z_+ \to Z_+$. 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 Q_+ of positive rational numbers is countably infinite. For we can define a surjection $g: Z_+ \times Z_+ \to Q_+$ by the equation

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

Because $Z_+ \times Z_+$ is countable, there is a surjection $f: Z_+ \to Z_+ \times Z_+$. Then the composite $g \circ f: Z_+ \to Q_+$ is a surjection, so that Q_+ is countable. And, of course, Q_+ is infinite because it contains Z_+.

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

There is a point in the proof of Theorem 7.1 where we stretched the principles of logic a bit. It occurred at the point, in proving the implication (2) $\Rightarrow$ (3), 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 §1-4, when we defined a^n.

But there is a problem here. After all, the induction principle states only that if Z_0 is an inductive set of positive integers and Z_0 contains 1, then $Z_0 = Z_+$. To use the principle to prove a theorem "by induction," one begins the proof with the statement "Let Z_0 be the set of all positive integers n for which the theorem is true," and then one goes ahead to prove that Z_0 contains 1 and that Z_0 is inductive, whence Z_0 must be all of 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 Z_0 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 Z_+, there is a unique function $h : Z_+ \to C$ satisfying the formula:

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

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

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

for example, is self-contradictory; although $h(i)$ necessarily belongs to the set $h(\{1, \ldots, i+1\})$, 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"; and 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* $h : Z_+ \to A$.

This is the principle we actually used in the proof of Theorem 7.1. 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 Theorem 7.1 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.

Now we give some further rules for determining whether or not a set is countable.

Theorem 7.4. *A countable union of countable sets is countable.*

Proof. Let $\{A_\alpha\}_{\alpha \in J}$ be an indexed family of countable sets, where the index set J is countable. The case $J = \varnothing$ is trivial; so let us assume that $J \neq \varnothing$. Assume also that each set A_α is nonempty, for convenience; this assumption does not change anything.

Because each A_α is countable, we can choose, for each α, a surjective function $f_\alpha : Z_+ \to A_\alpha$. Similarly, we can choose a surjective function $g : Z_+ \to J$. Now define

$$h : Z_+ \times Z_+ \longrightarrow \bigcup_{\alpha \in J} A_\alpha$$

by the equation

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

It is easy to check that h is surjective. Since $Z_+ \times Z_+$ is in bijective correspondence with Z_+, the countability of the union follows from Theorem 7.1. □

Theorem 7.5. *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 $f : Z_+ \to A$ and $g : Z_+ \to B$. Then the function $h : Z_+ \times Z_+ \to A \times B$ 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

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

Since 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 7.6. *Let* X *denote the two element set* $\{0, 1\}$. *Then the set* X^ω *is uncountable.*

Proof. We show that, given any function

$$g: Z_+ \longrightarrow X^\omega,$$

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 a point $\mathbf{y} = (y_1, y_2, \ldots, y_n, \ldots)$ 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 a point 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 following:

Theorem 7.7. *The set* $\mathcal{P}(Z_+)$ *of all subsets of* Z_+ *is uncountable.*

Proof. One proof consists of showing that $\mathcal{P}(Z_+)$ is in bijective correspondence with the set $\{0, 1\}^\omega$. This we leave to the exercises.

Another proof is more direct. We shall prove that if A is an arbitrary set,

there is no surjective function $g : A \longrightarrow \mathcal{P}(A)$. Then, in particular, the set $\mathcal{P}(Z_+)$ is not countable.

So, let $g : A \longrightarrow \mathcal{P}(A)$ be a function. 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 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 7.6. 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 R does not, in fact, depend on the infinite decimal expansion of R or indeed on *any* of the algebraic properties of R; it depends only on the order properties of R. We shall demonstrate the uncountability of R, using only its order properties, in a later section (§3-6).

Exercises

1. Show that Q is countably infinite.
2. Show that the maps f and g of Examples 1 and 2 are bijections.
3. Let A be a set; let X be the two-element set $\{0, 1\}$. Show that there is a bijective correspondence between the set $\mathcal{P}(A)$ of all subsets of A and the cartesian product X^A.
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 $f: \{0, 1\} \longrightarrow Z_+$.
   (b) The set B_n of all functions $f: \{1, \ldots, n\} \longrightarrow Z_+$.
   (c) The set $C = \bigcup_{n \in Z_+} B_n$.
   (d) The set D of all functions $f: Z_+ \longrightarrow Z_+$.
   (e) The set E of all functions $f: Z_+ \longrightarrow \{0, 1\}$.
   (f) The set F of all functions $f: Z_+ \longrightarrow \{0, 1\}$ 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 $f: Z_+ \longrightarrow Z_+$ that are eventually 1.
   (h) The set H of all functions $f: Z_+ \longrightarrow Z_+$ that are eventually constant.
   (i) The set I of all two-element subsets of Z_+.
   (j) The set J of all finite subsets of 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 $h : A \longrightarrow B$ 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 $f: A \longrightarrow C$ and $g: C \longrightarrow A$, 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 all *countable* subsets of X^ω. Show that X^ω and $\mathcal{B}$ have the same cardinality.

9. (a) The recursion formula

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

   is not one to which the principle of recursive definition applies. Show that nevertheless there does exist a function $h: Z_+ \longrightarrow R$ 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 $h : Z_+ \to R$ satisfying the recursion formula

$$h(1) = 1,$$
$$h(2) = 2,$$
$$h(n) = [h(n+1)]^2 + [h(n-1)]^2 \quad \text{for } n \geq 2.$$

*1-8 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, the case that was used in the proof of Theorem 7.1. That should make the underlying idea of the proof much clearer when we consider the general case.

So, given the infinite subset C of Z_+, let us consider the following recursion formula for a function $h : Z_+ \to C$:

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

We shall prove that there exists a unique function $h : Z_+ \to C$ satisfying this recursion formula.

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

Lemma 8.1. Given $n \in Z_+$, *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, and therefore it is capable of being proved by induction. Let Z_0 be the set of all n for which the lemma holds. We show that Z_0 contains 1 and is inductive. It then follows that $Z_0 = Z_+$.

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

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

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 (*). Hence $i > 1$, and for all $j < i$, we have $f(j) = g(j)$. Because f and g satisfy (*),

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

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

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 $h: Z_+ \to C$ *satisfying* (*) *for all* $i \in Z_+$.

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 $h: Z_+ \to C$ 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 $Z_+ \times C$. We must show that U is the rule for a function $h: Z_+ \to C$.

That is, we must show that each element i of 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 $n \geq i$. 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:

$$h(i) = f_n(i) \quad \text{for } i \leq n,$$

f_n satisfies (*) for all i 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 section of the positive integers into* A, *an element of* A. *Then there exists a unique function*

$$h : Z_+ \longrightarrow A$$

such that

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

The formula (*) is called a **recursion formula** for h. It determines $h(1)$, 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 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 *section* of Z_+ 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 $h : Z_+ \longrightarrow C$ such that $h(1) = a_0$, and for $i > 1$,

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

as desired.

EXAMPLE 2. Given $a \in R$, we "defined" a^n, in the exercises of §1-4, by the recursion formula

$$a^1 = a,$$

$$a^n = a^{n-1} \cdot a.$$

We wish to apply Theorem 8.4 to define a function $h : Z_+ \longrightarrow R$ rigorously such that $h(n) = a^n$. To apply this theorem, let a_0 denote the element a of R, and define ρ by the equation $\rho(f) = f(m) \cdot a$, where f maps a section of Z_+

into R and m is the largest element of the domain of f. Then there exists a unique function $h : Z_+ \longrightarrow R$ such that

$$h(1) = a_0,$$
$$h(i) = \rho(h|\{1, \ldots, i-1\}) \quad \text{for } i > 1.$$

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

$$a^1 = a,$$
$$a^i = a^{i-1} \cdot a,$$

as desired.

Exercises

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

$$\sum_{k=1}^{n} b_k = b_1 \qquad \text{for } n = 1,$$
$$\sum_{k=1}^{n} b_k = \left(\sum_{k=1}^{n-1} b_k\right) + b_n \quad \text{for } n > 1.$$

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

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

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

3. Obtain the definitions of a^n and $n!$ for $n \in Z_+$ as special cases of Exercise 2.

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

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

Define them rigorously by use of Theorem 8.4.

5. Show that there is a unique function $h : Z_+ \longrightarrow R_+$ satisfying the formula

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

6. (a) Show that there is no function $h : Z_+ \longrightarrow R_+$ satisfying the formula

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

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

§1-9

(b) Consider the recursion formula

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

Show that there exists a unique function $h : Z_+ \longrightarrow R_+$ satisfying this formula.

7. Prove Theorem 8.4.

1-9 *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* $f : Z_+ \longrightarrow A$.

(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 $f : Z_+ \longrightarrow A$. Let the image set $f(Z_+)$ 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

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

The map g is indicated schematically in Figure 12; one checks easily that it is a bijection.

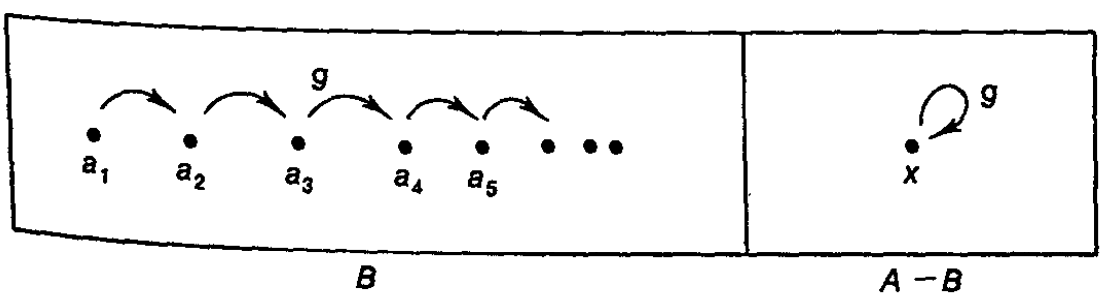


Figure 12

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 $f: Z_+ \to A$.

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

It is easy to see that f is injective. For suppose that $m < n$. Then $f(m)$ belongs to the set $f(\{1, \ldots, n-1\})$, 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 $f: Z_+ \to A$ recursively by the formula

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

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

In this respect this formula differs notably from the recursion formula we considered in proving Theorem 7.1. There we had an infinite subset C of Z_+, and we defined h by the formula

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

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

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 $f: Z_+ \to A$ 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 $Z_+ \times A$. Therefore, to prove the existence of the function f we must construct the appropriate subset of $Z_+ \times A$, 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 having exactly one element in common with each element of $\mathcal{A}$; that is, such that for each $A \in \mathcal{A}$, the set $C \cap A$ contains a single element.*

This axiom asserts the existence of a set that can be thought of as having been obtained by choosing one element from each of the sets A 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 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, because 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 having exactly one element in common with 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 $f: Z_+ \longrightarrow A$. 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 $f: Z_+ \longrightarrow A$ by the recursion formula

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

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

§ 1-9

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 having exactly one element in common with 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.

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 really 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 $f: Z_+ \longrightarrow X^\omega$, 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 Z_+.
   (b) The collection $\mathcal{B}$ of nonempty subsets of Z.
   (c) The collection $\mathcal{C}$ of nonempty subsets of the rational numbers Q.
   (d) The collection $\mathcal{D}$ of nonempty subsets of X^ω, where $X = \{0, 1\}$.
3. Suppose that A is a set and $\{f_n\}_{n \in Z_+}$ 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 $f: Z_+ \to A$ without using the choice axiom?

4. Show that the choice axiom is equivalent to the statement that for any indexed family $\{A_\alpha\}_{\alpha \in J}$ of nonempty sets, with $J \neq \emptyset$, the cartesian product

$$\prod_{\alpha \in J} A_\alpha.$$

is not empty.

5. There was a theorem in §1-7 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.)

6. (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?
   (c) Show that if A is any set, there is no injective map $f: \mathcal{P}(A) \to A$, where $\mathcal{P}(A)$ is the set of all subsets of A.

7. 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?

8. 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 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 $n \in Z_+$, 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.

*9. Show that $\mathcal{P}(Z_+)$ and R have the same cardinality.

A famous conjecture of set theory, called the *continuum hypothesis*, asserts that there exists no set having greater cardinality than Z_+ and lesser cardinality than R. The *generalized continuum hypothesis* asserts that given A infinite, 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].

§1-10

1-10 Well-Ordered Sets

One of the useful properties of the set 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 $\{1, 2\} \times Z_+$ 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 hard 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 $Z_+ \times Z_+$ in the dictionary order. Schematically, it can represented as an infinite sequence of infinite sequences. We assert that it is well-ordered. Let X be a nonempty subset of $Z_+ \times Z_+$. Let A be the subset of 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 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 13.)

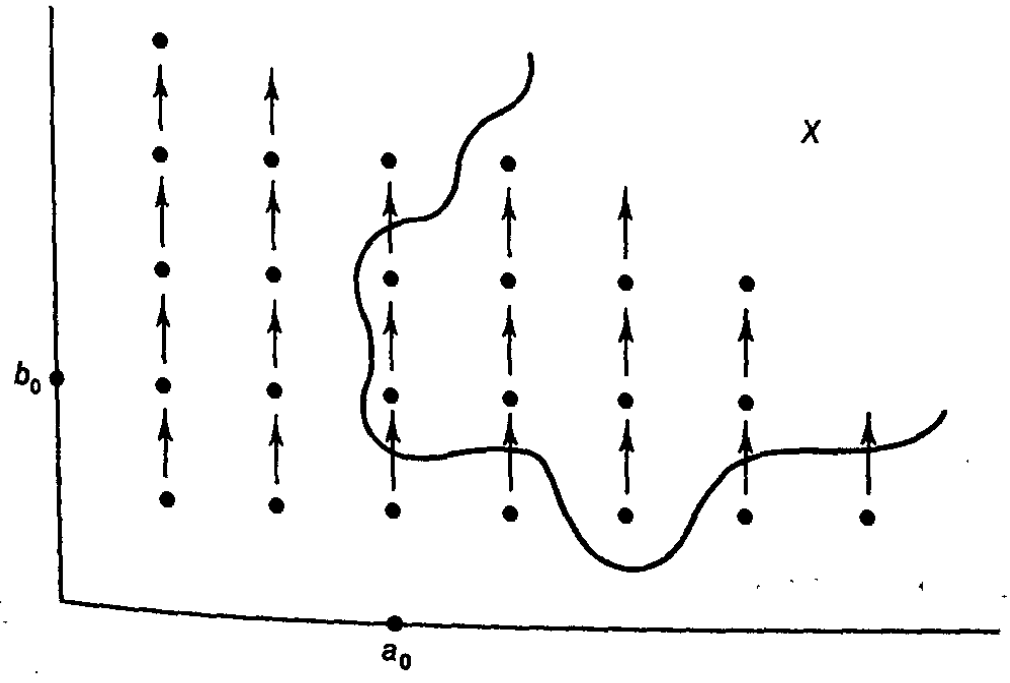


Figure 13

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 $Z_+ \times (Z_+ \times Z_+)$ is well-ordered in the dictionary order; it can be represented as an infinite sequence of infinite sequences of infinite sequences. Similarly, $(Z_+)^4$ is well-ordered in the dictionary order. And so on. But if you try to generalize to an infinite product of 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

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

can be used to define an order relation on A under which A has the same order type as the ordered set $\{1, \ldots, n\}$. 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* $\{1, \ldots, n\}$ *of* Z_+, *so it is necessarily well-ordered.*

Proof. This was given as an exercise in §1-6; we prove it here. First, one shows 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, one shows there is an order-preserving bijection of A with $\{1, \ldots, n\}$ 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

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

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

$$f(x) = f'(x) \quad \text{for } x \neq b,$$
$$f(b) = n. \quad \square$$

§ 1-10

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

$$Z_+,$$
$$\{1, \ldots, n\} \times Z_+,$$
$$Z_+ \times Z_+,$$
$$Z_+ \times (Z_+ \times Z_+)$$

are all countably infinite, but they all have different order types.

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

$$X = Z_+ \times Z_+ \times \cdots = (Z_+)^\omega$$

of 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 $n \geq 1$,

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

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.

We have seen that the dictionary order at least does not give a well-ordering of the set $(Z_+)^\omega$. Is there some other order relation on this set that is a well-ordering? No one has ever constructed a specific well-ordering of $(Z_+)^\omega$. 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.

In this book we are not going to prove the well-ordering theorem. It is long (although not exceedingly difficult) and primarily of interest to logicians. For those interested, a proof is outlined in the supplementary exercises at the end of the chapter.

One seldom needs the well-ordering theorem to prove theorems of topology in any case; we shall use it to prove a theorem only twice, once in the exercises of §4-5 and again in §6-3. For us, its primary function will be to provide us with a particular counterexample that will be very useful.

We shall not, in fact, need the full strength of the well-ordering theorem to construct the example we want. We shall need only the following weaker result, which we now assume:

Corollary. There exists an uncountable well-ordered set.

(It is of some interest to note that one can prove this weaker assertion without use of the choice axiom; see the supplementary exercises.)

To construct the desired example, we first need some terminology.

Definition. Let X be an ordered set. Given $\alpha \in X$, the set

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

is called the **section of X by α.**

Theorem 10.2. There exists an uncountable well-ordered set, every section of which is countable.

Proof. By assumption, there exists an uncountable well-ordered set X. From this fact it follows that there exists an uncountable well-ordered set Y, at least one *section* of which is uncountable. The set $\{1, 2\} \times X$ in the dictionary order is one such set; its section by any element of the form $2 \times x$ is uncountable. Consider the subset of Y consisting of those elements α for which the section S_α is uncountable; let Ω be the smallest such element. Then S_Ω is a well-ordered set, it is uncountable, and every section of S_Ω is countable. □

The well-ordered set of this theorem is called a **minimal uncountable well-ordered set**; the reason for the choice of terminology is fairly obvious. Its order type is, in fact, uniquely determined by the requirements of the theorem. (See the supplementary exercises.) We shall denote it throughout this book by S_Ω, and we shall use the symbol $\bar{S}_\Omega$ throughout to denote the well-ordered set $S_\Omega \cup \{\Omega\}$.

§ 1-10

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

Corollary 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 $\{1, 2\} \times Z_+$ and $Z_+ \times \{1, 2\}$ are well-ordered in the dictionary order. Do they have the same order type?
4. (a) Let 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 Z_-.
   (b) Show that if A is simply ordered and every countable subset of A is well-ordered, then A is well-ordered.
5. Let S_Ω be the minimal uncountable well-ordered set.
   (a) Show that S_Ω has no largest element.
   (b) Show that for every $\alpha < \Omega$, the set
   $$\{x \mid \alpha < x < \Omega\}$$
   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.
6. 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 $J_0 = J$.
7. Suppose that we call a subset J_0 of the well-ordered set J "semi-inductive" if for every element $\alpha \in J_0$, the immediate successor of α (if any) is in J_0. Show by means of an example that if J_0 contains the smallest element of J and is semi-inductive, J_0 need not equal all of 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. (a) Show that $(Z_+)^n$ has the order type of a section of $(Z_+)^{n+1}$, both in the dictionary order.

(b) Find a well-ordered set that for every n has a section having the order type of $(Z_+)^n$.

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*

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

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

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

satisfying $(*)$.

(d) Show by transfinite induction that for every $\alpha \in J$, there exists a function $h : S_\alpha \to C$ satisfying $(*)$.

(e) Prove the theorem.

11. Let A and B be two sets. Assuming 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.]

12. *Theorem (Principle of transfinite recursive definition). Let J be a well-ordered set with smallest element α_0. Let C be a set. Let $\mathcal{F}$ be the set of all functions mapping sections of J into C (including the empty function mapping S_{α_0} into C). Suppose that a function $\rho : \mathcal{F} \to C$ is given. Then there exists a unique function $h : J \to C$ such that*

$$h(x) = \rho(h \mid S_x)$$

for each $x \in J$.

[*Hint:* The proof follows the pattern outlined in Exercise 10.]

*1-11 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 another

†This section will be assumed in Chapter 5.

§1-11

consequence which is even more important in mathematics. It is the principle called the maximum principle, which we now discuss.

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 §1-3); 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 $\prec$.

Now we can state the maximum principle.

***Theorem* (*The maximum principle*).** *Let* A *be a set; let* $\prec$ *be a strict partial order on* A; *let* B *be a subset of* A *that is simply ordered by* $\prec$. *Then there exists a maximal simply ordered subset* C *of* A *containing* B.

Said differently, there exists a subset C of A containing B such that C is simply ordered by $\prec$ and such that no subset D of A properly containing C 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 14.

Now the maximum principle guarantees not only that $\mathcal{A}$ has a maximal simply ordered subcollection but also that starting with *any* simply ordered

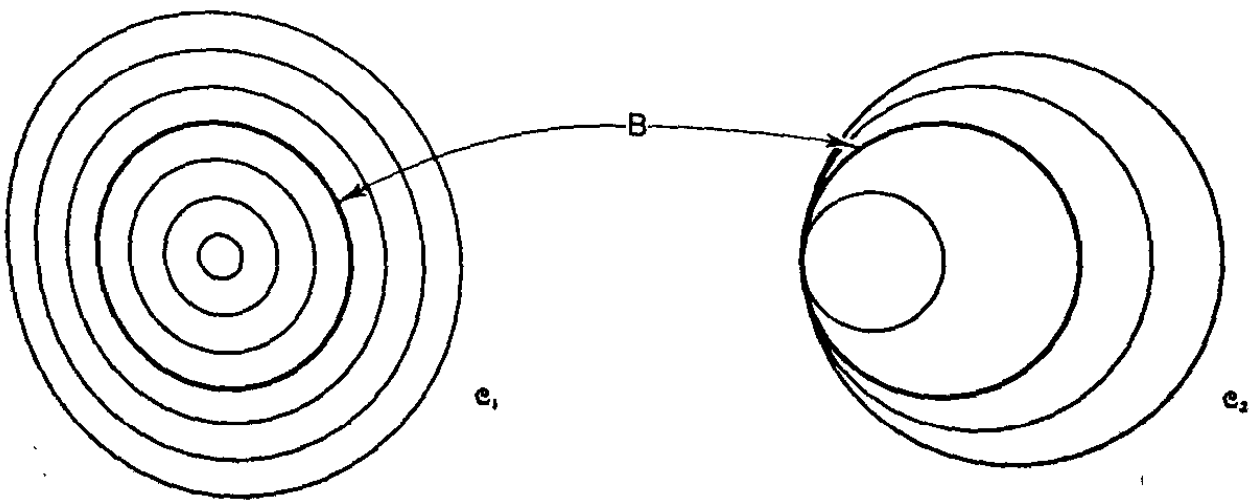


Figure 14

subcollection $\mathcal{B}$ of $\mathcal{A}$, there is a maximal simply ordered subcollection $\mathcal{C}$ of $\mathcal{A}$ that contains $\mathcal{B}$. If $\mathcal{B}$ consists of a single circular region B, it is easy to find several different maximal simply ordered subcollections of $\mathcal{A}$ that contain the element B. The collections $\mathcal{C}_1$ and $\mathcal{C}_2$ indicated in Figure 14 are two such. But if $\mathcal{B}$ consists of a more complicated collection, such as the three regions pictured in Figure 15, it is more work to find a maximal simply ordered subcollection containing $\mathcal{B}$. We leave it to you.

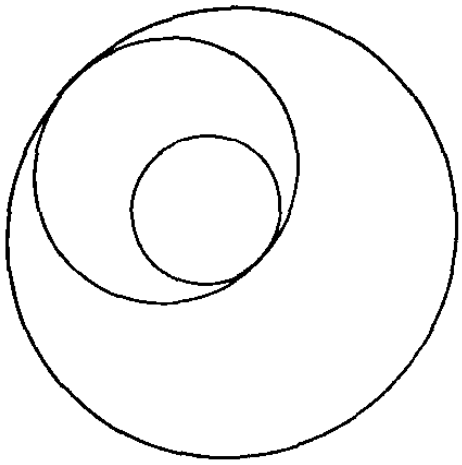


Figure 15

EXAMPLE 2. If (x_0, y_0) and (x_1, y_1) are two points of the plane 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 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 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. Take a box, and put into the box all the elements of B. Then examine the remaining elements of A one by one. First pick an element of A not in B. If it is comparable with everything in B, put it in the box; if not, throw it away. At the general step, you will have a collection of elements in the box and a collection of elements that you have 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 $A - B$ should happen to be *countable*, it is not hard to make this intui-

§ 1-11

tive proof into a real proof. Let us take the countably infinite case; the finite case is even easier. Index the elements of $A - B$ with the positive integers,

$$A - B = \{a_1, a_2, \ldots\},$$

letting distinct integers correspond to distinct points of $A - B$. This indexing gives a way of deciding what order to test the elements of $A - B$ in, and how to know when one has tested them all.

Specifically, we define a function $h : Z_+ \to \{0, 1\}$, which assigns the value 0 to i if we "put a_i in the box," and the value 1 if we "throw a_i away."

To be more formal, consider the following recursion formula:

$$h(i) = 0 \quad \text{if } a_i \text{ is comparable with every element of the set } B \cup \{a_j \mid 1 \leq j \leq i - 1 \text{ and } h(j) = 0\},$$

$$h(i) = 1 \quad \text{otherwise.}$$

By the principle of recursive definition, this formula determines a unique function $h : Z_+ \to \{0, 1\}$. It is easy to check that

$$B \cup \{a_j \mid h(j) = 0\}$$

is a maximal simply ordered subset of A.

If $A - B$ is not countable, a variant of this procedure will work, *if we assume the well-ordering theorem.* Instead of indexing the elements of $A - B$ with the set Z_+, we index them (in a bijective fashion) with the elements of some well-ordered set J:

$$A - B = \{a_\alpha \mid \alpha \in J\}.$$

For this we need the well-ordering theorem, so that we know there is a bijective correspondence between $A - B$ and a well-ordered set J. Then we can proceed as in the previous paragraph, letting the section S_α replace the section $\{1, \ldots, i - 1\}$ in the argument. Strictly speaking, you need to generalize the principle of recursive definition to well-ordered sets as well. (See Exercise 12 of the preceding section.)

This discussion is far from complete. But it does indicate that the well-ordering theorem and the maximum principle are related. It turns out that they are in fact *equivalent;* either of them implies the other. Furthermore, each of them is equivalent to the choice axiom.

We are not going to prove the maximum principle in this book; that would take us too far afield. For those interested, a proof is outlined in the supplementary exercises.

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 is 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 $a < b$ if $b - a$ is positive and rational. Show this is a strict partial order on R. What are the maximal simply ordered subsets?
2. Complete the discussion in Example 1.
3. (a) Let $<$ be a strict partial order on the set A. Define a relation on A by letting $a \leq b$ if either $a < b$ or $a = b$. Show that this relation has the following properties, which are called the **partial order axioms**:
   (i) $a \leq a$ for all $a \in A$.
   (ii) $a \leq b$ and $b \leq a \Rightarrow a = b$.
   (iii) $a \leq b$ and $b \leq c \Rightarrow a \leq c$.
   (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.
4. Let A be a set with a strict partial order $<$; let $x \in A$. Suppose that we wish to find a maximal simply ordered subset C of A that contains x. One plausible way of attempting to define C is to let C equal the set of all those elements of A that are *comparable* with x;

$$C = \{y \mid y \in A \text{ and either } x < y \text{ or } y < x\}.$$

   But this will not always work. In which of Examples 1 and 2 will this procedure succeed and in which will it not?
5. Complete the proof of the maximum principle in the case where $A - B$ is countable, by showing that $B \cup \{a_j \mid h(j) = 0\}$ is a maximal simply ordered subset of A.

*6. Given two points (x_0, y_0) and (x_1, y_1) of R^2, define

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

   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 R^2, and the curve $y = x^2$ is not. Find all maximal simply ordered subsets.

**Supplementary Exercises: Well-Ordering*

In the following exercises, we ask you to prove, without using the choice axiom, that there exists an uncountable well-ordered set. We also ask you to prove the equivalence of the choice axiom, the well-ordering theorem, and the maximum principle.

1. Check that the proof of the principle of transfinite recursive definition does not use the choice axiom. (See Exercise 12 of §1-10.)

2. Without using the choice axiom, prove the following:

Theorem. Let J and E be well-ordered sets. If J has the order type of a subset of E, then J has the order type of E or of a unique section of E, and not both.

Proof. Let e_0 be a fixed element of E. Define $h : J \to E$ by the formula

$$h(\alpha) = \begin{cases} \text{smallest } (E - h(S_\alpha)) & \text{if } h(S_\alpha) \neq E, \\ e_0 & \text{if } h(S_\alpha) = E. \end{cases}$$

(a) By hypothesis, there is an order-preserving map $i : J \to E$. Show that we have $h(\alpha) \leq i(\alpha)$ for all $\alpha \in J$; conclude that h satisfies the formula

$$(*) \qquad h(\alpha) = \text{smallest } (E - h(S_\alpha))$$

for all α.

(b) Show that $h(S_\alpha)$ is a section of E for each α; conclude that $h(J)$ equals E or a section of E.

(c) Show that h is order preserving.

(d) Show that if $k : J \to E$ is any order-preserving map such that $k(J)$ equals E or a section of E, then k satisfies $(*)$ and hence is unique.

3. Without using the choice axiom, prove the following:

Theorem. If A and B are two well-ordered sets, then either they have the same order type, or one has the order type of a section of the other.

[*Hint:* Consider the well-ordered set constructed in Exercise 8(a) of §1-10.]

Corollary. If A and B are two uncountable well-ordered sets all of whose sections are countable, then A and B have the same order type.

4. Without using the choice axiom, construct an uncountable well-ordered set, as follows. Let $\mathcal{A}$ be the collection of all pairs $(A, <)$, where A is a subset of 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 $h : E \to Z_+$ is a bijection, then h gives rise to a well-ordering of Z_+.]

This same argument, with 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.

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. Assuming Exercises 1 and 5, prove the following:

Theorem. The maximum principle is equivalent to the well-ordering theorem.

7. Assuming Exercises 1–3 and 5, 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:* If T_1 and T_2 are well-ordered sets and $h : T_1 \to T_2$ is order preserving and $h(T_1)$ equals T_2 or a section of T_2, then $h(S_x(T_1)) = S_{h(x)}(T_2)$ for all x. If T_1 and T_2 are towers, then $h(x) = x$ for all x.]

(b) 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 $T = X$.

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.

2-1 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 this 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 $\mathfrak{I}$ of subsets of X having the following properties:

(1) $\varnothing$ and X are in $\mathfrak{I}$.
(2) The union of the elements of any subcollection of $\mathfrak{I}$ is in $\mathfrak{I}$.
(3) The intersection of the elements of any finite subcollection of $\mathfrak{I}$ is in $\mathfrak{I}$.

A set X for which a topology $\mathfrak{I}$ has been specified is called **a topological space.**

Properly speaking, a topological space is an ordered pair $(X, \mathfrak{I})$ consisting of a set X and a topology $\mathfrak{I}$ on X, but we often omit specific mention of $\mathfrak{I}$ if no confusion will arise.

If X is a topological space with topology $\mathfrak{I}$, we say that a subset U of X is an **open set** of X if U belongs to the collection $\mathfrak{I}$. 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 1. The diagram in the upper right-hand corner indicates the topology in which

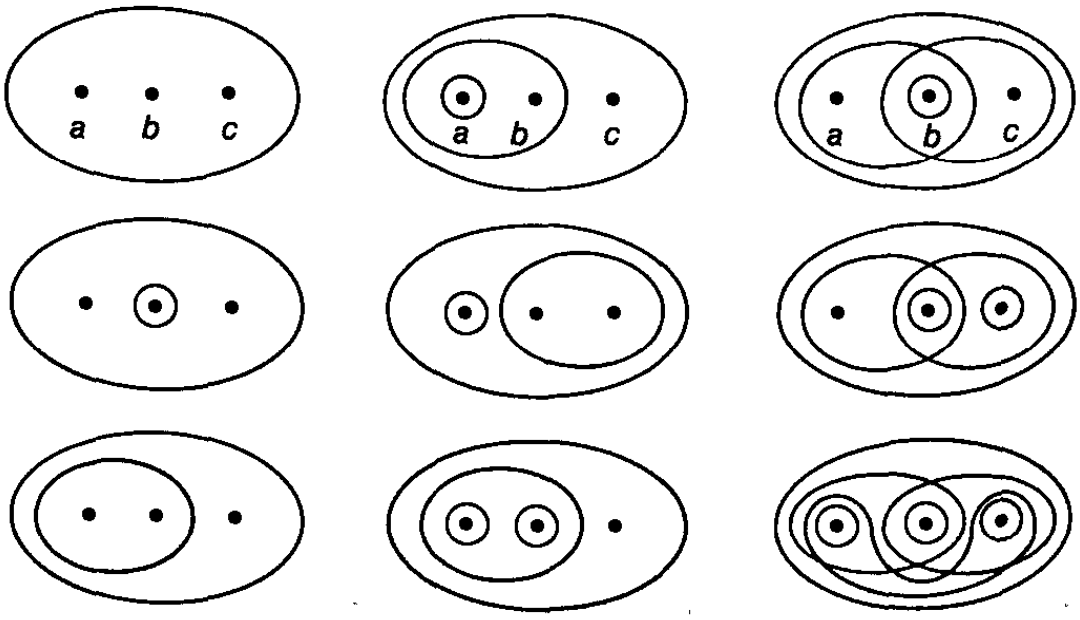


Figure 1

§2-1

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.

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 2 is a topology, for instance.

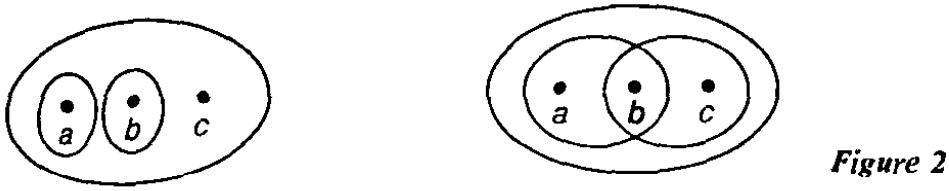


Figure 2

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 $\mathfrak{I}_f$ be the collection of all subsets U of X such that $X - U$ either is finite or is all of X. Then $\mathfrak{I}_f$ is a topology on X, called the **finite complement topology**. Both X and $\bigcirc$ are in $\mathfrak{I}_f$, since $X - X$ is finite and $X - \varnothing$ is all of X. If $\{U_\alpha\}$ is an indexed collection of elements of $\mathfrak{I}_f$, to show that $\bigcup U_\alpha$ is in $\mathfrak{I}_f$, 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 elements of $\mathfrak{I}_f$, to show that $\bigcap U_i$ is in $\mathfrak{I}_f$, 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 $\mathfrak{I}_c$ be the collection of all subsets U of X such that $X - U$ either is countable or is all of X. Then $\mathfrak{I}_c$ is a topology on X, as you can check.

Definition. Suppose that $\mathfrak{I}$ and $\mathfrak{I}'$ are two topologies on a given set X. If $\mathfrak{I}' \supset \mathfrak{I}$, we say that $\mathfrak{I}'$ is **finer** than $\mathfrak{I}$; if $\mathfrak{I}'$ *properly* contains $\mathfrak{I}$, we say that $\mathfrak{I}'$ is **strictly finer** than $\mathfrak{I}$.

We also say that $\mathfrak{I}$ is **coarser** than $\mathfrak{I}'$, or **strictly coarser**, in these two respective situations.

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 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.

2-2 *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 $B_3 \subset B_1 \cap B_2$.

Definition. If $\mathcal{B}$ is a basis for a topology on X, the **topology $\mathcal{T}$ generated by $\mathcal{B}$** is described 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 element of $\mathcal{B}$ is open in X under this definition, so that $\mathcal{B} \subset \mathcal{T}$. It is easy to check that this collection of subsets of X is 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 3. 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.

§ 2-2

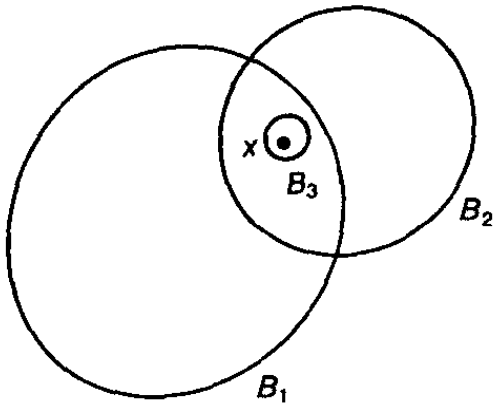


Figure 3

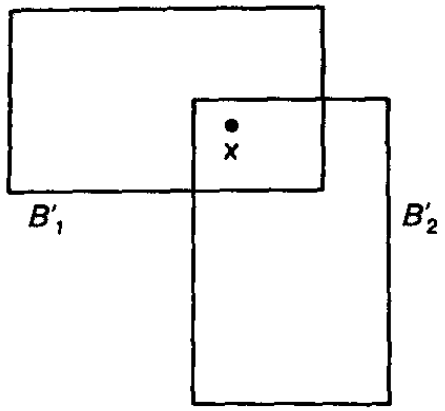


Figure 4

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 4; 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 $\mathfrak{I}$ 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 $\mathfrak{I}$, 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 $\mathfrak{I}$ and show that

$$U = \bigcup_{\alpha \in J} U_\alpha$$

belongs to $\mathfrak{I}$. 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 $\mathfrak{I}$ and show that $U_1 \cap U_2$ belongs to $\mathfrak{I}$. 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 5. Then $x \in B_3$ and $B_3 \subset U_1 \cap U_2$, so $U_1 \cap U_2$ belongs to $\mathfrak{I}$, by definition.

Finally, we show by induction that any finite intersection $U_1 \cap \cdots \cap U_n$ of elements of $\mathfrak{I}$ is in $\mathfrak{I}$. This fact is trivial for $n = 1$; we suppose it true for $n - 1$ and prove it for n. Now

$$(U_1 \cap \cdots \cap U_n) = (U_1 \cap \cdots \cap U_{n-1}) \cap U_n.$$

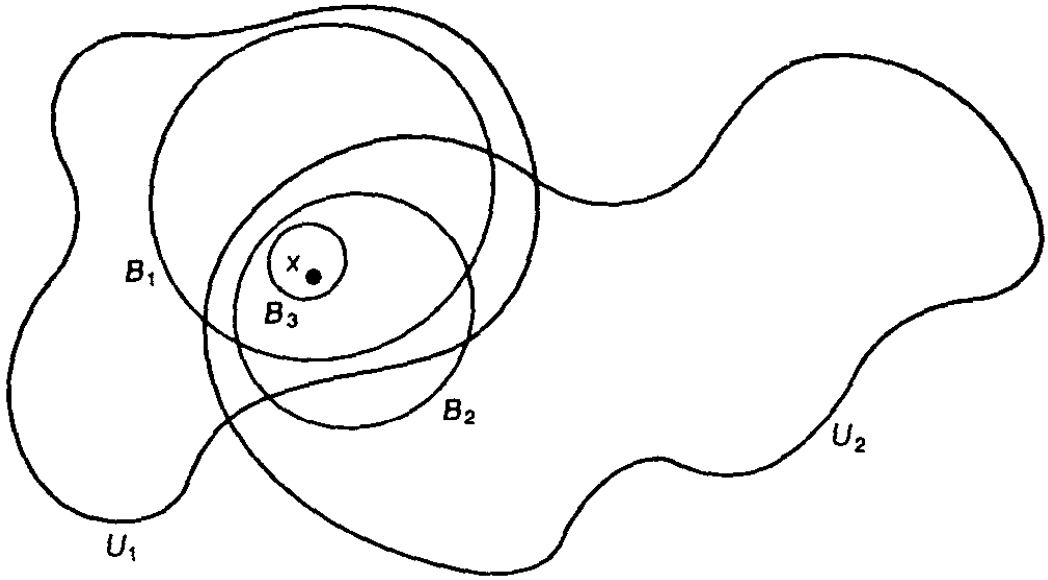


Figure 5

By hypothesis, $U_1 \cap \cdots \cap U_{n-1}$ belongs to $\mathfrak{I}$; by the result just proved, the intersection of $U_1 \cap \cdots \cap U_{n-1}$ and U_n also belongs to $\mathfrak{I}$.

Thus we have checked that the 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 2.1. *Let* X *be a set; let* $\mathcal{B}$ *be a basis for a topology* $\mathfrak{I}$ *on* X. *Then* $\mathfrak{I}$ *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 $\mathfrak{I}$. Because $\mathfrak{I}$ is a topology, their union is in $\mathfrak{I}$. Conversely, given $U \in \mathfrak{I}$, 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 $\mathcal{B}$. □

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 2.2. *Let* $\mathcal{B}$ *and* $\mathcal{B}'$ *be bases for the topologies* $\mathfrak{I}$ *and* $\mathfrak{I}'$, *respectively, on* X. *Then the following are equivalent:*

(1) $\mathfrak{I}'$ *is finer than* $\mathfrak{I}$.

(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* $x \in B' \subset B$.

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.

Proof of the lemma. (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 $x \in B' \subset B$. □

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 6 illustrates the proof. We shall treat this example more formally when we study metric spaces.

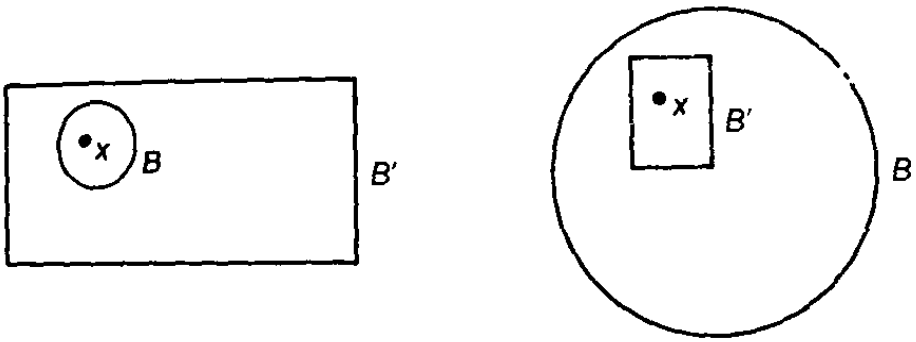


Figure 6

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 2.3. 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 $\mathcal{T}'$ denote the topology on X generated by $\mathcal{C}$; let $\mathcal{T}$ be the topology of X. The preceding lemma shows that $\mathcal{T}'$ is finer than $\mathcal{T}$. Conversely, since

each element of $\mathcal{C}$ is an element of $\mathfrak{I}$, so are arbitrary unions of elements of $\mathcal{C}$. Therefore, by Lemma 2.1, $\mathfrak{I}' \subset \mathfrak{I}$. We conclude that $\mathfrak{I}' = \mathfrak{I}$. □

There are two interesting topologies on the real line R which can be described in terms of bases:

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. 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 R. When R is given the lower limit topology, we denote it by R_l.

It is easy to see that both $\mathcal{B}$ and $\mathcal{B}'$ are bases; the intersection of two basis elements is either another basis element or is empty.

Whenever we consider R, we shall suppose that it is given the standard topology unless we specifically state otherwise. The lower limit topology will prove very useful to us in constructing counterexamples. The relation between these two topologies is the following:

Lemma 2.4. The lower limit topology $\mathfrak{I}'$ on R *is strictly finer than the standard topology $\mathfrak{I}$.*

Proof. Given a basis element (a, b) for $\mathfrak{I}$, and a point x of (a, b), the basis element $[x, b)$ for $\mathfrak{I}'$ contains x and lies in (a, b). Therefore, $\mathfrak{I}'$ is finer than $\mathfrak{I}$. On the other hand, given a basis element $[x, d)$ for $\mathfrak{I}'$, there is no open interval (a, b) satisfying the condition

$$x \in (a, b) \subset [x, d);$$

therefore, $\mathfrak{I}$ is not finer than $\mathfrak{I}'$. □

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 $\mathfrak{I}$ of all unions of finite intersections of elements of $\mathcal{S}$.

We must of course check that $\mathfrak{I}$ 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 $\mathfrak{I}$ of all unions of elements of $\mathcal{B}$ is a topology, by Lemma 2.1. 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 $\mathcal{B}$.

Subbases will prove useful to us only occasionally, so we shall not study them in detail.

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 §2-1. 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 $\mathfrak{I}_c$ given in Example 4 of §2-1 is a topology on the set X. Is the collection

$$\mathfrak{I}_\infty = \{U \mid X - U \text{ is infinite or empty or all of } X\}$$

a topology on X?

4. (a) If $\{\mathfrak{I}_\alpha\}$ is a collection of topologies on X, show that $\bigcap \mathfrak{I}_\alpha$ is a topology on X. Is $\bigcup \mathfrak{I}_\alpha$ a topology on X?

   (b) Let $\{\mathfrak{I}_\alpha\}$ be a collection of topologies on X. Show that there is a unique smallest topology on X containing all the collections $\mathfrak{I}_\alpha$, and a unique largest topology contained in all $\mathfrak{I}_\alpha$.

   (c) If $X = \{a, b, c\}$, let

$$\mathfrak{I}_1 = \{\varnothing, X, \{a\}, \{a, b\}\} \quad \text{and} \quad \mathfrak{I}_2 = \{\varnothing, X, \{a\}, \{b, c\}\}.$$

   Find the smallest topology containing $\mathfrak{I}_1$ and $\mathfrak{I}_2$, and the largest topology contained in $\mathfrak{I}_1$ and $\mathfrak{I}_2$.

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. Consider the following collections of subsets of R:

$$\mathcal{B}_1 = \{(a, b) \mid a < b\},$$
$$\mathcal{B}_2 = \{[a, b) \mid a < b\},$$
$$\mathcal{B}_3 = \{(a, b] \mid a < b\}, \text{ where } (a, b] = \{x \mid a < x \leq b\},$$
$$\mathcal{B}_4 = \mathcal{B}_1 \cup \{B - K \mid B \in \mathcal{B}_1\}, \text{ where } K = \{1/n \mid n \in Z_+\},$$
$$\mathcal{B}_5 = \{(a, +\infty) \mid a \in R\}, \text{ where } (a, +\infty) = \{x \mid x > a\},$$

$\mathcal{B}_6 = \{(-\infty, a) | a \in R\}$, where $(-\infty, a) = \{x | x < a\}$,
$\mathcal{B}_7 = \{B | R - B \text{ is finite}\}$.

(a) Show that each $\mathcal{B}_i$ is a basis for a topology on R.
(b) Compare these seven topologies with one another.
(c) Show that $\mathcal{B}_5 \cup \mathcal{B}_6$ is a subbasis that generates the same topology as $\mathcal{B}_1$.

7. (a) Apply Lemma 2.3 to show that the countable collection

$$\mathcal{B}'_1 = \{(a, b) | a < b, a \text{ and } b \text{ rational}\}$$

is a basis that generates the standard topology on R.

(b) Show that the collection

$$\mathcal{B}'_2 = \{[a, b) | a < b, a \text{ and } b \text{ rational}\}$$

is a basis that generates a topology different from the lower limit topology on R.

2-3 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:

$$(a, b) = \{x | a < x < b\},$$
$$(a, b] = \{x | a < x \leq b\},$$
$$[a, b) = \{x | a \leq x < b\},$$
$$[a, b] = \{x | a \leq x \leq b\}.$$

The notation used here is familar 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 R, as defined in the preceding section, is just the order topology derived from the usual order on R.

EXAMPLE 2. Consider the set $R \times R$ in the dictionary order; we shall denote the general element of $R \times R$ by $x \times y$, to avoid difficulty with notation. The set $R \times R$ has neither a largest nor a smallest element, so the order topology on $R \times R$ 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 7. The subcollection consisting of only intervals of the second type is also a basis for the order topology on $R \times R$, as you can check.

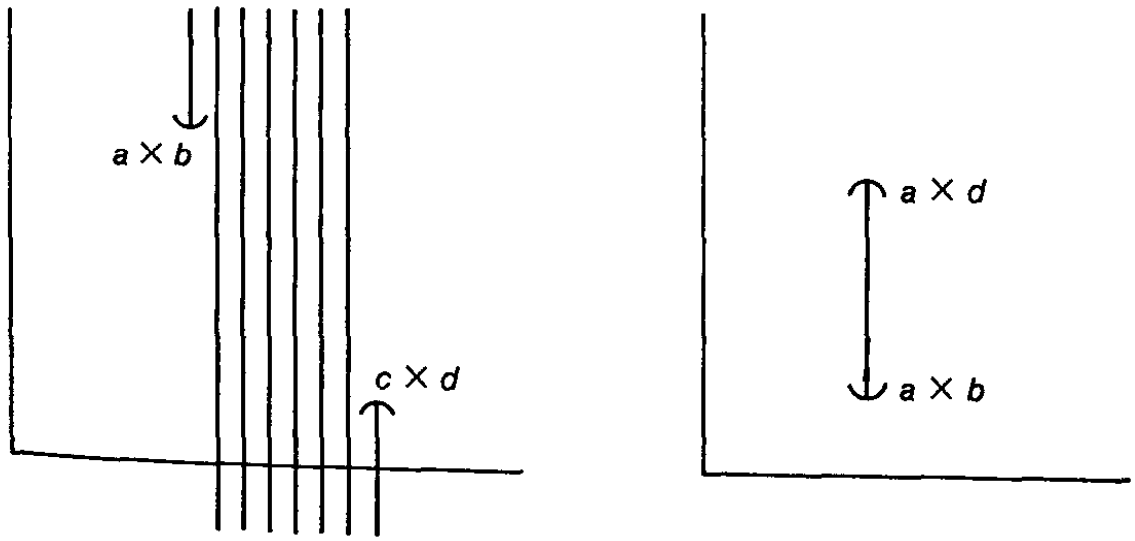


Figure 7

EXAMPLE 3. The positive integers Z_+ form an ordered set with a smallest element. The order topology on 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 $X = \{1, 2\} \times Z_+$ 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 **ciosed** 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; we leave this for you to check.

2-4 *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 8.

Note that the collection $\mathcal{B}$ is not a topology on $X \times Y$. The union of the

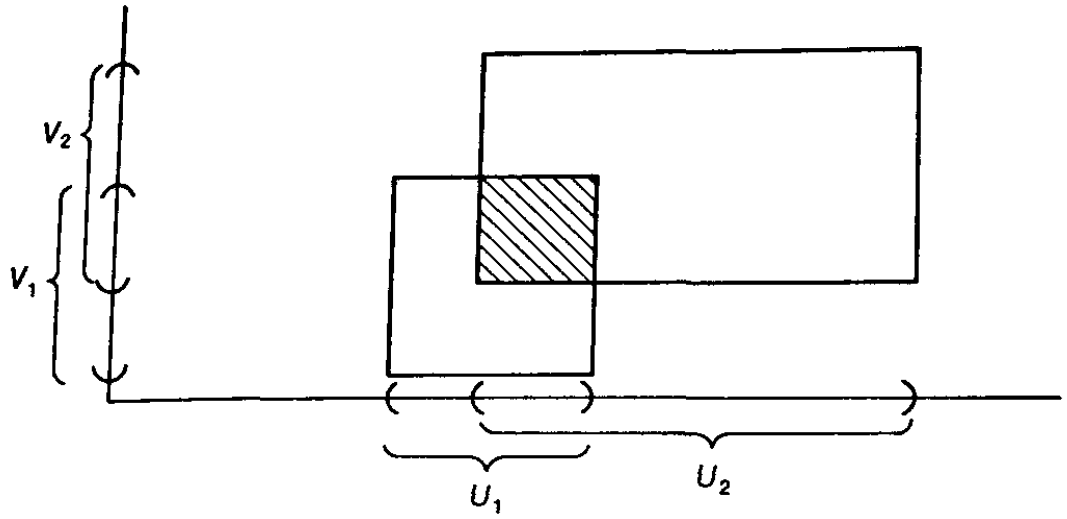


Figure 8

two rectangles pictured in Figure 8, for instance, is not a product of two sets, so it cannot belong to $\mathcal{B}$; but 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 4.1. 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 $X \times Y$.

Proof. We apply Lemma 2.3. 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 2.3, so $\mathcal{D}$ is a basis for $X \times Y$. □

EXAMPLE 1. We have a standard topology on R: the order topology. The product of this topology with itself is called the **standard topology** on $R \times R = R^2$. It has as basis the collection of all products of open sets of 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 R will also serve as a basis for the topology of R^2. Each such set can be pictured as the interior of a rectangle in R^2. Thus the standard topology on R^2 is just the one we considered in Example 2 of §2-2.

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 9. This fact leads to the following theorem:

Theorem 4.2. The collection

$$\mathcal{S} = \{\pi_1^{-1}(\mathrm{U}) \mid \mathrm{U} \text{ open in } \mathrm{X}\} \cup \{\pi_2^{-1}(\mathrm{V}) \mid \mathrm{V} \textit{ open in } \mathrm{Y}\}$$

is a subbasis for the product topology on $\mathrm{X} \times \mathrm{Y}$.

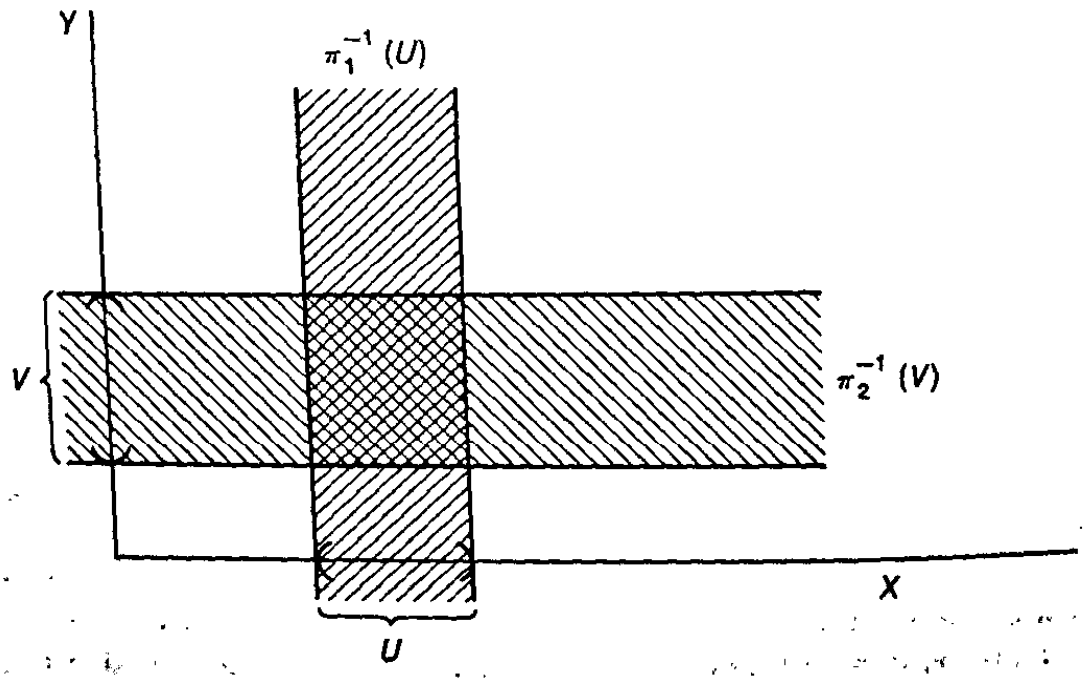


Figure 9

§2-5

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. □

2-5 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

$$\varnothing = Y \cap \varnothing \qquad \text{and} \qquad Y = Y \cap X,$$

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

$$(U_1 \cap Y) \cap \cdots \cap (U_n \cap Y) = (U_1 \cap \cdots \cap U_n) \cap Y,$$

$$\bigcup_{\alpha \in J} (U_\alpha \cap Y) = \left(\bigcup_{\alpha \in J} U_\alpha\right) \cap Y.$$

Lemma 5.1. 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 2.3 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 5.2. 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 $Y \cap V$. □

EXAMPLE 1. Consider the subset $Y = [0, 1]$ of the real line 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 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 R.

Note that sets of the first three types are the basis elements 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 R) and its order topology are the same.

Lest you think this remark utterly trivial, here is a subset of R for which the two topologies are *not* the same:

EXAMPLE 2. Let Y be the subset $[0, 1) \cup \{2\}$ of 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.

The anomaly illustrated in Example 2 does not occur if Y is an interval or a ray in the ordered set X; one has the following theorem, whose proof is left to the exercises:

Theorem 5.3. If X *is an ordered set in the order topology and if* Y *is an interval or a ray in* X, *then the subspace topology and the order topology on* Y *are the same.*

To avoid any ambiguity, let us agree that whenever we consider a subset of an ordered set X, we shall assume that it is given the subspace topology unless we specifically state otherwise.

Fortunately, there is no such ambiguity involving subspaces when one is dealing with the product topology:

Theorem 5.4. 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* $X \times Y$.

§2-5

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. □

Exercises

1. Show that if Y is a subspace of X, and A is a subset of Y, then the subspace topology on A as a subspace of Y is the same as the subspace topology on A 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 R. Which of the following sets are open in Y? Which are open in R?

$$A = \{x \mid \tfrac{1}{2} < |x| < 1\},$$
$$B = \{x \mid \tfrac{1}{2} < |x| \leq 1\},$$
$$C = \{x \mid \tfrac{1}{2} \leq |x| < 1\},$$
$$D = \{x \mid \tfrac{1}{2} \leq |x| \leq 1\},$$
$$E = \{x \mid 0 < |x| < 1 \text{ and } 1/x \notin Z_+\}.$$

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.
   (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.
   (c) What can you say if $\mathcal{T}' \supset \mathcal{T}$ and $\mathcal{U}' \supset \mathcal{U}$ and $\mathcal{U}' \neq \mathcal{U}$?
6. (a) Show that the collection of open rays in an ordered set A is a subbasis for the order topology on A.
   (b) Let X be an ordered set in the order topology. Let Y be an interval or ray in X; let $(-\infty, a)$ and $(a, +\infty)$ be open rays in X. Show that if $a \in Y$, then each of the sets $(-\infty, a) \cap Y$ and $(a, +\infty) \cap Y$ is an open ray of the ordered set Y, while if $a \notin Y$, each of these sets is either empty or all of Y.

(c) Conclude that if Y is an interval or ray in X, then the order topology and the subspace topology on Y are the same.

7. 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 R^2.

8. Show that the dictionary order topology on the set $R \times R$ is the same as the product topology $R_d \times R$, where R_d denotes the set R in the discrete topology. Compare this topology with the standard topology on R^2.

9. Let R denote the reals in their usual topology; let R_l denote the reals in the lower limit topology. If L is a straight line in the plane, describe the topology L inherits as a subspace of $R_l \times R$ and as a subspace of $R_l \times R_l$. In each case it is a familiar topology.

10. Let I denote the subspace $[0, 1]$ of R. Compare the product topology on $I \times I$, the dictionary order topology on $I \times I$, and the topology $I_d \times I$, where I_d denotes I in the discrete topology.

2-6 *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 R is closed because its complement

$$R - [a, b] = (-\infty, a) \cup (b, +\infty),$$

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 R is neither open nor closed.

EXAMPLE 2. In the plane R^2, the set

$$\{x \times y \mid x \geq 0 \quad \text{and} \quad y \geq 0\}$$

is closed, because its complement is the union of the two sets

$$(-\infty, 0) \times R \quad \text{and} \quad R \times (-\infty, 0),$$

each of which is a product of open sets of R and is therefore open in R^2.

§2-6

EXAMPLE 3. In the discrete topology on the set X, every set is open; it follows that every set is closed as well.

EXAMPLE 4. 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 R with Y. Similarly, (2, 3) is open as a subset of Y; it is even open as a subset of 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 6.1. *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 $i = 1, \ldots, n$, 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 6.2. *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 10.) 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 11.) Then $Y - A$ is open in Y, so that by definition it equals

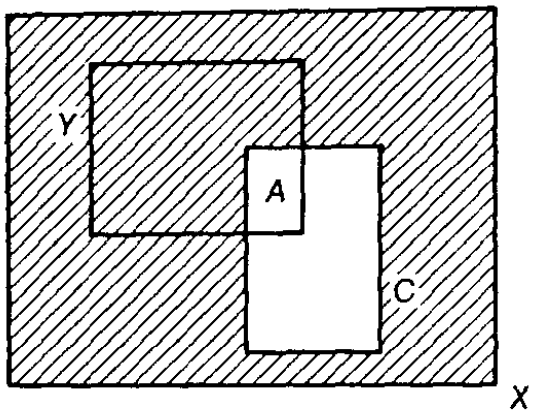


Figure 10

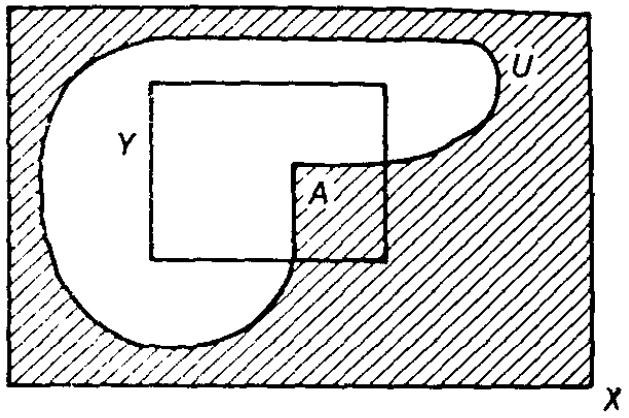


Figure 11

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:

Theorem 6.3. *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.

§2-6

The interior of A is denoted by Int A or by $\mathring{A}$, and the closure of A is denoted by Cl A or by $\bar{A}$. Obviously $\mathring{A}$ is an open set and $\bar{A}$ is a closed set; furthermore,

$$\mathring{A} \subset A \subset \bar{A}.$$

If A is open, $A = \mathring{A}$; 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 6.4. *Let* Y *be a subspace of* X*; let* A *be a subset of* Y*; let* $\bar{\text{A}}$ *denote the closure of* A *in* X*. Then the closure of* A *in* Y *equals* $\bar{\text{A}} \cap$ 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 6.2. 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 6.2, $B = C \cap Y$ 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 closure 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 6.5. *Let* A *be a subset of the topological space* X*.*

(a) *Then* $x \in \bar{\text{A}}$ *if and only if every open set* U *containing* x *intersects* A*.*

(b) *Supposing the topology of* X *is given by a basis, then* $x \in \bar{\text{A}}$ *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 5. Let X be the real line 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 $B = \{1/n | n \in Z_+\}$, then $\bar{B} = \{0\} \cup B$. If $C = \{0\} \cup (1, 2)$, then $\bar{C} = \{0\} \cup [1, 2]$. If Q is the set of rational numbers, then $\bar{Q} = R$. If Z_+ is the set of positive integers, then $\bar{Z}_+ = Z_+$. If R_+ is the set of positive reals, then the closure of R_+ is the set $R_+ \cup \{0\}$. (This is the reason we introduced the notation $\bar{R}_+$ for the set $R_+ \cup \{0\}$, back in §1-2.)

EXAMPLE 6. Consider the subspace $Y = (0, 1]$ of the real line R. The set $A = (0, \frac{1}{2})$ is a subset of Y; its closure in 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 7. Consider the real line 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 R is a limit point of A.

If $B = \{1/n | n \in Z_+\}$, then 0 is the only limit point of B. Every other point x of 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 Q is the set of rational numbers, every point of R is a limit point of Q. If Z_+ is the set of positive integers, no point of R is a limit point of Z_+. If R_+ is the set of positive reals, then every point of $\{0\} \cup R_+$ is a limit point of R_+.

Comparison of Examples 5 and 7 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 6.6. Let A *be a subset of the topological space* X*; let* A′ *be the set of all limit points of* A. *Then*

$$\bar{\mathrm{A}} = \mathrm{A} \cup \mathrm{A}'.$$

Proof. If x is in A', every neighborhood of x intersects A (in a point different from x). Therefore, by Theorem 6.5, 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 6.7. 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 order topology on R the interval (a, b) always has b as a limit point. But this is not true in order topologies in general.

For another example, consider the fact that in R or 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 one of the very first examples we had of a topological space; the topology on the three-point set $\{a, b, c\}$ indicated in Figure 12. In this space, the one-point set b is *not* closed, for its complement is not open.

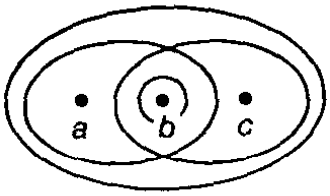


Figure 12

Topologies like this one are not really very interesting to mathematicians, 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 6.8. *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. □

Theorem 6.9. *Let* X *be a Hausdorff space; 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 $\{x_1, \ldots, x_m\}$ be the points of

§2-6

$U \cap (A - \{x\})$. The set $X - \{x_1, \ldots, x_m\}$ is an open set of X, since the finite point set $\{x_1, \ldots, x_m\}$ is closed; then

$$U \cap (X - \{x_1, \ldots, x_m\})$$

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. □

The Hausdorff condition is stronger than is needed to prove Theorems 6.8 and 6.9. The proofs would still hold if one assumed only the following weaker condition, which is usually called the T_1 axiom:

Given two distinct points a and b of X, each has a neighborhood not containing the other.

We shall give a few exercises involving this axiom; apart from that, we shall not use the T_1 axiom in this book.

In order to prove many of the interesting theorems of topology one needs the Hausdorff condition anyhow. Furthermore, most of the spaces that are important to mathematicians are Hausdorff spaces. The following theorem, whose proof is left to the exercises, gives some substance to the latter remark.

Theorem 6.10. 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.

For this reason 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
$$\mathfrak{I} = \{X - C \mid C \in \mathcal{C}\}$$
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) $\overline{A \cup B} = \bar{A} \cup \bar{B}$.
   (b) $\overline{\bigcup A_\alpha} \supset \bigcup \bar{A}_\alpha$; 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; let A' denote the set of limit points of A. 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) $\overline{A - B} = \bar{A} - \bar{B}$.
   (d) $(A \cup B)' = A' \cup B'$.
   (e) $(A \cap B)' = A' \cap B'$.

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. (a) Show that the T_1 axiom is equivalent to the requirement that finite point sets be closed.
   (b) Given a set X, show that the finite complement topology $\mathfrak{I}_f$, defined in Example 3 of §2-1, satisfies the T_1 axiom and is contained in every T_1 topology on X. Does $\mathfrak{I}_f$ satisfy the Hausdorff axiom?

§2-7

15. Consider the seven topologies on R given in Exercise 6 of §2-2.
(a) Determine the closure of the set $K = \{1/n | n \in Z_+\}$ under each of these topologies.
(b) Which of these topologies satisfy the Hausdorff axiom? The T_1 axiom?

16. Consider the two topologies on R given in Exercise 7 of §2-2. Determine the closures of the sets

$$A = (0, \sqrt{2}) \quad \text{and} \quad B = (\sqrt{2}, 3)$$

under each of these topologies.

17. Consider the set $X = [0, 1] \times [0, 1]$ in the dictionary order topology. Determine the closures of the following subsets of X:

$$A = \{(1/n) \times 0 | n \in Z_+\},$$
$$B = \{(1 - 1/n) \times (\tfrac{1}{2}) | n \in Z_+\},$$
$$C = \{x \times 0 | 0 < x < 1\},$$
$$D = \{x \times \tfrac{1}{2} | 0 < x < 1\},$$
$$E = \{\tfrac{1}{2} \times y | 0 < y < 1\}.$$

18. If $A \subset X$, we define the **boundary** of A by the equation

$$\text{Bd } A = \bar{A} \cap (\overline{X - A}).$$

(a) Show that Int A and Bd A are disjoint, and $\bar{A} = \text{Int } A \cup \text{Bd } A$.
(b) Show that Bd $A = \varnothing \Leftrightarrow A$ is both open and closed.
(c) Show that U is open $\Leftrightarrow$ Bd $U = \bar{U} - U$.
(d) If U is open, is it true that $U = \text{Int } (\bar{U})$? Justify your answer.

19. Find the boundary and the interior of each of the following subsets of R^2:
(a) $A = \{x \times y | y = 0\}$
(b) $B = \{x \times y | x > 0 \text{ and } y \neq 0\}$
(c) $C = A \cup B$
(d) $D = \{x \times y | x \text{ is rational}\}$
(e) $E = \{x \times y | 0 < x^2 + y^2 \leq 1\}$
(f) $F = \{x \times y | x \neq 0 \text{ and } y \leq 1/x\}$

*20. (Kuratowski) Consider the collection of all subsets A of the topological space X. The operations of closure $A \longrightarrow \bar{A}$ and complementation $A \longrightarrow X - A$ 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 R (in its usual topology) for which the maximum of 14 is obtained.

2-7 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 $f: X \longrightarrow Y$ 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,"

$$f: R \longrightarrow R.$$

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 R, and given $\epsilon > 0$, the interval $V = (f(x_0) - \epsilon, f(x_0) + \epsilon)$

§2-7

is an open set of the range space R. Therefore, $f^{-1}(V)$ is an open set in the domain space 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:

$f: R \longrightarrow R^2$ (curves in the plane)

$f: R \longrightarrow R^3$ (curves in space)

$f: R^2 \longrightarrow R$ (functions $f(x, y)$ of two real variables)

$f: R^3 \longrightarrow R$ (functions $f(x, y, z)$ of three real variables)

$f: R^2 \longrightarrow R^2$ (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 R denote the set of real numbers in its usual topology, and let R_l denote the same set in the lower limit topology. Let

$$f: R \longrightarrow R_l$$

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 R_l equals itself, which is not open in R. On the other hand, the identity function

$$g: R_l \longrightarrow R$$

is continuous, because the inverse image of (a, b) is itself, which is open in R_l.

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 in §2-10.

Theorem 7.1. Let X *and* Y *be topological spaces; let* $f: X \longrightarrow Y$. *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 *in* Y, *the set* $f^{-1}(B)$ *is closed in* X.

Proof. (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 $f(y)$. Hence $f(x) \in \overline{f(A)}$, 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 $\bar{A} \subset A$. By elementary set theory, we have $f(A) \subset B$. Therefore, if x is a point of $\bar{A}$,

$$f(x) \in f(\bar{A}) \subset \overline{f(A)} \subset \bar{B} = B,$$

so that $x \in f^{-1}(B) = A$. Thus $\bar{A} \subset A$ as desired.

(3) $\Rightarrow$ (1). Let V be an open set in Y. Let $B = Y - V$; then B is closed in Y. Since (3) holds, $f^{-1}(B)$ is closed in X. By elementary set theory,

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

so that $f^{-1}(V)$ is open, as desired. □

Homeomorphisms

Let X and Y be topological spaces; let $f : X \longrightarrow Y$ be a bijection. If both the function f and the inverse function

$$f^{-1} : Y \longrightarrow X$$

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 $f^{-1} : Y \longrightarrow X$ 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 13. So another way to define a homeomorphism is to say that it is a bijective correspondence $f : X \longrightarrow Y$ such that $f(U)$ is open if and only if U is open.

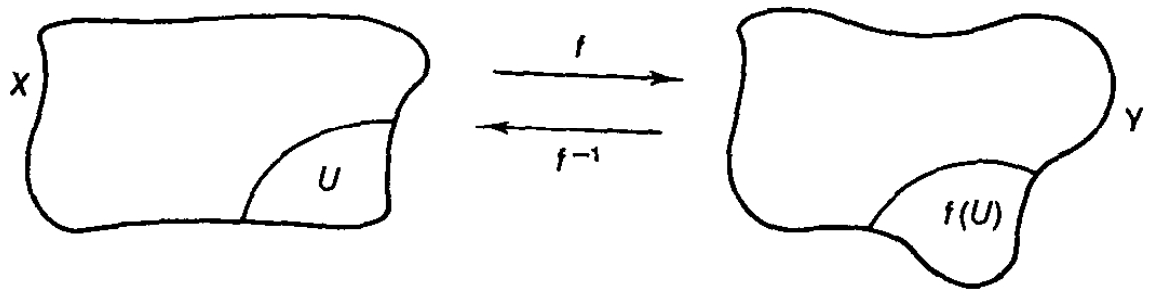


Figure 13

This remark shows that a homeomorphism $f : X \longrightarrow Y$ 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.

§2-7

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 $f: X \longrightarrow Y$ 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 $f': X \longrightarrow Z$ 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 $f: X \longrightarrow Y$ is a **topological imbedding**, or simply an **imbedding**, of X in Y.

EXAMPLE 4. The function $f: R \longrightarrow R$ given by $f(x) = 3x + 1$ is a homeomorphism. See Figure 14. If we define $g: R \longrightarrow R$ by the equation

$$g(y) = \tfrac{1}{3}(y - 1)$$

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 $F: (-1, 1) \longrightarrow R$ defined by

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

is a homeomorphism. See Figure 15. We have already noted in Example 9 of §1-3 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

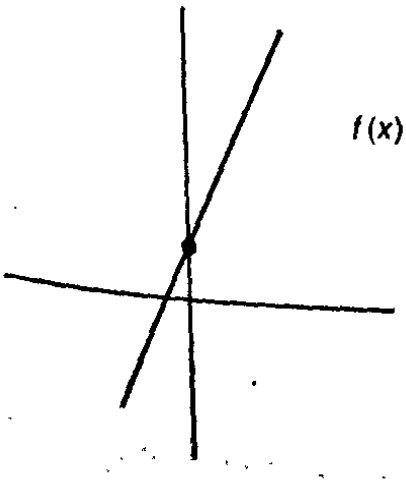


Figure 14

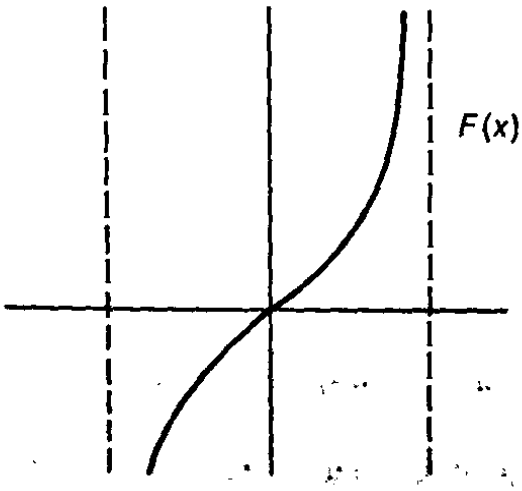


Figure 15

in R, and vice versa. As a result, F is automatically a homeomorphism of $(-1, 1)$ with 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 R.

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 $f: X \longrightarrow Y$ can be continuous without being a homeomorphism. One such function is the identity map $g: R_l \longrightarrow R$ considered in Example 3. Another is the following: Let S^1 denote the unit circle,

$$S^1 = \{x \times y | x^2 + y^2 = 1\},$$

considered as a subspace of the plane 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 R^2 such that $V \cap S^1 \subset f(U)$. See Figure 16.

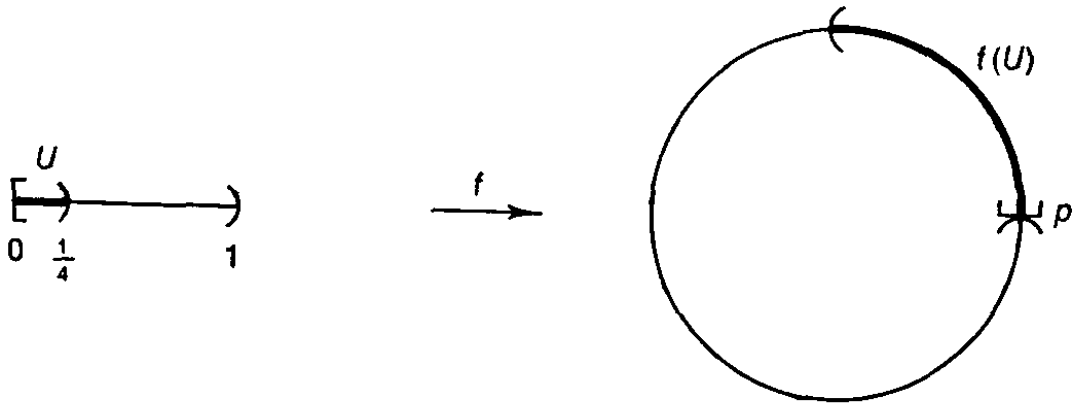


Figure 16

EXAMPLE 7. Consider the function

$$g: [0, 1) \longrightarrow R^2$$

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 anal-

ysis, 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 7.2 (***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.

(g) (*Continuity at each point*) *The map* $f: X \to Y$ *is continuous if 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 (g) holds for a particular point x of X, we say that f is continuous at the point x.

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.

(g) 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. □

Theorem 7.3 (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 [Theorem 7.2(f)].

EXAMPLE 8. Let us define a function $h: R \to R$ 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 R, the function h is continuous. (See Figure 17.) 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 R into 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 $[0, 1)$.

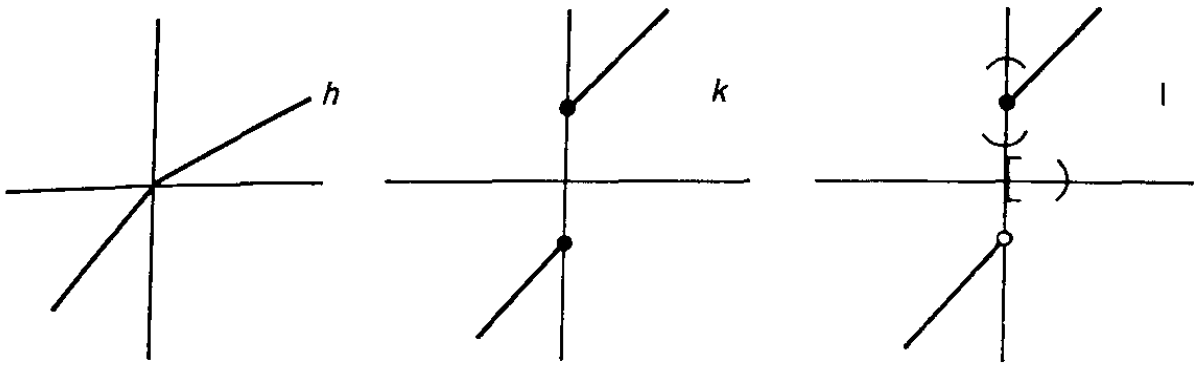


Figure 17

Theorem 7.4 (Maps into products). Let $\mathrm{f} : \mathrm{A} \longrightarrow \mathrm{X} \times \mathrm{Y}$ *be given by the equation*

$$\mathrm{f}(\mathrm{a}) = (\mathrm{f}_1(\mathrm{a}), \mathrm{f}_2(\mathrm{a})).$$

Then f *is continuous if and only if the functions*

$$\mathrm{f}_1 : \mathrm{A} \longrightarrow \mathrm{X} \quad \textit{and} \quad \mathrm{f}_2 : \mathrm{A} \longrightarrow \mathrm{Y}$$

are continuous.

The maps f_1 and f_2 are called the **coordinate functions** of f.

Proof. Let $\pi_1 : X \times Y \longrightarrow X$ and $\pi_2 : X \times Y \longrightarrow Y$ 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 13.)

EXAMPLE 9. In calculus, a *parametrized curve* in the plane is defined as 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 §2-10.

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 §2-10.

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.

§2-7

(a) Show that i is continuous $\Leftrightarrow \mathcal{T}'$ is finer than $\mathcal{T}$.
(b) Show that i is a homeomorphism $\Leftrightarrow \mathcal{T}' = \mathcal{T}$.
(c) If $n \in Z_+$, define a topology $\mathcal{T}_n$ on R by adjoining to the usual basis of open sets the one-point set $\{n\}$. Show that $(R, \mathcal{T}_1)$ and $(R, \mathcal{T}_2)$ are homeomorphic, but that $\mathcal{T}_1 \neq \mathcal{T}_2$.

4. Given $x_0 \in X$ and $y_0 \in Y$, show that the maps $f: X \longrightarrow X \times Y$ and $g: Y \longrightarrow X \times Y$ defined by

$$f(x) = x \times y_0 \quad \text{and} \quad g(y) = x_0 \times y$$

are imbeddings.

5. (a) Let X and Y be two ordered sets in the order topology. Show that if the map $f: X \longrightarrow Y$ is bijective and order preserving, it is a homeomorphism.
(b) Let $n \in Z_+$. Suppose you are given that for each real number $x \geq 0$, there is a unique real number $b \geq 0$ such that $b^n = x$. Denote b by $\sqrt[n]{x}$. Show that the function $g: \bar{R}_+ \longrightarrow \bar{R}_+$ defined by $g(x) = \sqrt[n]{x}$ is continuous.
(c) Let X be the subspace $(-\infty, -1) \cup [0, \infty)$ of R. Define $f: X \longrightarrow R$ by the equation

$$f(x) = \begin{cases} x + 1 & \text{if } x < -1, \\ x & \text{if } x \geq 0. \end{cases}$$

Show that f is bijective and order preserving and continuous. Is f a homeomorphism?

6. Show that the subspace (a, b) of R is homeomorphic with $(0, 1)$, and the subspace $[a, b]$ of R is homeomorphic with $[0, 1]$.

7. Find a function $f: R \longrightarrow R$ that is continuous at precisely one point.

8. (a) Suppose that $f: R \longrightarrow R$ is "continuous from the right," that is,

$$\lim_{x \to a+} f(x) = f(a),$$

for each $a \in R$. Show that f is continuous when considered as a function from R_l to R.
(b) What sort of functions $f: R \longrightarrow R$ are continuous when considered as maps from R to R_l? As maps from R_l to R_l?

9. Let Y be an ordered set in the order topology. Let $f, g: X \longrightarrow Y$ be continuous.
(a) Show that the set $\{x \mid f(x) \leq g(x)\}$ is closed in X.
(b) Let $h: X \longrightarrow Y$ be the function

$$h(x) = \min \{f(x), g(x)\}.$$

Show that h is continuous. [*Hint:* Use the pasting lemma.]

10. Let $\{A_\alpha\}$ be a collection of subsets of X; let $X = \bigcup_\alpha A_\alpha$. Let $f: X \longrightarrow Y$; suppose that $f \mid A_\alpha$ 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.

11. Let $f: A \longrightarrow B$ and $g: C \longrightarrow D$ be continuous functions. Let us define a map $f \times g: A \times C \longrightarrow B \times D$ by the equation

$$(f \times g)(a \times c) = f(a) \times g(c).$$

Show that $f \times g$ is continuous.

12. Let $F: X \times Y \longrightarrow Z$. We say that F is **continuous in each variable separately** if for each y_0 in Y, the map $h: X \longrightarrow Z$ defined by $h(x) = F(x \times y_0)$ is continuous, and for each x_0 in X, the map $k: Y \longrightarrow Z$ 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.

13. Let $F: R \times R \longrightarrow R$ 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 $g: R \longrightarrow R$ defined by $g(x) = F(x \times x)$.
(c) Show that F is not continuous.

14. Let $A \subset X$; let $f: A \longrightarrow Y$ be continuous; let Y be Hausdorff. Show that if f may be extended to a continuous function $g: \bar{A} \longrightarrow Y$, then g is uniquely determined by f.

15. Recall that a map $f: X \longrightarrow Y$ is an *open map* if for every set U that is open in X, the set $f(U)$ is open in Y. Which of the statements (a)–(f) of Theorem 7.2 remain true if one replaces the word "continuous" throughout by the word "open"?

2-8 *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 arbitrary cartesian products. There are two ways of generalizing the definition; the one that will later prove to be the more important we shall call the *product topology*.

One way to impose a topology on a product space is the following; it is a direct generalization of the way we defined a basis for the product topology on $X \times Y$:

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

$$\prod_{\alpha \in J} X_\alpha$$

§2-8

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:

$$(\prod_{\alpha \in J} U_\alpha) \cap (\prod_{\alpha \in J} V_\alpha) = \prod_{\alpha \in J} (U_\alpha \cap V_\alpha).$$

This topology is *not* the most useful one for the product space $\prod X_\alpha$, as we shall see.

A second way to generalize the previous definition is to 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**.

How does the product topology differ from the box topology? It is easier to answer this question if we look at 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 $\beta_1, \ldots, \beta_n$ 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 $i = 1, \ldots, n$. 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 an element of $\mathcal{B}$.

There is another way to describe this basis element which is particularly useful. Note that 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 $\beta_1, \ldots, \beta_n$. 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 $\alpha \neq \beta_1, \ldots, \beta_n$.

All this is summarized in the following theorem:

***Theorem 8.1** (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:

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:

***Theorem 8.2.** 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 on $\prod_{\alpha \in J} X_\alpha$.

***Theorem 8.3.** 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.*

§2-8

Theorem 8.4. If each space X_α is a Hausdorff space, then $\prod X_\alpha$ is a Hausdorff space in both the box and product topologies.

The proofs of these theorems follow the pattern of the proofs already given for $X \times Y$, so the details are left to you.

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 8.5. 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 continous, 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. See Example 2 following.

EXAMPLE 1. Consider euclidean n-space R^n. A basis for R consists of all open intervals in R; hence a basis for the topology of 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 R^n is a finite product, the box and product topologies agree. Whenever we consider R^n, we will assume that it is given this topology, unless we specifically state otherwise.

EXAMPLE 2. Now consider R^ω, the countably infinite product of R with itself. Recall that

$$R^\omega = \prod_{n \in Z_+} X_n,$$

where $X_n = R$ for each n. Let us define a function $f: R \longrightarrow R^\omega$ 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 $f_n : R \longrightarrow R$ is continuous; therefore, the function f is continuous if R^ω is given the product topology. But f is not continuous if R^ω is given the box topology. Consider, for example, the basis element

$$B = (-1, 1) \times (-\tfrac{1}{2}, \tfrac{1}{2}) \times (-\tfrac{1}{3}, \tfrac{1}{3}) \times \cdots$$

for the box topology. We assert that $f^{-1}(B)$ is not open in R. If $f^{-1}(B)$ were open in 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 8.2.
2. Prove Theorem 8.3.
3. Prove Theorem 8.4.
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. Let $\{X_\alpha\}$ be a family of spaces; let $A_\alpha \subset X_\alpha$ for each α.
   (a) Show that if A_α is closed in X_α, then $\prod A_\alpha$ is closed in $\prod X_\alpha$.
   (b) Show that
   $$\overline{\prod A_\alpha} = \prod \bar{A}_\alpha.$$
   (c) Which of (a) and (b) remain true if you use the box topology instead of the product topology?
6. A sequence (x_n) of points of X is said to converge to the point x of X if for every neighborhood U of x there is an N such that $x_n \in U$ for $n \geq N$. Let $(\mathbf{x}_n)$ be a sequence of points of the product space $\prod_{\alpha \in J} X_\alpha$. Show that the sequence $(\mathbf{x}_n)$ converges to $\mathbf{x}$ if and only if the sequence $\pi_\alpha(\mathbf{x}_n)$ converges to $\pi_\alpha(\mathbf{x})$ for each $\alpha \in J$. Is this true if you use the box topology instead of the product topology?
7. Let R^∞ be the subset of R^ω consisting of all sequences that are "eventually zero," that is, all $(x_1, x_2, \ldots)$ such that $x_i \neq 0$ for only finitely many values of i. What is the closure of R^∞ in R^ω in the box and product topologies? Justify your answer.
8. Show that the product topology is the coarsest (smallest) topology on $\prod X_\alpha$ relative to which each projection function π_β is continuous.
9. 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 $f_\alpha : A \longrightarrow X_\alpha$.
   (a) Show there is a unique coarsest topology $\mathfrak{I}$ 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 $\mathfrak{I}$.

(c) Show that a map $g: Y \to A$ is continuous relative to $\mathfrak{I}$ 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 $\mathfrak{I}$ is an open set of Z.

2-9 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
$$d: X \times X \to R$$
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 18.

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 $B(y, \delta) \subset B_1 \cap B_2$.

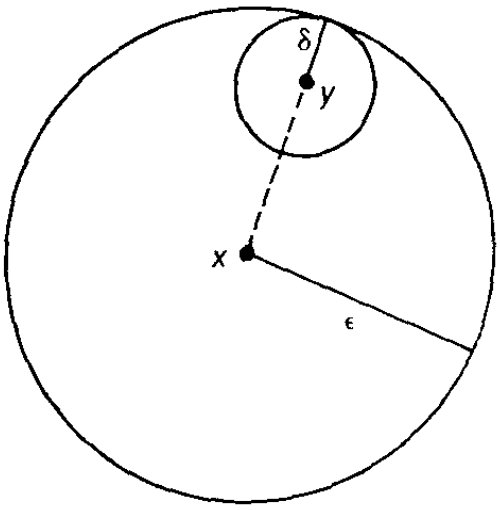


Figure 18

Using what we have just proved, we can rephrase the definition of the metric topology as follows:

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 $\boldsymbol{R}$ 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,

$$(a, b) = B(x, \epsilon),$$

§2-9

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 which will guarantee that 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 R^n and R^ω 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. Therefore, they are not topological properties. 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, the **diameter** of A is defined to be the number

$$\text{diam } A = \text{lub } \{d(a_1, a_2) \mid a_1, a_2 \in A\}.$$

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 $\bar{d}$ which gives the topology of X, relative to which *every* subset of X is bounded. It is defined as follows:

Theorem 9.1. Let X *be a metric space with metric* d. *Define* $\bar{d} : X \times X \to R$ *by the equation*

$$\bar{d}(x, y) = \min \{d(x, y), 1\}.$$

Then $\bar{d}$ *is a metric that induces the topology of* X.

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 $\bar{d}$.

The fact that $\bar{d}$ and d induce the same topology follows from the inclusions

$$B_d(x, \epsilon) \subset B_{\bar{d}}(x, \epsilon),$$

$$B_{\bar{d}}(x, \delta) \subset B_d(x, \epsilon),$$

where $\delta = \min\{\epsilon, 1\}$. We merely apply the following lemma. □

Lemma 9.2. *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 2.2 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 2.2 applies to show $\mathcal{T}'$ is finer than $\mathcal{T}$. □

Now let us show that R^n and R^ω are metrizable.

Definition. Given $\mathbf{x} = (x_1, \ldots, x_n)$ in R^n, we define the **norm** of $\mathbf{x}$ by the equation

$$\|\mathbf{x}\| = (x_1^2 + \cdots + x_n^2)^{1/2};$$

and we define the **euclidean metric** d on 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

$$\rho(\mathbf{x}, \mathbf{y}) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}.$$

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 the euclidean metric on R^n.

To show that ρ is a metric is easier. Only the triangle inequality is non-

trivial. From the triangle inequality for 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 $R = R^1$, these two metrics coincide with the standard metric for R. In the plane 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.

Each of these metrics induces the usual topology on R^n:

Theorem 9.3. The topologies on R^n *induced by the euclidean metric* d *and the square metric* ρ *are the same as the product topology on* R^n.

Proof. Let $\mathbf{x} = (x_1, \ldots, x_n)$ and $\mathbf{y} = (y_1, \ldots, y_n)$ be two points of 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 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 $\mathbf{x} = (x_1, \ldots, x_n)$ 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 $\epsilon = \min \{\epsilon_1, \ldots, \epsilon_n\}$. 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 $\mathbf{y} \in B_\rho(\mathbf{x}, \epsilon)$, 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. □

The preceding theorem is a metrization theorem for R^n. We seek to generalize it to R^ω.

As a first try for a metric on R^ω, it is natural to try to generalize either the euclidean metric or the square metric, by defining

$$d(\mathbf{x}, \mathbf{y}) = [\textstyle\sum_{i=1}^{\infty} (x_i - y_i)^2]^{1/2}$$

and

$$\rho(\mathbf{x}, \mathbf{y}) = \text{lub}\,\{|x_i - y_i|\}.$$

But these formulas simply do not make sense on all of R^ω; the series need not converge, and the set may not be bounded.

We can avoid the difficulty with the metric ρ, however, by first replacing the metric $|x - y|$ on R by its bounded counterpart $\bar{d}$. This idea gives us our second try for a metric on R^ω.

Suppose that we take the standard bounded metric on R,

$$\bar{d}(a, b) = \min\{|a - b|, 1\};$$

and, given two points $\mathbf{x} = (x_1, x_2, \ldots)$ and $\mathbf{y} = (y_1, y_2, \ldots)$ of R^ω, define

$$\bar{\rho}(\mathbf{x}, \mathbf{y}) = \text{lub}\,\{\bar{d}(x_i, y_i)\}.$$

It is easy to check that $\bar{\rho}$ is a metric on R^ω. Unfortunately, however, it does not induce the product topology, as we shall see. Therefore, it is of no use for proving R^ω metrizable. It is, however, very important in its own right; it will appear frequently in this book (and in other parts of mathematics).

There is nothing special about R^ω as far as this metric is concerned; one can define it on 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 R^J, let us define a metric $\bar{\rho}$ on R^J by the equation

$$\bar{\rho}(\mathbf{x}, \mathbf{y}) = \text{lub}\,\{\bar{d}(x_\alpha, y_\alpha) \mid \alpha \in J\},$$

where $\bar{d}$ is the standard bounded metric on R. The metric $\bar{\rho}$ is called the **uniform metric** on R^J, and the topology it induces is called the **uniform topology.**

The relation between this topology and the product topology is the following:

Theorem 9.4. *The uniform topology on* $\mathbf{R}^J$ *is finer than the product topology; they are different if* $\mathbf{J}$ *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 $\alpha_1, \ldots, \alpha_n$ be the indices for which $U_\alpha \neq R$. Then for each α_i, choose $\epsilon_i > 0$ so that

$$B_{\bar{d}}(x_{\alpha_i}, \epsilon_i) \subset U_{\alpha_i};$$

this we can do because U_{α_i} is open in R. Let $\epsilon = \min\{\epsilon_1, \ldots, \epsilon_n\}$; then

$$B_{\bar{\rho}}(\mathbf{x}, \epsilon) \subset \prod U_\alpha.$$

For if $\mathbf{z}$ is a point of 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 is easy to show that these two topologies are different if J is infinite; we leave it to you. □

We still have not found what we want, a metric that gives the product topology on R^ω. But we are almost there. It turns out that by modifying the uniform metric only slightly one can obtain the desired metric:

Theorem 9.5. Let $\bar{d}(a, b) = \min\{|a - b|, 1\}$ *be the standard bounded metric on* R. *If* $\mathbf{x}$ *and* $\mathbf{y}$ *are two points of* R^ω, *define*

$$D(\mathbf{x}, \mathbf{y}) = \text{lub}\left\{\frac{\bar{d}(x_i, y_i)}{i}\right\}.$$

Then D *is a metric that induces the product topology on* R^ω.

Proof. The properties of a metric are satisfied trivially except for the triangle inequality, which is proved by noting that for all i,

$$\frac{\bar{d}(x_i, z_i)}{i} \leq \frac{\bar{d}(x_i, y_i)}{i} + \frac{\bar{d}(y_i, z_i)}{i} \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z}),$$

whence

$$\text{lub}\left\{\frac{\bar{d}(x_i, z_i)}{i}\right\} \leq D(\mathbf{x}, \mathbf{y}) + D(\mathbf{y}, \mathbf{z}).$$

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

$$V = (x_1 - \epsilon, x_1 + \epsilon) \times \cdots \times (x_N - \epsilon, x_N + \epsilon) \times R \times R \times \cdots.$$

We assert that $V \subset B_D(\mathbf{x}, \epsilon)$: Given any $\mathbf{y}$ in R^ω,

$$\frac{\bar{d}(x_i, y_i)}{i} \leq \frac{1}{N} \quad \text{for } i \geq N.$$

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 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

$$U = \prod_{i \in Z_+} U_i$$

for the product topology, where U_i is open in R for $i = \alpha_1, \ldots, \alpha_n$ and $U_i = R$ 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 R centered about x_i and lying in U_i for $i = \alpha_1, \ldots, \alpha_n$; choose each $\epsilon_i \leq 1$. Then define

$$\epsilon = \min \{\epsilon_i / i \mid i = \alpha_1, \ldots, \alpha_n\}.$$

We assert that

$$\mathbf{x} \in B_D(\mathbf{x}, \epsilon) \subset U.$$

Let 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 $i = \alpha_1, \ldots, \alpha_n$, 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. □

It is natural to ask whether this theorem can be generalized still further. Can one perhaps prove that R^J is metrizable for arbitrary J? The answer is "no," as we shall see in the next section. The situation is no better if we use the box topology instead of the product topology; in fact, under the box topology not even R^ω is metrizable, as we shall see.

Exercises

1. (a) In 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 R^n. Sketch the basis elements under d' when $n = 2$.

(b) More generally, given $p \geq 1$, define

$$d'(\mathbf{x}, \mathbf{y}) = [\textstyle\sum_{i=1}^{n} |x_i - y_i|^p]^{1/p}$$

for $\mathbf{x}, \mathbf{y} \in R^n$. Assume that d' is a metric. Show that it induces the usual topology on R^n.

2. Let X be a set; let d be a metric on X. Show that the topology on X induced by d is the coarsest topology relative to which the function $d: X \times X \to R$ is continuous.

3. Show that $R \times R$ in the dictionary order topology is metrizable.

4. (a) Compare the box, product, and uniform topologies on R^ω.

§2-9

(b) In which topologies are the following functions from R to R^ω continuous?

$$f(t) = (t, 2t, 3t, \ldots),$$
$$g(t) = (t, t, t, \ldots),$$
$$h(t) = (t, \tfrac{1}{2}t, \tfrac{1}{3}t, \ldots).$$

(c) In which topologies do the following sequences converge?

$$\begin{aligned}
\mathbf{w}_1 &= (1, 1, 1, 1, \ldots), & \mathbf{x}_1 &= (1, 1, 1, 1, \ldots),\\
\mathbf{w}_2 &= (0, 2, 2, 2, \ldots), & \mathbf{x}_2 &= (0, \tfrac{1}{2}, \tfrac{1}{2}, \tfrac{1}{2}, \ldots),\\
\mathbf{w}_3 &= (0, 0, 3, 3, \ldots), & \mathbf{x}_3 &= (0, 0, \tfrac{1}{3}, \tfrac{1}{3}, \ldots),\\
&\cdots & &\cdots\\
\mathbf{y}_1 &= (1, 0, 0, 0, \ldots), & \mathbf{z}_1 &= (1, 1, 0, 0, \ldots),\\
\mathbf{y}_2 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0, 0, \ldots), & \mathbf{z}_2 &= (\tfrac{1}{2}, \tfrac{1}{2}, 0, 0, \ldots),\\
\mathbf{y}_3 &= (\tfrac{1}{3}, \tfrac{1}{3}, \tfrac{1}{3}, 0, \ldots), & \mathbf{z}_3 &= (\tfrac{1}{3}, \tfrac{1}{3}, 0, 0, \ldots),\\
&\cdots & &\cdots
\end{aligned}$$

5. Let R^∞ be the subset of R^ω consisting of all sequences that are eventually zero. What is the closure of R^∞ in R^ω in the uniform topology? Justify your answer.

6. Let $\bar{\rho}$ be the uniform metric on R^ω. Given $\mathbf{x} = (x_1, x_2, \ldots) \in R^\omega$ 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. Let X be a subset of R^ω having the property that for every pair $\mathbf{x}, \mathbf{y}$ of points of X, the series

$$\sum_{i=1}^{\infty} (x_i - y_i)^2$$

converges. On X we have the three topologies it inherits as a subspace of R^ω in the box, uniform, and product topologies. We also have the topology given by the metric

$$d(\mathbf{x}, \mathbf{y}) = \left[\sum_{i=1}^{\infty} (x_i - y_i)^2\right]^{1/2}.$$

(The fact that d is a metric follows from Exercise 9.)
(a) What can you say in general about how these four topologies on the set X compare?
(b) What can you say if X is the subset

$$\tilde{R}^n = \{(x_1, x_2, \ldots) \mid x_i = 0 \text{ for } i > n\}?$$

(c) What can you say if X is the subset

$$R^\infty = \bigcup \tilde{R}^n?$$

(d) What can you say if X is the **Hilbert cube**

$$H = \prod_{n \in \mathbb{Z}_+} [0, 1/n]?$$

8. Show that the euclidean metric d on R^n is a metric, as follows: If $\mathbf{x}, \mathbf{y} \in R^n$ and $c \in R$, define

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, \ldots, x_n + y_n),$$
$$c\mathbf{x} = (cx_1, \ldots, cx_n),$$
$$\mathbf{x} \cdot \mathbf{y} = x_1 y_1 + \cdots + x_n y_n.$$

(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 $\mathbf{x}, \mathbf{y} \neq \mathbf{0}$, 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.

9. Let ℓ^2 denote the subset of R^ω consisting of all sequences $(x_1, x_2, \ldots)$ 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 $\mathbf{x}, \mathbf{y} \in \ell^2$, then $\sum |x_i y_i|$ converges. [*Hint:* Use (b) of Exercise 8 to show that the partial sums are bounded.]
(b) Let $c \in R$. Show that if $\mathbf{x}, \mathbf{y} \in \ell^2$, then so are $\mathbf{x} + \mathbf{y}$ and $c\mathbf{x}$.
(c) Show that

$$d(\mathbf{x}, \mathbf{y}) = [\textstyle\sum_{i=1}^{\infty} (x_i - y_i)^2]^{1/2}$$

is a well-defined metric on ℓ^2. It is called the *ℓ^2-metric.*

10. Show that $d'(x, y) = d(x, y)/(1 + d(x, y))$ is a bounded metric for X if d is a metric for X.

2-10 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, Z_+ and R), and others are not. See Example 3.

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 R^n and R^ω 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 R^ω, so we leave it to the exercises.

§2-10

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 §2-7. One is the process of taking sums, differences, products, and quotients of continuous real-valued functions. The other is taking limits of uniformly convergent sequences of continuous functions.

Theorem 10.1. 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*

$$d_X(x, y) < \delta \implies d_Y(f(x), f(y)) < \epsilon.$$

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 remind you that a *sequence* of points of a set X is simply a function mapping Z_+ into X; that is, it is an element of the space X^ω. We usually denote a sequence by

$$(x_n) \quad \text{or} \quad (x_1, x_2, \ldots).$$

Definition. A sequence $(x_1, x_2, \ldots)$ of points of X is said to **converge** to the point x of X if for every neighborhood U of x there exists a positive integer N such that x_i lies in U for all $i \geq N$.

If the sequence (x_n) converges to x, we write

$$x_n \longrightarrow x.$$

Of course, a sequence need not converge at all. But if it does converge, it converges to only one point, *provided that X is Hausdorff*. For if (x_n) converges to x, and if $y \neq x$, we need merely choose disjoint neighborhoods U and V of x and y, respectively, and note that since U contains the points x_i for all but finitely many values of i, the set V cannot.

It is intuitively 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 10.2* (*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 6.5. 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 10.3. *Let* $f: X \to Y$; *let* X *be metrizable. The function* f *is continuous if and only if for every convergent sequence* $x_n \to x$ *in* X, *the sequence* $f(x_n)$ *converges to* $f(x)$.

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 10.2 and Theorem 10.3 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 $\{U_n\}_{n \in Z_+}$ 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 conntability axiom.**

If X has a countable basis $\{U_n\}$ at x, then the proof of Lemma 10.2 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 10.3 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 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 10.4. The addition, subtraction, and multiplication operations are continuous functions from $\mathrm{R} \times \mathrm{R}$ *into* R; *and the quotient operation is a continuous function from* $\mathrm{R} \times (\mathrm{R} - \{0\})$ *into* 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 10.5. If X *is a topological space, and if* $\mathrm{f}, \mathrm{g} : \mathrm{X} \to \mathrm{R}$ *are continuous functions, then* $\mathrm{f} + \mathrm{g}$, $\mathrm{f} - \mathrm{g}$, *and* $\mathrm{f} \cdot \mathrm{g}$ *are continuous. If* $\mathrm{g}(\mathrm{x}) \neq 0$ *for all* x, *then* f/g *is continuous.*

Proof. The map $h : X \to R \times R$ defined by

$$h(x) = f(x) \times g(x)$$

is continuous, by Theorem 7.4. The function $f + g$ equals the composite of h and the addition operation

$$+ : R \times R \longrightarrow R;$$

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 for a sequence of functions.

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 $n \geq N$ 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 10.6 (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 $x \in X$,

$$d(f_n(x), f(x)) < \epsilon/4.$$

Finally, using continuity of f_N, choose a neighborhood U of x_0 such that f_N carries U into the $\epsilon/2$ 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

$$d(f(x), f_N(x)) < \epsilon/4 \quad \text{(by choice of } N\text{)},$$
$$d(f_N(x), f_N(x_0)) < \epsilon/2 \quad \text{(by choice of } U\text{)},$$
$$d(f_N(x_0), f(x_0)) < \epsilon/4 \quad \text{(by choice of } N\text{)}.$$

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 R^X of all functions $f : X \to R$, in the uniform metric $\bar{\rho}$. It is not hard to see that a sequence of functions $f_n : X \to R$ 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 $(R^X, \bar{\rho})$. We leave the proof to the exercises.

We conclude the section with some examples of spaces that are not metrizable.

EXAMPLE 1. *R^ω in the box topology is not metrizable.*

We shall show that the sequence lemma does not hold for R^ω. Let A be the subset of R^ω consisting of those points all of whose coordinates are positive:

$$A = \{(x_1, x_2, \ldots) \mid x_i > 0 \text{ for all } i \in Z_+\}.$$

Let $\mathbf{0}$ be the "origin" in R^ω, that is, the point $(0, 0, \ldots)$ 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

$$(\tfrac{1}{2}b_1, \tfrac{1}{2}b_2, \ldots)$$

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

$$\mathbf{a}_n = (x_{1n}, x_{2n}, \ldots, x_{in}, \ldots).$$

Every coordinate x_{in} is positive, so we can construct a basis element B' for the box topology on 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 R with itself is not metrizable.*

Let J be an uncountable index set; we show that R^J does not satisfy the sequence lemma (in the product topology).

Let A be the subset of R^J consisting of all points (x_α) such that $x_\alpha = 0$ for finitely many values of α and $x_\alpha = 1$ for all other values of α. Let $\mathbf{0}$ be the "origin" in 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 $U_\alpha \neq R$ for only finitely many values of α, say for $\alpha = \alpha_1, \ldots, \alpha_n$. Let (x_α) be the point of A defined by letting $x_\alpha = 0$ for $\alpha = \alpha_1, \ldots, \alpha_n$ 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. Each point $\mathbf{a}_n$ is a point of the product space having only finitely many coordinates equal to 0. 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 zero. 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 R, and let U be the open set $\pi_\beta^{-1}(U_\beta)$ in 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 $\mathbf{0}$.

EXAMPLE 3. *The well-ordered set $\bar{S}_\Omega$ is not metrizable in the order topology.*

Recall that S_Ω is the minimal uncountable well-ordered set constructed in §1-10, and that $\bar{S}_\Omega$ denotes the set $S_\Omega \cup \{\Omega\}$. Note first that, in the order topology, Ω is a limit point of S_Ω. (This is the reason we introduced the notation $\bar{S}_\Omega$ for the set $S_\Omega \cup \{\Omega\}$, back in §1-10.) For if $(a, \Omega]$ is any basis element containing Ω, it must intersect S_Ω; otherwise we would have

$$S_\Omega = S_a \cup \{a\},$$

although S_Ω is uncountable and S_a is countable.

We assert that there is no sequence of points of S_Ω converging to Ω. For if (x_n) is any sequence of points of S_Ω, then the set $\{x_n | n \in Z_+\}$ is a countable subset of S_Ω and as such has an upper bound b that lies in S_Ω (Corollary 10.3 of Chapter 1). Then $(b, \Omega]$ is a basis element containing Ω that contains *no* point whatever of the sequence (x_n).

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 $f: X \longrightarrow Y$ 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. Show that a countable product $\prod X_n$ of metrizable spaces is metrizable. [*Hint:* Let d_n be a metric for X_n that is bounded by 1. Define $D(\mathbf{x}, \mathbf{y}) = \text{lub}\,\{d_n(x_n, y_n)/n\}$.]

4. Show that the spaces S_Ω and R_l satisfy the sequence lemma. (This does not, of course, imply that they are metrizable.)

5. *Theorem. Let $x_n \longrightarrow x$ and $y_n \longrightarrow y$ in the space R. Then*

$$x_n + y_n \longrightarrow x + y,$$

$$x_n - y_n \longrightarrow x - y,$$

$$x_n y_n \longrightarrow xy,$$

and provided that each $y_n \neq 0$ and $y \neq 0$,

$$x_n / y_n \longrightarrow x/y.$$

[*Hint:* Apply Theorem 10.3 and Lemma 10.4; recall that if $x_n \longrightarrow x$ and $y_n \longrightarrow y$, then $x_n \times y_n \longrightarrow x \times y$.]

6. Define $f_n: [0, 1] \longrightarrow R$ 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 $f_n: X \longrightarrow R$ be a sequence of functions. Let $\bar{\rho}$ be the uniform metric on the space R^X. Show that the sequence (f_n) converges uniformly to the function $f: X \longrightarrow R$ if and only if the sequence (f_n) converges to f as elements of the metric space $(R^X, \bar{\rho})$.

8. Let X be a topological space and let Y be a metric space. Let $f_n: X \longrightarrow Y$ 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 $f_n : R \longrightarrow R$ be the function

$$f_n(x) = \frac{1}{n^3[x - (1/n)]^2 + 1}.$$

(See Figure 19.) Let $f: R \longrightarrow R$ be the zero function.

(a) Show that $f_n(x) \longrightarrow f(x)$ for each $x \in R$.

(b) Show that f_n does not converge uniformly to f. (This shows that the converse of Theorem 10.6 does not hold; the limit function f may be continuous even though the convergence is not uniform.)

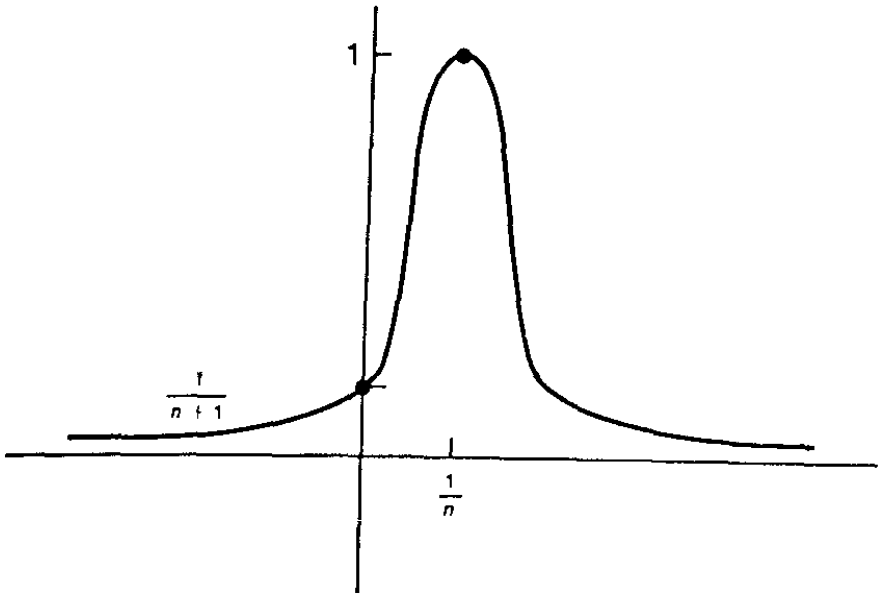


Figure 19

10. Using the closed set formulation of continuity (Theorem 7.1), show that the following are closed subsets of R^2:

$$A = \{x \times y | xy = 1\},$$
$$S^1 = \{x \times y | x^2 + y^2 = 1\},$$
$$B^2 = \{x \times y | x^2 + y^2 \leq 1\}.$$

The set B^2 is called the (closed) **unit ball** in 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 $s_n \longrightarrow s$, 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 $f_n : X \to R$, let

$$s_n(x) = \sum_{i=1}^{n} f_i(x).$$

Prove the Weierstrass *M*-test for uniform convergence:

If $|f_i(x)| \leq b_i$ *for all* $x \in X$ *and all* i, *and if the series* $\sum b_i$ *converges, then the sequence* (s_n) *converges uniformly to a function* s.

[*Hint:* Let $r_n = \sum_{i=n+1}^{\infty} b_i$. 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 R, as follows: Use the metric $d(a, b) = |a - b|$ on R and the metric

$$\rho((x, y), (x_0, y_0)) = \max \{|x - x_0|, |y - y_0|\}$$

on R^2.

(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 $\epsilon > 0$, let

$$3\delta = \min \{\epsilon/(|x_0| + |y_0| + 1), \sqrt{\epsilon}\}$$

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 $R - \{0\}$ to R. [*Hint:* Given $x_0 \neq 0$ and $\epsilon > 0$, let $\delta = \min \{|x_0|/2, x_0^2\epsilon/2\}$. Note that $d(1/x, 1/x_0) = |x - x_0|/|xx_0|$. If $|x - x_0| < \delta$, then $|xx_0 - x_0^2| < |x_0|^2/2$, so $xx_0 - x_0^2 > -x_0^2/2$ and $xx_0 > x_0^2/2 > 0$.]

(d) Show that the subtraction and quotient operations are continuous.

*2-11 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 20. 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 21. Formalizing these constructions involves the concept of quotient topology.

†This section will be used when we prove the Jordan curve theorem in Chapter 8.

§2-11

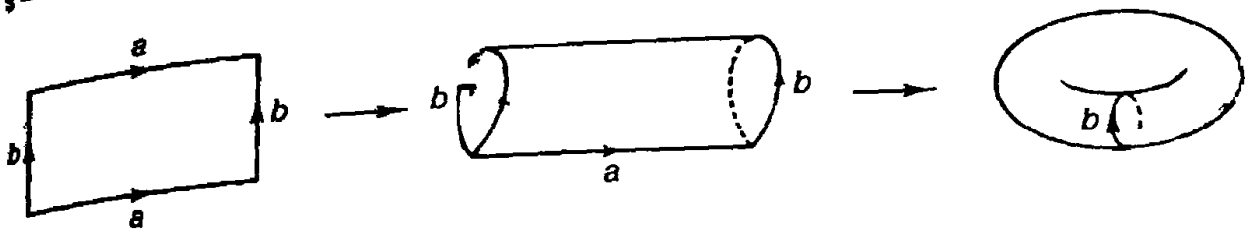


Figure 20

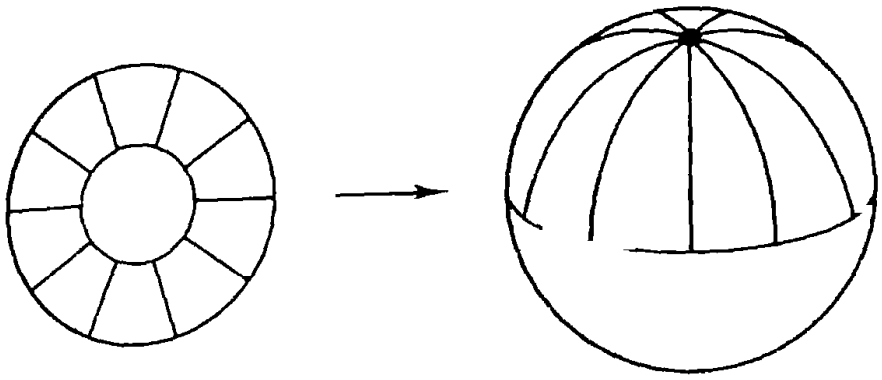


Figure 21

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 the 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 $p^{-1}(p(C))$. 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 2.)

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 $\mathfrak{I}$ on A relative to which p is a quotient map; it is called the **quotient topology** induced by p.

The topology $\mathfrak{I}$ 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 $\mathfrak{I}$ 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

$$p^{-1}(\bigcup_{\alpha \in J} U_\alpha) = \bigcup_{\alpha \in J} p^{-1}(U_\alpha),$$
$$p^{-1}(\bigcap_{i=1}^{n} U_i) = \bigcap_{i=1}^{n} p^{-1}(U_i).$$

EXAMPLE 1. Let $\pi_1 : X \times Y \to X$ be the projection map; π_1 is continuous and surjective. If $U \times V$ is a basis element for $X \times Y$, the image set $\pi_1(U \times V) = U$ is open in X. It follows readily that π_1 is an open map. In general, π_1 is not a closed map; the projection $\pi_1 : R \times R \to R$ carries the closed set $\{x \times y | xy = 1\}$ onto the nonclosed set $R - \{0\}$, for instance.

EXAMPLE 2. Let X be the subspace $[0, 1] \cup [2, 3]$ of R, and let Y be the subspace $[0, 2]$ of 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 a surjective, continuous, closed map. It is not, however, an open map; the image of the open set $[0, 1]$ of X is not open in Y.

EXAMPLE 3. Let p be the map of the real line 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.

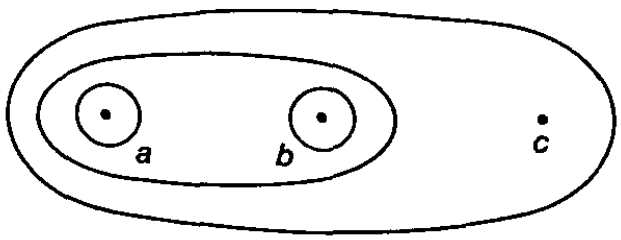


Figure 22

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.

§2-11

Some mathematicians call X^* a *decomposition space* or an *identification space* of X. For they think of X^* as having been obtained by "identifying all the elements in each partition class to a single point."

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 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 open sets in X of the form $p^{-1}(U)$ are pictured by the shaded regions in Figure 23; each is an open set of X that equals a union of equivalence classes.

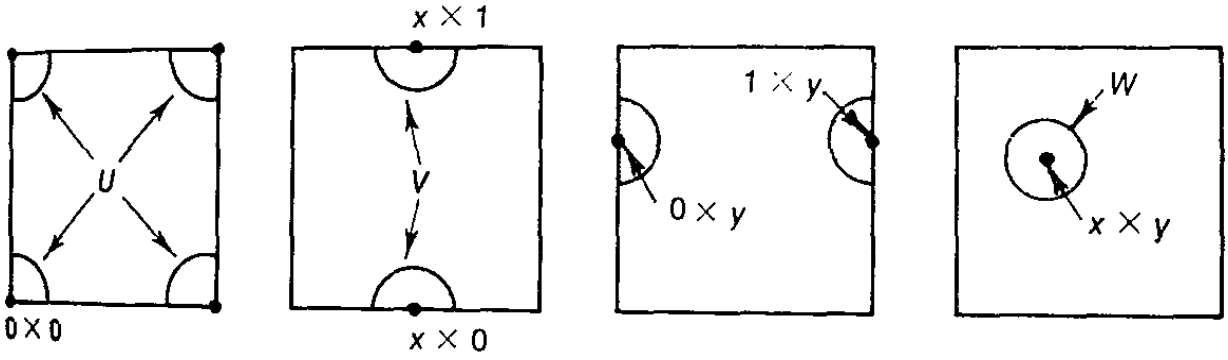


Figure 23

The image of each of these sets under p is an open set of X^*, as indicated in Figure 24. 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.

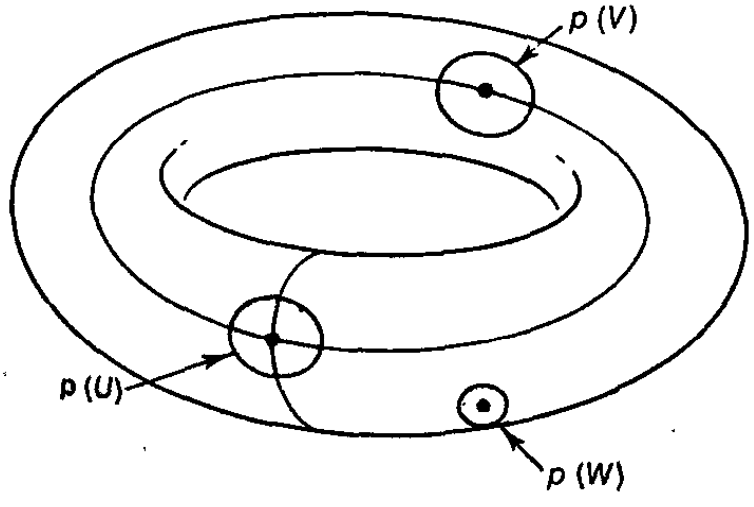


Figure 24

EXAMPLE 5. Let X be the closed unit ball

$$\{x \times y | x^2 + y^2 \leq 1\}$$

in 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 $S^1 = \{x \times y | x^2 + y^2 = 1\}$. Typical open sets in X of the form $p^{-1}(U)$ are pictured by the shaded regions in Figure 25. One can show that X^* is homeomorphic with the subspace of R^3 called the unit 2-sphere, defined by

$$S^2 = \{(x, y, z) | x^2 + y^2 + z^2 = 1\}.$$

It is pictured in Figure 26.

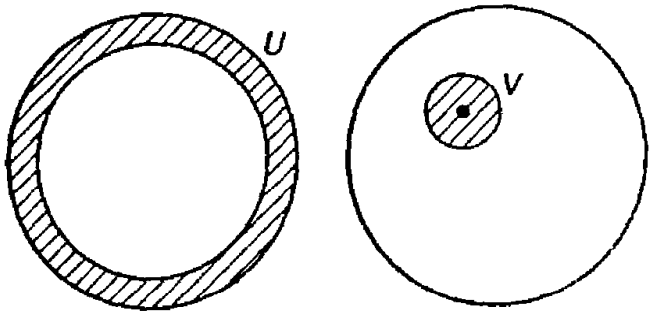


Figure 25

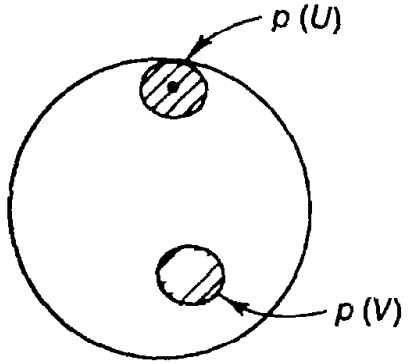


Figure 26

Now we explore the relationship between this new concept, that of quotient topology, and the concepts we have treated previously. It is interesting to note that many of these concepts do not behave the way one might wish.

It is not hard to see that *subspaces* do not behave well: If $p : X \longrightarrow Y$ is a quotient map and A is a subspace of X, then the map $p' : A \longrightarrow p(A)$ obtained by restricting both the domain and range of p need not be a quotient map. See Example 6 below. If it should happen, however, that A is open in X and p is an *open* map, then p' is a quotient map; the same is true if both A and p are closed. This we leave to you to check.

EXAMPLE 6. Consider the subspace $A = [0, 1] \cup (2, 3]$ of R; it is a subspace of the space X of Example 2. Suppose that we restrict the map p of Example 2 to A. Then $q = p | A : A \longrightarrow [0, 2]$ is continuous and surjective, but it is not a quotient map. The set

$$q^{-1}((1, 2]) = (2, 3],$$

for instance, is closed in the domain space A; but $(1, 2]$ is not closed in the image space $Y = [0, 2]$.

Composites of maps behave nicely; it is trivial to check that the composite of two quotient maps is a quotient map.

On the other hand, the *product* of two quotient maps need not be a quo-

§2-11

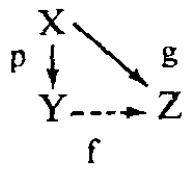
tient map: If $p: A \to B$ and $q: C \to D$ are quotient maps, it does not follow that the map

$$p \times q: A \times C \to B \times D$$

defined by $(p \times q)(a \times c) = p(a) \times q(c)$ is a quotient map. See Example 8. One needs further conditions on either the maps or the spaces in order for this to be true. One such, a condition on the spaces, is called *local compactness*; we shall study it later. (See the exercises of §3-8.) 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.

About *continuous functions* on quotient spaces there is something to be said. 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 11.1. Let $p: X \to Y$ *be a quotient map. Let* Z *be a space and let* $g: X \to Z$ *be a continuous map that is constant on each set* $p^{-1}(\{y\})$, *for* $y \in Y$. *Then* g *induces a continuous map* $f: Y \to Z$ *such that* $f \circ p = g$.



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 $x \in X$, $f(p(x)) = g(x)$. To show that f is continuous, let V be an open set in Z. Continuity of g implies that

$$g^{-1}(V) = p^{-1}(f^{-1}(V))$$

is open in X. Because p is a quotient map, $f^{-1}(V)$ must be open in Y. □

The *Hausdorff axiom* does not behave well; even though one starts with a Hausdorff space X, there is no reason that the quotient space needs to be Hausdorff. There is a simple condition for one-point sets to be closed in a quotient space: If $p: X \to Y$ is a quotient map, then one-point sets are closed in Y if and only if each set $p^{-1}(\{y\})$ is closed in X. Therefore, if X^* is a partition of X into closed sets, all one-point sets are closed in the quotient space X^*. Conditions that will ensure that X^* is Hausdorff are harder to find. There is a somewhat complicated condition called *upper semicontinuity* that one can impose on the partition X^* to ensure that it is Hausdorff.

(See Exercise 13 of §4-2.) There is, however, one situation in which it is easy to see that X^* is a Hausdorff space:

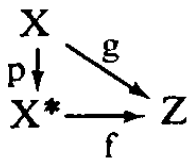
Theorem 11.2. *Let* $g: X \to Z$ *be a surjective continuous map. Let* X^* *be the following collection of subsets of* X:

$$X^* = \{g^{-1}(\{z\}) \mid z \in Z\}.$$

Give X^* *the quotient topology.*

(a) *If* Z *is Hausdorff, so is* X^*.

(b) *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.*



Proof. By the preceding theorem, g induces a continuous map $f: X^* \to Z$; it is clear that f is bijective. If Z is Hausdorff, then 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^*.

Suppose that f is a homeomorphism. Then both f and the projection $p: X \to X^*$ are quotient maps, so that $g = f \circ p$ is a quotient map. Conversely, suppose that g is a quotient map. Given an open set V of X^*, we compute

$$g^{-1}(f(V)) = p^{-1}(V),$$

which is open in X (because p is continuous). Therefore, $f(V)$ is open in Z, because g is a quotient map. Hence f maps open sets to open sets, so it is a homeomorphism. □

When dealing with quotient spaces, one must be careful not to rely too much on one's intuition as to what the spaces look like. Consider the following example:

EXAMPLE 7. Let X be the union of all straight lines in the plane of the form $L_n = R \times \{n\}$, for $n \in Z_+$. Let Z be the union of all straight lines in the plane passing through the origin and having positive integral slope. That is, let $Z = \bigcup L'_n$, where

$$L'_n = \{x \times (nx) \mid x \in R\}.$$

Both are subspaces of R^2. Let $g: X \to Z$ be the map

$$g(x \times n) = x \times (nx),$$

which carries each line L_n linearly onto L'_n. See Figure 27. We assert that the map g is not a quotient map.

§2-11

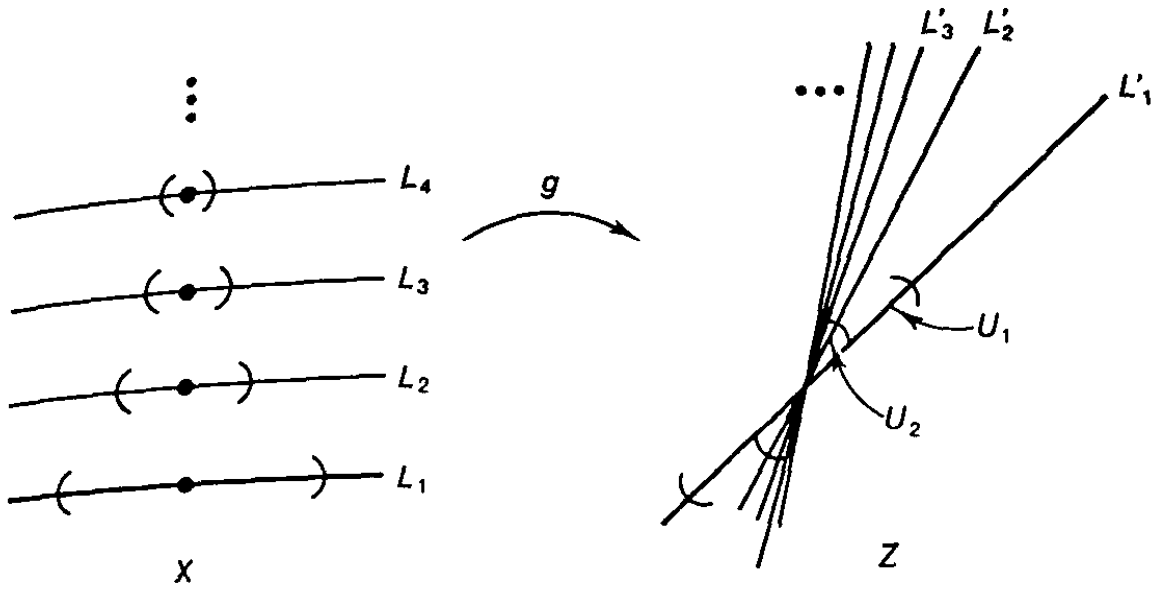


Figure 27

That is, we assert that the quotient topology on Z induced by the map g is not the same as the subspace topology on Z.

To prove this fact, let us take, for each n, an open interval U_n of length $1/n$ in the line L'_n, centered at the origin; and let us define $U = \bigcup U_n$. Then U is open in Z in the quotient topology induced by g, since the set $g^{-1}(U)$ is open in X. But U is not open in the subspace topology on Z; in that topology, the origin is a limit point of the complement $Z - U$ of U.

To say the same thing in a different way, suppose that we form a quotient space X^* from X by collapsing the set $0 \times Z_+$ to a single point. It is tempting to think of X^* as being essentially the same as the subspace Z of the plane. But it is not. The map g induces a bijective continuous map $h: X^* \longrightarrow Z$. But since g is not a quotient map, h is not a homeomorphism.

EXAMPLE 8. *The product of two quotient maps need not be a quotient map.* Let us take the spaces X and Z of the preceding example, but now *let us give Z the quotient topology induced by g.*

Consider the space R^ω (in the product topology). Let $i: R^\omega \longrightarrow R^\omega$ be the identity map; being a homeomorphism, it is a quotient map. We shall prove that the map

$$f = g \times i: X \times R^\omega \longrightarrow Z \times R^\omega$$

is not a quotient map. Recall that $X = R \times Z_+$.

Given $n \in Z_+$, define U_n to be the following subset of $(R \times Z_+) \times R^\omega$:

$$U_n = \{(t \times n) \times (x_1, x_2, \ldots); |tx_n| < 1\}.$$

The set U_n is open in $X \times R^\omega$, because it is the cartesian product of the open set

$$W = \{t \times x_n; |tx_n| < 1\}$$

in the tx_n plane, the open set $\{n\}$ in Z_+, and the open sets R in all the other coordinates. See Figure 28.

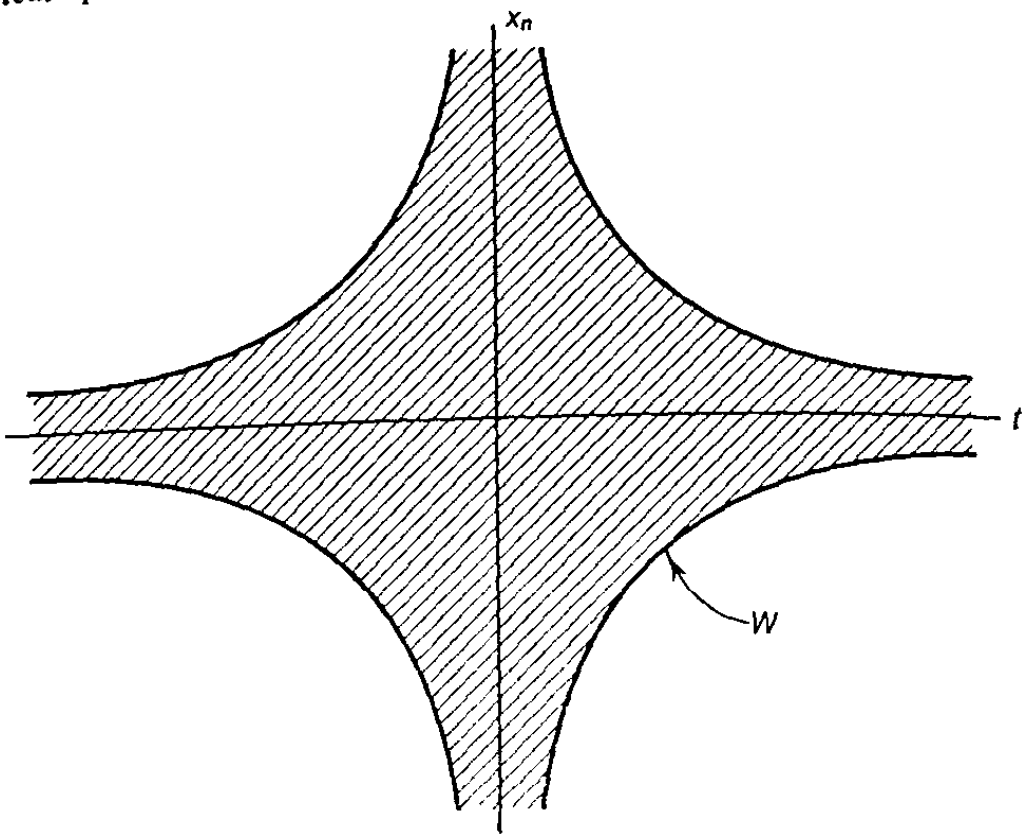


Figure 28

Define U to be the open set

$$U = \bigcup_{n \in Z_+} U_n$$

of $X \times R^\omega$. We assert the following:

(a) $f^{-1}(f(U)) = U$.

(b) $f(U)$ is not open in $Z \times R^\omega$.

It follows, since $f^{-1}(f(U))$ is open in $X \times R^\omega$ and $f(U)$ is not open in $Z \times R^\omega$, that the map f is not a quotient map.

(a) If $f^{-1}(f(U))$ is different from U, then there must exist two points $\mathbf{x} \in U$ and $\mathbf{y} \notin U$ such that $f(\mathbf{x}) = f(\mathbf{y})$. But in the present situation, $f(\mathbf{x}) = f(\mathbf{y})$ for $\mathbf{x} \neq \mathbf{y}$ only if

$$\mathbf{x} = (0 \times m) \times (x_1, x_2, \ldots),$$
$$\mathbf{y} = (0 \times n) \times (x_1, x_2, \ldots),$$

for some point $(x_1, x_2, \ldots)$ of R^ω. And these two points both belong to U, since $|0 \cdot x_m| < 1$ and $|0 \cdot x_n| < 1$.

(b) Suppose that $f(U)$ is open in $Z \times R^\omega$. The set $f(U)$ contains $\mathbf{0} \times \mathbf{0}$, so it contains a basis element $V \times \prod W_i$ about $\mathbf{0} \times \mathbf{0}$. Here V is open in Z, and W_i is open in R; and $W_i = R$ for all but finitely many i. Choose N so that $W_N = R$. Since $f(U)$ contains $V \times \prod W_i$, the set $U = f^{-1}(f(U))$ contains $f^{-1}(V \times \prod W_i) = g^{-1}(V) \times \prod W_i$. Now $g^{-1}(V)$ is an open set in X containing the entire set $0 \times Z_+$. In particular, $g^{-1}(V)$ contains some point $t_0 \times N$ for $t_0 \neq 0$. Choose $a_N > 1/|t_0|$, and consider the point

$$(t_0 \times N) \times (0, \ldots, 0, a_N, 0, \ldots)$$

§2-11

of $X \times R^\omega$. Because $W_N = R$, this point lies in $g^{-1}(V) \times \prod W_i$. Because $|t_0 a_N|$ is *not* less than 1, it does not lie in U. But U contains $g^{-1}(V) \times \prod W_i$. Contradiction.

Exercises

1. Check the details of Example 3.
2. Let $\pi_1 : R \times R \to R$ be projection on the first coordinate.
   (a) Let X be the subspace $(0 \times R) \cup (R \times 0)$ of $R \times R$; let $g = \pi_1 | X$. Show that g is a closed map but not an open map.
   (b) Let Y be the subspace $(\bar{R}_+ \times R) \cup (R \times 0)$ of $R \times R$; let $h = \pi_1 | Y$. Show that h is neither open nor closed, but it is a quotient map. [*Hint:* $h^{-1}(U) \cap (R \times 0) = U \times 0$.]
3. Let $p : X \to Y$ be surjective and continuous; let A be a subspace of X. Show that if A is open in X and p is an open map, then $p|A$ is an open map; conclude that $p' : A \to p(A)$ is an open map. Show that if A and p are closed, $p|A$ is closed, and so is p'.
4. Let X be a space; let A be a set; let $p : X \to A$ be a surjective map. Show that the quotient topology on A induced by p is the finest (largest) topology relative to which p is continuous.
5. (a) Define an equivalence relation on the plane X 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 collection of equivalence classes, in the quotient topology. X^* is homeomorphic to a familiar space; what is it?
   (b) Repeat (a) for the equivalence relation

$$x_0 \times y_0 \sim x_1 \times y_1 \quad \text{if } x_0^2 + y_0^2 = x_1^2 + y_1^2.$$

6. (a) Define $g : R^2 \to R$ by the equation $g(x \times y) = x + y^2$. Show that g is a quotient map.
   (b) Define $g : R^2 \to \bar{R}_+$ by the equation $g(x \times y) = x^2 + y^2$. Show that g is a quotient map.
   (c) Compare (a) and (b) with Exercise 5.
7. Let Z be the subspace $(R \times 0) \cup (0 \times R)$ of R^2. Define $g : R^2 \to Z$ by the equations

$$g(x \times y) = x \times 0 \quad \text{if } x \neq 0,$$
$$g(0 \times y) = 0 \times y.$$

   (a) Is g a quotient map? Is g continuous?
   (b) Show that in the quotient topology induced by g, the space Z is not Hausdorff.
8. Let C_n be the subspace of R^2 defined by

$$C_n = \{x \times y | (x - 1/n)^2 + y^2 = (1/n)^2\}.$$

Let Y be the subspace

$$Y = \bigcup_{n \in Z_+} C_n$$

of R^2, and let X be the subspace $C_1 \times Z_+$ of $R^2 \times R$. Define $g : X \longrightarrow Y$ by the equation

$$g((x \times y) \times n) = (x/n, y/n).$$

Show that g is continuous and surjective, but g is not a quotient map.

9. Let $\{X_\alpha\}$ be a family of spaces and let $\{p_\alpha\}$ be a family of maps $p_\alpha : X_\alpha \longrightarrow A$, where A is a set.

(a) Show there is a unique finest topology $\mathfrak{I}$ on A relative to which each map p_α is continuous.

(b) Show that a map $f: A \longrightarrow Y$ is continuous relative to $\mathfrak{I}$ if and only if each map $f \circ p_\alpha$ is continuous.

*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 requirements 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.

1. Show that G is a topological group if the map of $G \times G$ into G sending $x \times y$ into $x^{-1} \cdot y$ is continuous, and conversely.

2. Show that the following are topological groups:

(a) $(Z, +)$

(b) $(R, +)$

(c) $(R_+, \cdot)$

(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* $\mathrm{GL}(n)$, under the operation of matrix multiplication. ($\mathrm{GL}(n)$ is the set of all nonsingular n by n matrices, topologized by considering it as a subset of $R^{(n^2)}$ in the obvious way.)

(f) $(R^J, +)$ in the product, uniform, and box topologies.

3. Let $(G, \cdot)$ be a topological group; let α be an element of G. Show that the maps $f_\alpha, g_\alpha : G \longrightarrow G$ 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.)

4. Let H be a subgroup of G. If $x \in G$, define $xH = \{x \cdot h | h \in H\}$; 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. If G is a topological group, 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 the quotient map $p : G \to G/H$ is open.
   (c) Show that if H is a closed set in the topology of G, then one-point sets are closed in G/H.
   (d) Show that if H is a normal subgroup of G, then G/H is a topological group.

5. $(Z, +)$ is a normal subgroup of $(R, +)$. The quotient R/Z is a familiar topological group; what is it?

6. Let G be a topological group with identity element e. 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) If U is a neighborhood of e, show there is a neighborhood V of e such that $V \cdot V^{-1} \subset U$.
   (b) If A is a closed set not containing e, show there is a neighborhood V of e such that the open sets V and
   $$A \cdot V = \bigcup_{\alpha \in A} f_\alpha(V)$$
   are disjoint. (f_α is the map defined in Exercise 3.)
   (c) Suppose that one-point sets are closed in G. Show that G is Hausdorff. In fact, show that G satisfies the following stronger separation axiom, which is called the **regularity axiom**: Given a closed set C and a point x not in C, there exist disjoint open sets containing C and x, respectively.

7. 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:* Consider first the case where A is a closed set in G/H not containing $p(e)$. Let $B = p^{-1}(A)$. Find a neighborhood U of e disjoint from B, and let $V^{-1} \cdot V \subset U$. Then consider the sets $V \cdot H$ and $V \cdot B$.]

8. Show that if H is a subgroup of G, so is the closure of H.

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 $f: [a, b] \longrightarrow R$ 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 $f: [a, b] \longrightarrow R$ 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 $f: [a, b] \longrightarrow R$ 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

§3-1

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.

3-1 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 which 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 1.1. 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 $A \cap \bar{B} = \varnothing$; whence 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 R. Each of the sets $[-1, 0)$ and $(0, 1]$ is nonempty and open in Y (although not in 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 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 Q are not connected. Indeed, the only connected subspaces of Q are the one-point sets: If Y is a subspace of 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 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 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

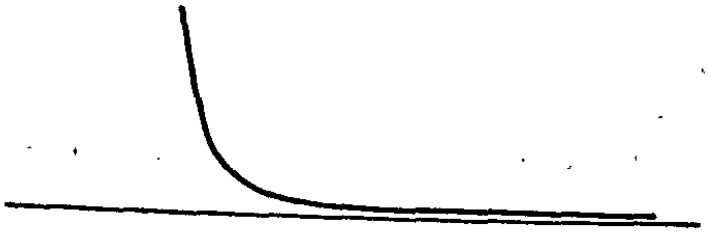


Figure 1

§3-1

theorems that tell how to form new connected spaces from given ones. Then in the next section we shall apply these theorems to construct some specific connected spaces, such as intervals in R, and balls and cubes in R^n. First, a lemma:

Lemma 1.2. If the sets C *and* D *form a separation of* X, *and if* Y *is a connected subset 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 1.3. The union of a collection of connected sets that have a point in common is connected.

Proof. Let $\{A_\alpha\}$ be a collection of connected subsets of a space X; let p be a point of $\bigcap A_\alpha$. We prove that the set $Y = \bigcup A_\alpha$ 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 the set 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; whence $\bigcup A_\alpha \subset C$, contradicting the fact that D is nonempty. □

Theorem 1.4. Let A *be a connected subset 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 set 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 1.2, 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 1.5. The image of a connected space under a continuous map is connected.

Proof. Let $f: X \longrightarrow Y$ be a continuous map; let X be connected. We wish to prove the image set $Z = f(X)$ 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

$$g: X \longrightarrow Z.$$

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 1.6. *The 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 sets that have the point $x \times b$ in common. See Figure 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 sets that have the point $a \times b$ in common. Since this union equals $X \times Y$, the space $X \times Y$ is connected.

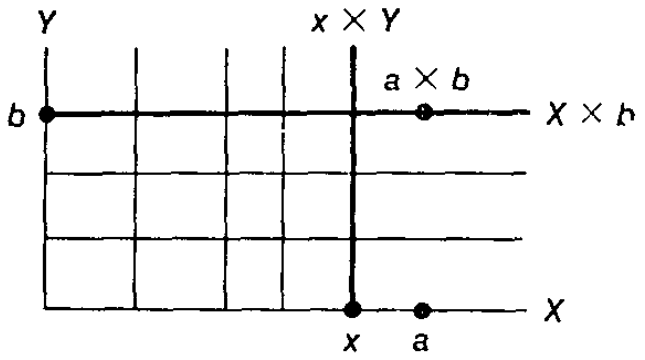


Figure 2

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 $(X_1 \times \cdots \times X_{n-1}) \times X_n$.

The proof for an arbitrary product now proceeds as follows. Let $\{X_\alpha\}_{\alpha \in J}$ be an indexed family of connected spaces, and let

$$X = \prod_{\alpha \in J} X_\alpha.$$

Choose a "base point" $\mathbf{b} = (b_\alpha)_{\alpha \in J}$ for X.

Given any finite set $\{\alpha_1, \ldots, \alpha_n\}$ of indices in J, let us define a subspace $X(\alpha_1, \ldots, \alpha_n)$ of X; it consists of all points $(x_\alpha)_{\alpha \in J}$ such that $x_\alpha = b_\alpha$ for $\alpha \neq \alpha_1, \ldots, \alpha_n$. We assert that $X(\alpha_1, \ldots, \alpha_n)$ is homeomorphic with the finite product

$$X_{\alpha_1} \times \cdots \times X_{\alpha_n},$$

and hence is connected: For the obvious bijection between these spaces, mapping

$$(x_{\alpha_1}, \ldots, x_{\alpha_n}) \longrightarrow (y_\alpha)_{\alpha \in J},$$

§3-1

where $y_\alpha = x_\alpha$ for $\alpha = \alpha_1, \ldots, \alpha_n$ and $y_\alpha = b_\alpha$ for all other values of α, carries a basis element for $X_{\alpha_1} \times \cdots \times X_{\alpha_n}$ to a basis element for $X(\alpha_1, \ldots, \alpha_n)$.

It follows that the subspace Y of X which is the *union* of these subspaces is connected; this is the subspace

$$Y = \bigcup X(\alpha_1, \ldots, \alpha_n),$$

where the union extends over all finite subsets $\{\alpha_1, \ldots, \alpha_n\}$ of J. For the spaces $X(\alpha_1, \ldots, \alpha_n)$ are connected, and they all contain the base point $\mathbf{b} = (b_\alpha)$.

It is tempting to think that the proof is finished. But it is not, for Y is not all of X. The space Y consists of all points $(x_\alpha)_{\alpha \in J}$ of X having the property that $x_\alpha = b_\alpha$ for all but finitely many values of α. So in some sense Y is only a very small part of X. But now the fact that we are using the product topology for X comes in. We assert that under the product topology for X, the closure of Y equals all of X. Once we prove this fact, the connectedness of X follows from Theorem 1.4.

Let us take an arbitrary point (x_α) of X, and an arbitrary basis element $U = \prod U_\alpha$ about (x_α), and prove that U intersects Y. Each set U_α is open in X_α, and $U_\alpha = X_\alpha$ except for finitely many indices, say $\alpha = \alpha_1, \ldots, \alpha_n$. Construct a point (y_α) of X by setting

$$y_\alpha = \begin{cases} x_\alpha & \text{for } \alpha = \alpha_1, \ldots, \alpha_n, \\ b_\alpha & \text{for all other values of } \alpha. \end{cases}$$

Then (y_α) is a point of Y, because it belongs to the space $X(\alpha_1, \ldots, \alpha_n)$. Also (y_α) is a point of U, because $y_\alpha = x_\alpha \in U_\alpha$ for $\alpha = \alpha_1, \ldots, \alpha_n$ and $y_\alpha = b_\alpha \in X_\alpha$ for all other values of α. Hence U intersects Y, as desired. □

This theorem is not true if one uses the box topology on X. See Exercise 11.

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. Show that if $\prod X_\alpha$ is connected and nonempty, then each X_α is connected.
3. Let $\{A_n\}$ be a sequence of connected subsets of X, such that $A_n \cap A_{n+1} \neq \varnothing$ for all n. Show that $\bigcup A_n$ is connected.
4. Let $\{A_\alpha\}$ be a collection of connected subsets of X; let A be a connected subset of X. Show that if $A \cap A_\alpha \neq \varnothing$ for all α, then $A \cup (\bigcup A_\alpha)$ is connected.
5. Show that if X is an infinite set, it is connected in the finite complement topology $\mathcal{T}_f = \{A \mid X - A \text{ is finite or all of } X\}$.

6. A space is **totally disconnected** if its only connected subsets are one-point sets. Show that a finite Hausdorff space is totally disconnected.

7. Is it true that if X has the discrete topology, then X is totally disconnected? Does the converse hold?

8. Is it true that if X is connected, then for every nonempty proper subset A of X, we have $\text{Bd } A \neq \emptyset$? Does the converse hold? (Recall that $\text{Bd } A = \bar{A} \cap \overline{X - A}$.)

9. Let $A \subset X$. Show that if C is a connected subset of X that intersects both A and $X - A$, then C intersects $\text{Bd } A$.

10. Is the space R_l connected? Justify your answer.

11. Show that R^ω is not connected in the box topology by showing that the set A of all bounded sequences is both open and closed. What happens if R^ω has the uniform topology?

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.

3-2 Connected Sets in 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 every interval and every ray in R is connected.

One application is the intermediate value theorem of calculus, suitably generalized. As another application, we show that such familiar spaces as balls and spheres in euclidean space are connected. The proof of this fact leads to a new notion, called *path connectivity*, which we also discuss.

The fact that intervals and rays in 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 R, 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 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 2.1. *If* L *is a linear continuum in the order topology, then* L *is connected and so is every interval and ray in* L.

Proof. Let Y be a subset of L that equals either L or an interval or a ray in L. The set Y is "convex," in the sense that if a and b are any two points of

§3-2

Y and $a < b$, then the entire interval $[a, b]$ of points of L is contained in the set Y.

Let A and B be two disjoint nonempty sets that are open in Y. We shall show that $Y \neq A \cup B$, thereby demonstrating that there exists no separation of Y.

Choose a point a of A and a point b of B; assume the notation so chosen that $a < b$. (Do not make the mistake of assuming that *every* point of A is less than *every* point of B; that need not be true.) Because Y is "convex," we have $[a, b] \subset Y$. We shall find a point of $[a, b]$ that belongs to neither A nor B.

Consider the sets

$$A_0 = A \cap [a, b] \quad \text{and} \quad B_0 = B \cap [a, b].$$

These sets are open in $[a, b]$ in the subspace topology (which is the same as the order topology). Let

$$c = \text{lub } A_0.$$

We shall show that c belongs to neither A_0 nor 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. See Figure 3. 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.

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 4. 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. □

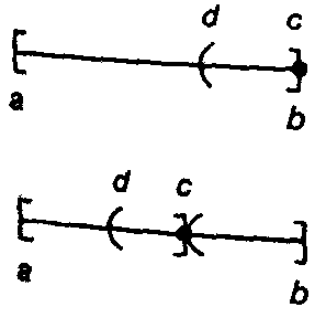


Figure 3

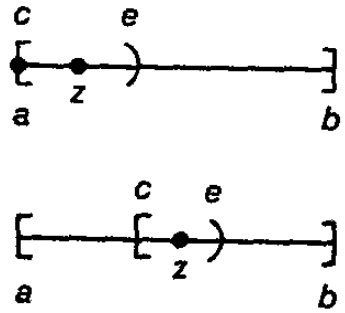


Figure 4

Corollary 2.2. The real line R *is connected and so is every interval and ray in* R.

As an application, we prove the intermediate value theorem of calculus, suitably generalized.

Theorem 2.3 (Intermediate value theorem). *Let* $f : X \to Y$ *be a continuous map of the connected space* X *into the ordered set* Y, *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 R and Y to be 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 R is the space $I \times I$ in the dictionary order topology, where $I = [0, 1]$. 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 $b = \text{lub}\, \pi_1(A)$. 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 5. If $b \notin \pi_1(A)$, then $b \times 0$ is the least upper bound of A; no element

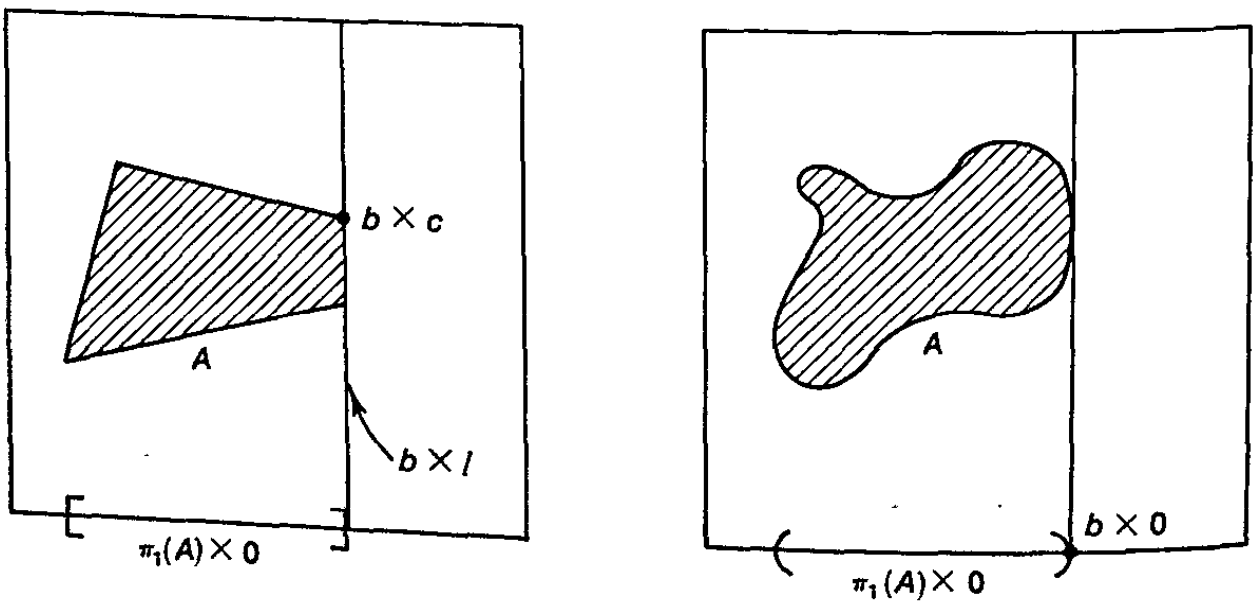


Figure 5

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 $\pi_1(A)$.

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 $(0, 1)$ between each element of X and its immediate successor, provided X has no largest element.

Connectedness of intervals in 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 necessarily 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 below.

EXAMPLE 3. Define the unit ball B^n in R^n by the equation

$$B^n = \{\mathbf{x}; \|\mathbf{x}\| \leq 1\},$$

where

$$\|\mathbf{x}\| = \|(x_1, \ldots, x_n)\| = (x_1^2 + \cdots + x_n^2)^{1/2}.$$

The unit ball is path connected; given any two points $\mathbf{x}$ and $\mathbf{y}$ of B^n, the straight-line path $f: [0, 1] \to R^n$ 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 $[0, 1]$,

$$\|f(t)\| \leq (1 - t)\|\mathbf{x}\| + t\|\mathbf{y}\| \leq 1.$$

A similar argument shows that every open ball $B_d(\mathbf{x}, \epsilon)$ and every closed ball $\bar{B}_d(\mathbf{x}, \epsilon)$ in R^n is path connected.

EXAMPLE 4. Define **punctured euclidean space** to be the space $R^n - \{\mathbf{0}\}$, where $\mathbf{0}$ is the origin in 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 R^n by the equation

$$S^{n-1} = \{\mathbf{x}; \|\mathbf{x}\| = 1\}.$$

If $n > 1$, it is path connected. For the map $g: R^n - \{0\} \longrightarrow S^{n-1}$ 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 space $I \times I$ in the dictionary order topology is connected but not path connected.*

Being a linear continuum, $I \times I$ is connected. Let $p = 0 \times 0$ and $q = 1 \times 1$. We suppose there is a path $f: [a, b] \longrightarrow I \times I$ joining p and q and derive a contradiction. The image set $f([a, b])$ must contain every point $x \times y$ of $I \times I$, 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 6. Choose, for each $x \in I$, a rational number q_x belonging to U_x. Since the sets

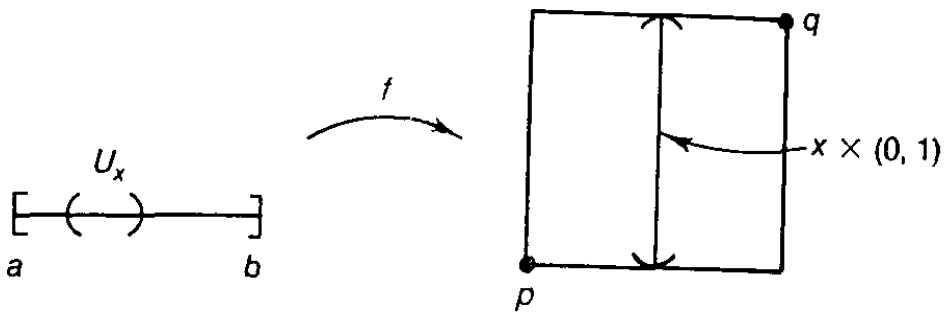


Figure 6

U_x are disjoint, the map $x \longrightarrow q_x$ is an injective mapping of I into Q. This contradicts the fact that the interval I is uncountable (which we shall prove later).

EXAMPLE 7. Here is another connected space that is not path connected; it is a subspace of the plane. Let K denote the set $\{1/n \mid n \in Z_+\}$, and define

$$C = ([0, 1] \times 0) \cup (K \times [0, 1]) \cup (0 \times [0, 1]).$$

The space C is called the **comb space**, for obvious reasons. See Figure 7. The space D obtained by deleting from C the points of the vertical interval $0 \times (0, 1)$ is called the **deleted comb space**.

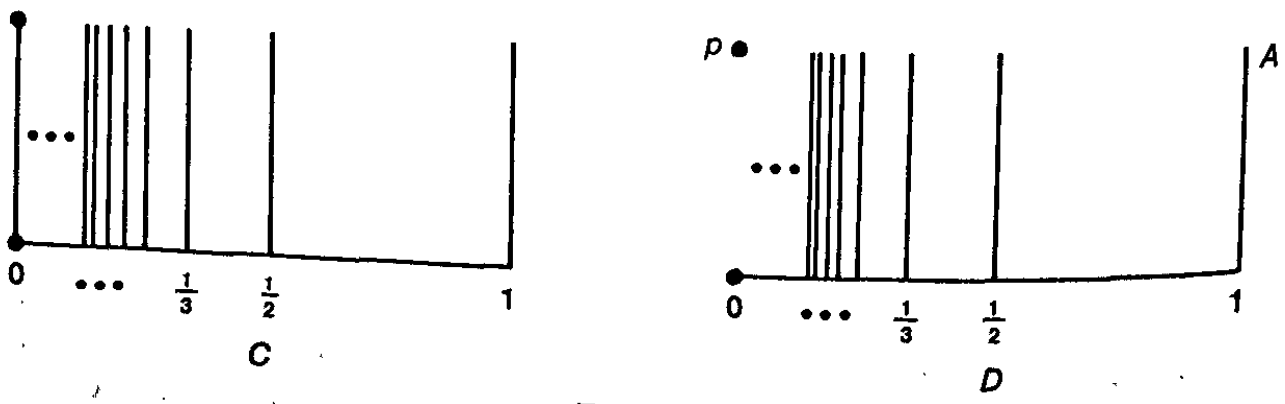


Figure 7

§ 3-2

The comb space C is clearly path connected. The deleted comb space D is readily seen to be *connected*, for it is the union of the path connected set

$$A = ([0, 1] \times 0) \cup (K \times [0, 1])$$

and the limit point $p = 0 \times 1$ of A. We shall prove that D is not path connected.

Suppose that $f: [a, b] \longrightarrow D$ is a path in D beginning at p. We assert that the set $f^{-1}(\{p\})$ is both open and closed in $[a, b]$; from this it follows by connectivity that $f^{-1}(\{p\}) = [a, b]$. We conclude that there exists no path in D joining p to a point of A.

To show that $f^{-1}(\{p\})$ is closed is trivial, since $\{p\}$ is closed and f is continuous. To show that $f^{-1}(\{p\})$ is open, we choose a neighborhood V of p in R^2 which does not intersect the x-axis, as illustrated in Figure 8. Then, given an

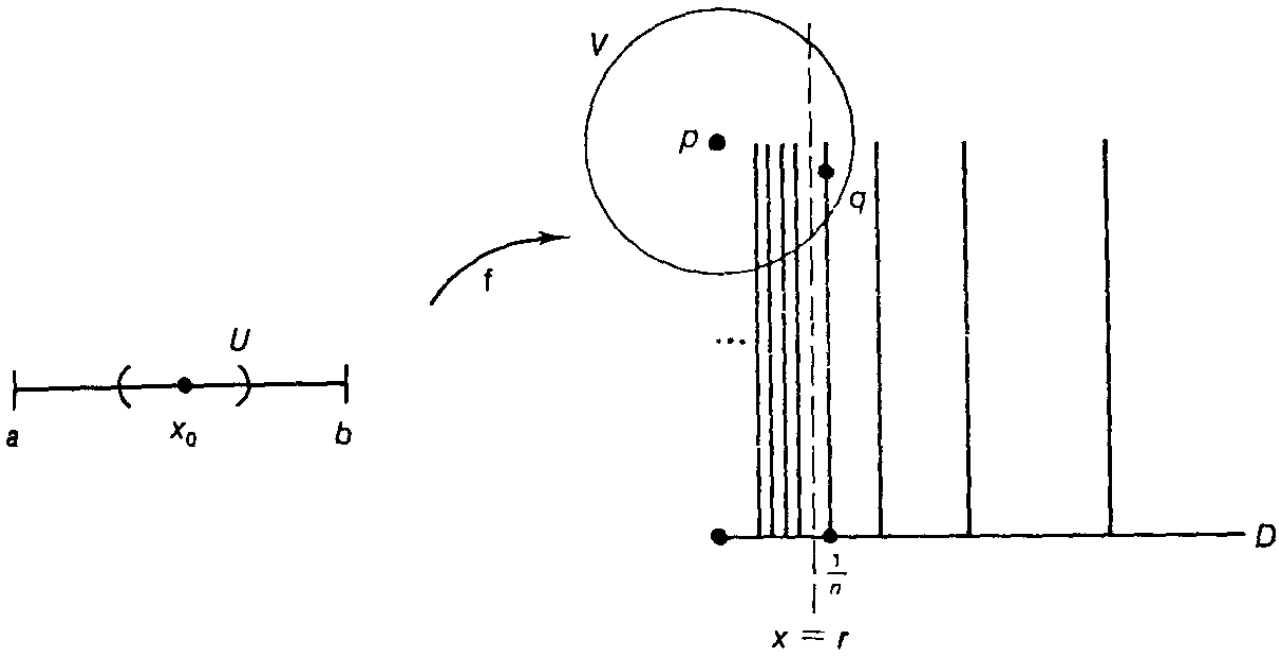


Figure 8

arbitrary point x_0 of $f^{-1}(\{p\})$, we choose a basis element U about x_0 such that $f(U) \subset V$. We assert that U is contained in $f^{-1}(\{p\})$, so $f^{-1}(\{p\})$ is open.

Note that U is connected, being a basis element for the order topology on $[a, b]$. Therefore, $f(U)$ is connected. Then $f(U)$ cannot contain any point different from p: Given a point $q = (1/n) \times t_0$ of D different from p and lying in V, choose r so that $1/(n + 1) < r < 1/n$. Then consider the following two disjoint open subsets of R^2:

$$(-\infty, r) \times R \quad \text{and} \quad (r, +\infty) \times R.$$

Because $f(U)$ lies in D and does not touch the x-axis, it does not intersect the line $x = r$; therefore, it lies in the union of these two sets. Since $f(U)$ is connected and contains the point p of the first set, it cannot contain the point q of the second set. Thus $f(U) = \{p\}$, as asserted.

EXAMPLE 8. Let S denote the following subset of the plane:

$$S = \{x \times \sin(1/x) \mid 0 < x \leq 1\}.$$

Because S is the image of the connected set $(0, 1]$ under a continuous map, S is connected. Therefore, its closure $\bar{S}$ in R^2 is also connected. The set $\bar{S}$ is a classical example in topology called the **topologist's sine curve**. It is illustrated in Figure 9. The topologist's sine curve is connected but not path connected; the proof is similar to the proof we gave for the deleted comb space.

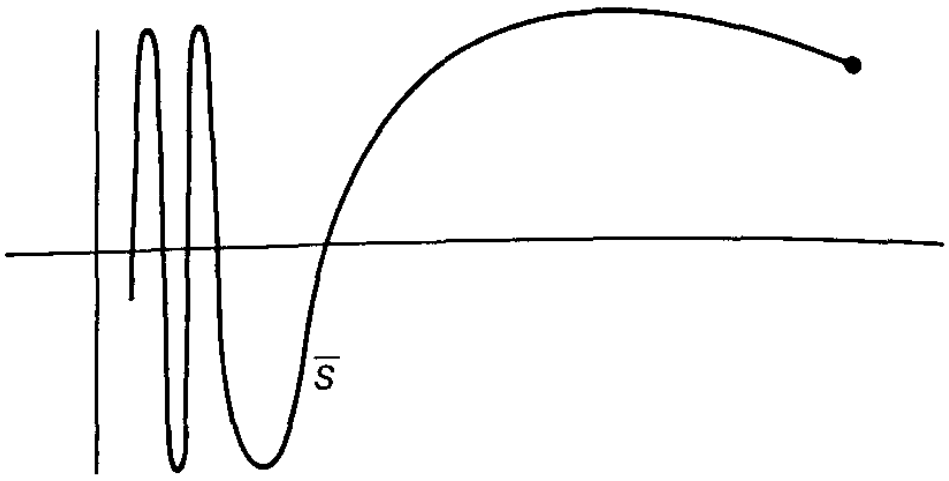


Figure 9

Exercises

1. (a) Show that no two of the spaces $(0, 1)$, $(0, 1]$, and $[0, 1]$ are homeomorphic.
   (b) Suppose that $f: X \longrightarrow Y$ and $g: Y \longrightarrow X$ are imbeddings. Show by means of an example that X and Y need not be homeomorphic.
2. Let n be a positive integer. Given that the function $f: R \longrightarrow R$ defined by $f(x) = x^n$ is continuous, show that for each $a \geq 0$ there is exactly one $b \geq 0$ such that $b^n = a$.
3. Let $f: X \longrightarrow X$ 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) $Z_+ \times [0, 1)$
   (b) $[0, 1) \times Z_+$
   (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) 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 $f: X \longrightarrow Y$ is continuous and X is path connected, is $f(X)$ necessarily path connected?

§3-3

(d) If $\{A_\alpha\}$ is a collection of path-connected subsets of X and if $\bigcap A_\alpha \neq \emptyset$, is $\bigcup A_\alpha$ necessarily path connected?

8. Show that R^n and R are not homeomorphic if $n > 1$.

9. Assume that R is uncountable. Show that if A is a countable subset of R^2, then $R^2 - A$ is path connected. [*Hint:* How many lines are there passing through a given point of R^2?]

10. Show that if U is an *open* connected subset of 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 subset of X, are Int A and Bd A necessarily connected? Does the converse hold? Justify your answers.

12. Show that the topologist's sine curve is connected but not path connected.

13. 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 R, but it cannot be imbedded in R.

Proof.

(a) Let X be an ordered set with smallest element a_0. Let

$$A = \{a_0\} \cup \{x \mid [a_0, x) \text{ has the order type of } [0, 1)\}.$$

Show that if for each $x \in X$ there is a $y > x$ such that $[x, y)$ has the order type of $[0, 1)$, then A is open in X. Show that if X satisfies the "sequence lemma," then A is closed in X. [*Hint:* If a is a limit point of A not in A, there is an increasing sequence of points of A converging to a.]

(b) Show that L is locally homeomorphic to R; that is, each point of L has a neighborhood homeomorphic to an open subset of R.

(c) Show that L is path connected.

(d) Show that L cannot be imbedded in R, or indeed in R^n for any n. [*Hint:* Any subspace Y of R^n has a countable basis for its topology; therefore, any collection of disjoint open sets in Y is countable.]

*3-3 Components and Path Components†

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 subset of X containing both x and y. The equivalence classes are called the **components** (or "connected components") of X.

Symmetry and reflexivity of the relation are obvious. Transitivity follows

†This section will be assumed in Chapter 8.

by noting that if A is a connected set containing x and y, and if B is a connected set containing y and z, then $A \cup B$ is a set containing x and z, which is connected because A and B have the point y in common.

The components of X can also be described as follows:

Theorem 3.1. The components of X *are connected disjoint subsets of* X *whose union is* X, *such that each connected subset of* X *intersects only one of them.*

Proof. Being equivalence classes, the components of X are disjoint and their union is X. Each connected set A in 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 $C_1 = C_2$.

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 set A_x containing x_0 and x. By the result of the preceding paragraph, $A_x \subset C$. Therefore,

$$C = \bigcup_{x \in C} A_x.$$

Since the sets 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 (because any two closed intervals $[a, b]$ and $[c, d]$ in R are homeomorphic) there is also a path g from x to y having $[c, d]$ as its domain. 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 7.3 of Chapter 2.

One has the following theorem, whose proof is immediate:

Theorem 3.2. The path components of X *are path-connected disjoint subsets of* X *whose union is* X, *such that each path-connected subset of* X *intersects only one of them.*

EXAMPLE 1. The components of the subspace

$$Y = [-1, 0) \cup (0, 1]$$

§3-4

of the real line R are the two sets $[-1, 0)$ and $(0, 1]$. These are also the path components of Y.

EXAMPLE 2. The deleted comb space D of the previous section is a space having a single component (because it is connected) and two path components. If we form a space Y by adjoining to the deleted comb space D all the irrational points of the interval $0 \times [0, 1]$, we obtain a space having only one component but uncountably many path components.

*3-4 Local Connectedness†

Connectivity is a useful property for a space to possess. But for some purposes, it is more important that the space satisfy a connectivity condition *locally*. Roughly speaking, local connectivity 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**.

To phrase it differently, X is locally connected if there is a basis for X consisting of connected sets. Local connectedness and connectedness of a space are not related to one another; a space may possess one or both of these properties, or neither.

Definition. 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 1. Each interval and each ray in the real line is both connected and locally connected. The subspace $[-1, 0) \cup (0, 1]$ of R is not connected, but it is locally connected. The deleted comb space (Example 7 of §3-2) is connected but not locally connected. The rationals Q are neither connected nor locally connected.

EXAMPLE 2. R^n is locally path connected, since each of the basis elements $(a_1, b_1) \times \cdots \times (a_n, b_n)$ is path connected. Similarly, R^ω is locally path connected; each of its standard basis elements is path connected.

†This section will be assumed in Chapter 8.

EXAMPLE 3. If we form a space Y by adjoining to the deleted comb space D all the points of the form $0 \times (1/n)$ for $n \in Z_+$, we obtain a space that is locally connected at the origin but not locally path connected at the origin.

The most important facts about locally connected spaces are given in the following theorems:

Theorem 4.1. *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 4.2. *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 4.3. *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 $P \neq C$. 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. □

§34

Exercises

1. What are the components and path components of R_l?
2. (a) What are the components and path components of R^ω (in the product topology)?
   (b) Consider R^ω in the uniform topology. Show that $\mathbf{x}$ and $\mathbf{y}$ lie in the same component of R^ω if and only if the sequence
   $$\mathbf{x} - \mathbf{y} = (x_1 - y_1, x_2 - y_2, \ldots)$$
   is bounded.
   (c) Consider R^ω in the box topology. Show that $\mathbf{x}$ and $\mathbf{y}$ lie in the same component of R^ω if the sequence $\mathbf{x} - \mathbf{y}$ is "eventually zero."
   *(d) Prove the converse of (c).
3. In a space X, 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 $y \in B$. Show this is an equivalence relation. We shall call the equivalence classes the **quasicomponents** of X. Show that each component of X lies in a quasicomponent of X.
4. Determine the quasicomponents, components, and path components of the following subspaces of R^2. [Here K denotes the set $\{1/n \mid n \in Z_+\}$ and $-K$ denotes the set $\{-1/n \mid n \in Z_+\}$.]
   $A = (K \times [0, 1]) \cup (0 \times [0, 1])$,
   $B = A - \{0 \times \frac{1}{2}\}$,
   $C = B \cup ([0, 1] \times 0)$,
   $D = (K \times [0, 1]) \cup (-K \times [-1, 0]) \cup ([0, 1] \times -K) \cup ([-1, 0] \times K)$.
5. Show that if X is locally connected, then the quasicomponents of X are the same as the components of X.
6. If $f: X \to Y$ is continuous and X is locally connected, is $f(X)$ necessarily locally connected? What if f is both continuous and open?
7. Show that $I \times I$ in the dictionary order topology is locally connected but not locally path connected. What are the path components of this space?
8. Let X be locally path connected. Show that every connected open set in X is path connected.
9. Let X denote the rational points of the interval $[0, 1] \times 0$ of 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 R^2 that is path connected but is locally connected at none of its points.
10. A space X is said to be **connected im kleinen at x** if for every neighborhood U of x, there is a connected subset A of U that contains a neighborhood of x. Show that if X is connected im kleinen at each of its points, then X is locally connected. [*Hint:* Show that components of open sets are open.]

11. Consider the "infinite broom" X pictured in Figure 10. Show that X is not locally connected at p, but X is connected im kleinen at p. [*Hint:* Any connected neighborhood of p must contain all the points a_i.]

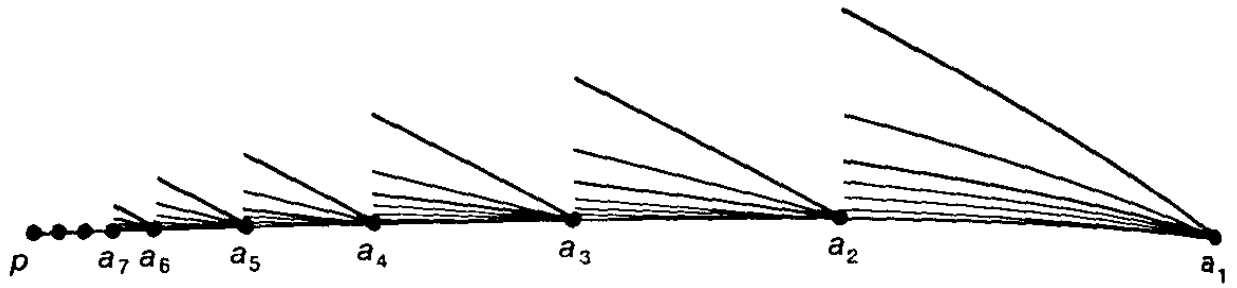


Figure 10

3-5 *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 R is not compact, for the covering of R by open intervals

$$\mathcal{A} = \{(n, n+2) \mid n \in Z\}$$

contains no finite subcollection that covers R.

EXAMPLE 2. The following subspace of R is compact:

$$X = \{0\} \cup \{1/n \mid n \in Z_+\}.$$

Given an open covering $\mathcal{A}$ of X, there is an element U of $\mathcal{A}$ containing 0. The

§3-5

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

$$\mathcal{A} = \{(1/n, 1] \mid n \in Z_+\}$$

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 given 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 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 5.1. 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

$$\{A_{\alpha_1} \cap Y, \ldots, A_{\alpha_n} \cap Y\}$$

covers Y. Then $\{A_{\alpha_1}, \ldots, A_{\alpha_n}\}$ 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 $\{A_{\alpha_1}, \ldots, A_{\alpha_n}\}$ covers Y. Then $\{A'_{\alpha_1}, \ldots, A'_{\alpha_n}\}$ is a subcollection of $\mathcal{A}'$ that covers Y. □

Theorem 5.2. *Every closed subset of a compact space is compact.*

Proof. Let Y be a closed subset 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 $X - Y$,

$$\mathcal{B} = \mathcal{A} \cup \{X - Y\}.$$

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 5.3. *Every compact subset of a Hausdorff space is closed.*

Proof. Let Y be a compact subset 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 $X - Y$. 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 $V_{y_1}, \ldots, V_{y_n}$ cover Y. The open set

$$V = V_{y_1} \cup \cdots \cup V_{y_n}$$

contains Y, and it is disjoint from the open set

$$U = U_{y_1} \cap \cdots \cap U_{y_n}$$

formed by taking the intersection of the corresponding neighborhoods of x_0. See Figure 11. For if z is a point of V, then $z \in V_{y_i}$ for some i, whence $z \notin U_{y_i}$ and so $z \notin U$.

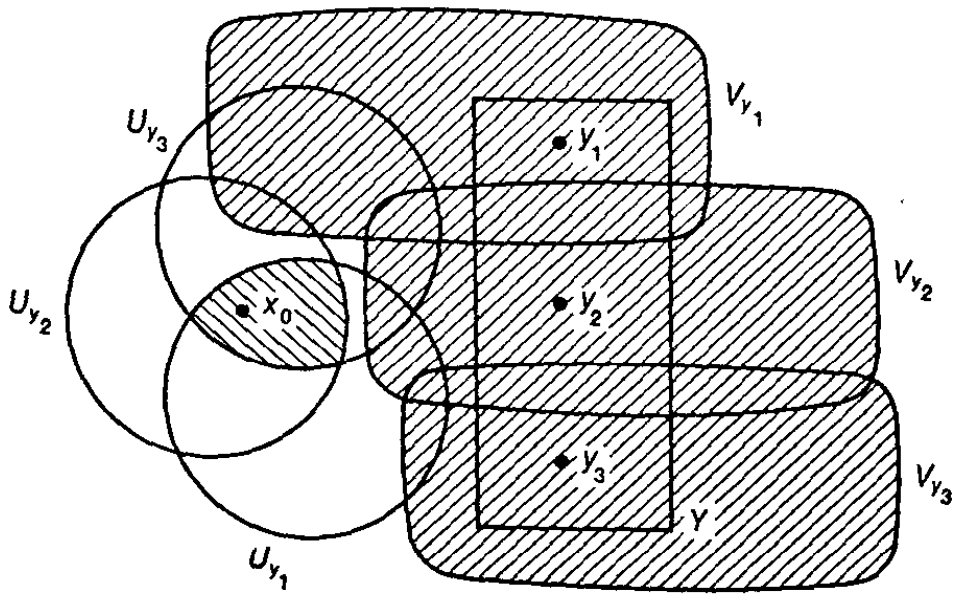


Figure 11

Therefore, U is a neighborhood of x_0 disjoint from Y. Hence $X - Y$ is open, as desired. □

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:

§3-5

Lemma 5.4. If Y *is a compact subset 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 R is compact, it follows from Theorem 5.2 that any closed subset of $[a, b]$ is compact. On the other hand, it follows from Theorem 5.3 that the intervals $(a, b]$ and (a, b) in R cannot be compact (which we knew already) because they are not closed in the Hausdorff space R.

EXAMPLE 6. One needs the Hausdorff condition in the hypothesis of Theorem 5.3. Consider, for example, the topology $\mathfrak{I}_f$ on the real line consisting of $\emptyset$ and all complements of finite sets. The only proper subsets of R that are closed in this topology are the finite sets. But *every* subset of R is compact in this topology, as you can check.

Theorem 5.5. The image of a compact space under a continuous map is compact.

Proof. Let $f: X \longrightarrow Y$ 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

$$f^{-1}(A_1), \ldots, f^{-1}(A_n)$$

cover X. Then the sets $A_1, \ldots, A_n$ cover $f(X)$. □

One important use of the preceding theorem is as a tool for constructing homeomorphisms:

Theorem 5.6. Let $f: X \longrightarrow Y$ *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 5.2. Therefore, by the theorem just proved, $f(A)$ is compact. Since Y is Hausdorff, $f(A)$ is closed in Y, by Theorem 5.3. □

Theorem 5.7. 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 × 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. Hence we can cover $x_0 \times Y$ by finitely many such basis elements

$$U_1 \times V_1, \ldots, U_n \times V_n.$$

(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 $W \times Y$. See Figure 12. 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

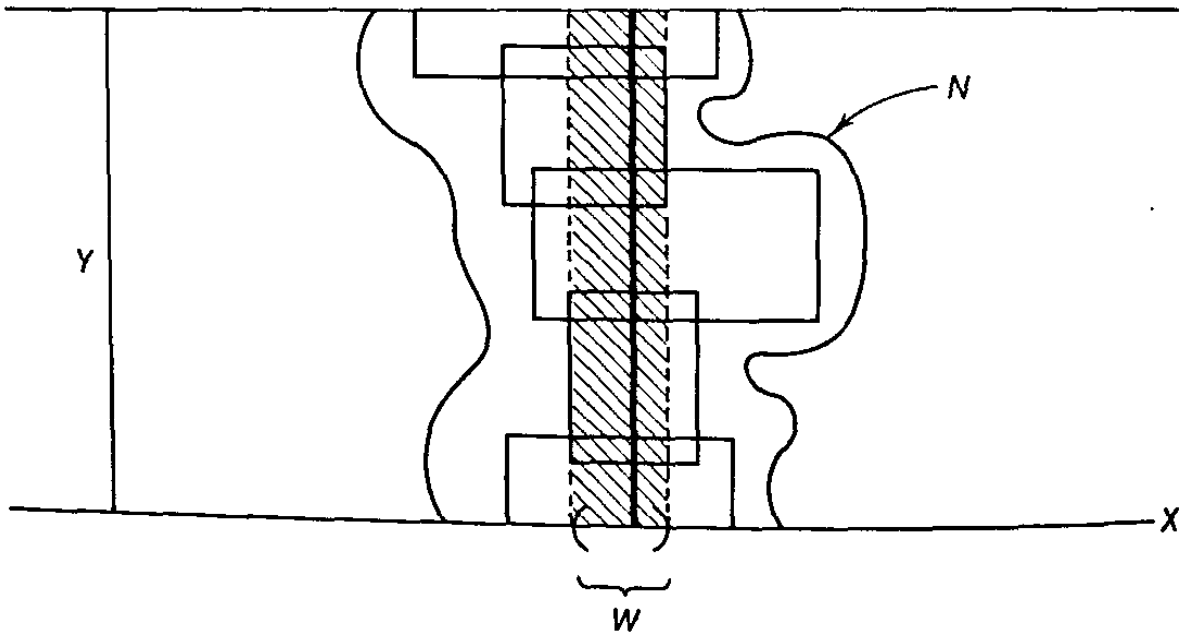


Figure 12

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.

§3-5

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 $A_1, \ldots, A_m$ 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 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 $A_1, \ldots, A_m$ 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

$$\{W_1, \ldots, W_k\}$$

covering X. The union of the tubes

$$W_1 \times Y, \ldots, W_k \times Y$$

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 5.8 (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 R^2, and let

$$N = \{x \times y; |x| < 1/(y^2 + 1)\}.$$

Then N is an open set containing the set $0 \times R$, but it contains no tube about $0 \times R$. It is illustrated in Figure 13.

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, we proved it first for finite products and derived the general case from that. In proving that cartesian products of compact spaces are compact, however, there is no way to go from finite products to infinite ones. The infinite case demands an entirely 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. How-

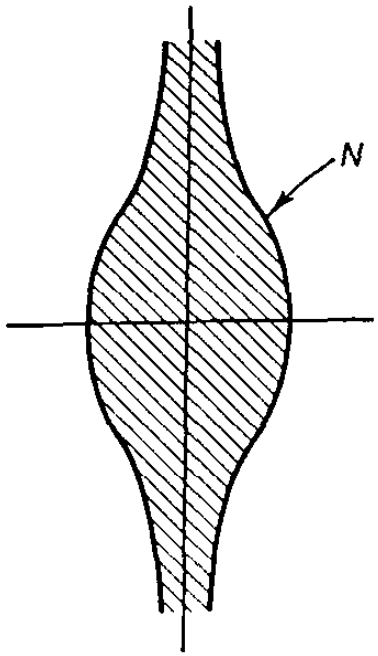


Figure 13

ever, you can study it now if you wish; the section in which it is proved (§5-1) 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 satisfy the **finite intersection condition** 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 5.9. *Let* X *be a topological space. Then* X *is compact if and only if for every collection* $\mathcal{C}$ *of closed sets in* X *satisfying the finite intersection condition, 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 $\{A_1, \ldots, A_n\}$ 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.

§3-5

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 above, 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 $C_1 \supset C_2 \supset \ldots C_n \supset C_{n+1} \supset \ldots$ of closed sets in a compact space X. If each of the sets C_n is nonempty, then the collection $\mathcal{C} = \{C_n\}_{n \in Z_+}$ automatically satisfies the finite intersection condition, as you can easily check. Then the intersection

$$\bigcap_{n \in Z_+} C_n$$

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 7 when we prove the Baire category theorem. Actually, in proving the Tychonoff theorem, we shall use the following slight reformulation of it, whose proof we leave to you.

Corollary 5.10. The space X *is compact if and only if for every collection* $\mathcal{A}$ *of subsets of* X *satisfying the finite intersection condition, the intersection*

$$\bigcap_{A \in \mathcal{A}} \bar{A}$$

of their closures is nonempty.

Exercises

1. (a) Let $\mathfrak{I}$ and $\mathfrak{I}'$ be two topologies on the set X; suppose that $\mathfrak{I}' \supset \mathfrak{I}$. 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 $\mathfrak{I}$ and $\mathfrak{I}'$, then either $\mathfrak{I} = \mathfrak{I}'$ or they are not comparable.

2. (a) Show that every subset of R is compact in the topology $\mathfrak{I}_f$. (See Example 6.)
   (b) Is $[0, 1]$ compact as a subspace of R in the topology
   $$\mathfrak{I}_c = \{A \mid R - A \text{ is countable or all of } R\}?$$
   In the lower limit topology R_l?
3. Show that a finite union of compact sets is compact.
4. Show that every compact subset of a metric space is bounded in that metric and is closed. Find a metric space in which not every closed bounded subset is compact.
5. Let A and B be disjoint compact subsets of the Hausdorff space X. Show that there exist disjoint open sets U and V containing A and B, respectively.
6. Show that if $f: X \longrightarrow Y$ 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. Prove Corollary 5.10.
8. Show that if Y is compact, then the projection $\pi_1 : X \times Y \longrightarrow X$ is a closed map.
9. *Theorem. Let $f: X \longrightarrow Y$; 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 $f(x_0)$, find a tube about $x_0 \times (Y - V)$ not intersecting G_f.]
10. Generalize the tube lemma as follows:
    Theorem. Let A and B be subsets of X and Y, respectively; let N be an open set in $X \times Y$ containing $A \times B$. If B is compact, then there exists an open set U in X such that
    $$A \times B \subset U \times B \subset N.$$
    If A and B are both compact, then there exist open sets U and V in X and Y, respectively, such that
    $$A \times B \subset U \times V \subset N.$$
11. (a) Prove the following partial converse to the uniform limit theorem:
    Theorem. Let $f_n : X \longrightarrow R$ be a sequence of continuous functions, with $f_n(x) \longrightarrow f(x)$ 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 Exercise 9 of §2-10.]
12. *Theorem. Let X be a compact Hausdorff space. Let $\mathcal{A}$ be a collection of closed connected subsets of X which is simply ordered by proper inclusion. Then*
    $$Y = \bigcap_{A \in \mathcal{A}} A$$
    is connected.

§ 3-6

[*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.]

13. Here is an exercise for those who have studied topological groups:

Theorem. 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:* Let $f: G \times B \to G$ be the map $f(x \times y) = x \cdot y^{-1}$. If $c \notin A \cdot B$, find a tube $W \times B$ about $c \times B$ lying in $f^{-1}(G - A)$.]

3-6 Compact Sets in the Real Line

The theorems of the preceding section enable us to construct new compact sets from old ones; but in order to get very far we have to find some specific compact sets. The natural place to start is the real line. We prove that every closed interval in R is compact. As an application, we prove the maximum value theorem of calculus, suitably generalized. Other applications include a characterization of all compact subsets of 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 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 6.1. *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. 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 $[a, b]$.

Step 1. First we prove the following: If x is a point of $[a, b]$ different from b, then there is a point $y > x$ in $[a, b]$ 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. 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 14. Since z is in C, the interval $[a, z]$ can be covered by finitely many, say n, elements of $\mathcal{A}$. Note that $[z, c]$ lies in the single element A of $\mathcal{A}$, whence $[a, c] = [a, z] \cup [z, c]$ can be covered by $n + 1$ elements of $\mathcal{A}$. Thus c is in C, contrary to assumption.

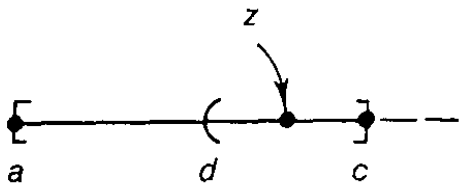


Figure 14

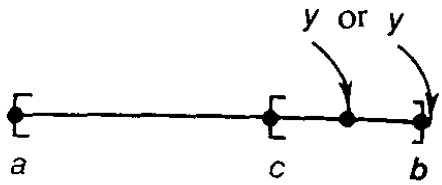


Figure 15

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 15. 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 6.2. *Every closed interval in* R *is compact.*

Now we characterize compact subsets of R^n:

Theorem 6.3. *A subset* A *of* 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 5.3, it is closed. Consider the collection of open sets

$$\{B_\rho(\mathbf{0}, m) \mid m \in Z_+\},$$

whose union is all of R^n. Some finite subcollection covers A. It follows that $A \subset B_\rho(0, M)$ 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 $\rho(x_0, 0) = b$. The triangle inequality implies that $\rho(x, 0) \leq N + b$ 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 R^n are compact, since they are closed and bounded. The set

$$A = \{x \times (1/x) \mid 0 < x \leq 1\}$$

is closed in 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 R^2, but it is not compact because it is not closed.

Now we prove the maximum value theorem of calculus, in suitably generalized form.

Theorem 6.4 (*Maximum and minimum 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 maximum value theorem of calculus is the special case of this theorem that occurs when we take X to be a closed interval in R and Y to be 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

$$\{(-\infty, a_1), \ldots, (-\infty, a_n)\}$$

covers A. If a_i is the largest of the elements $a_1, \ldots, a_n$, 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. □

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 R.

In fact, we prove the following more general result:

Theorem 6.5. *Let* X *be a (nonempty) compact Hausdorff space. If every point of* X *is a limit point of* X, *then* X *is uncountable.*

Proof. Step 1. First we show that, given a (nonempty) open set U of X, and given $x \in X$, there exists a (nonempty) open set V contained in U such that $\bar{V}$ does not contain x.

The point x may or may not be in U. But in either case, we can choose a point y in U that is different from x. This is possible if x is in U because x is a limit point of X (so that U must contain a point y different from x). And it is possible if x is not in U because U is nonempty. Let W_1 and W_2 be disjoint neighborhoods of x and y, respectively; then $V = U \cap W_2$ is the desired open set, whose closure does not contain x. See Figure 16.

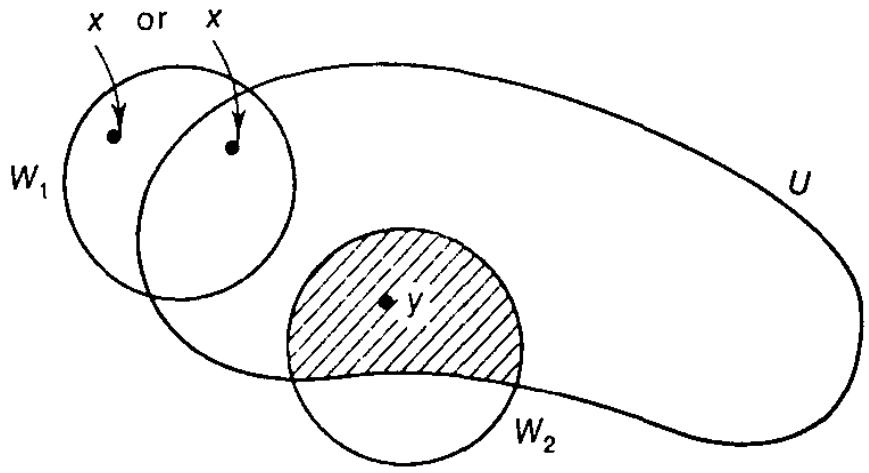


Figure 16

Step 2. We show that given $f: Z_+ \to X$, the function f is not surjective. It follows that X is uncountable.

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

$$\bar{V}_1 \supset \bar{V}_2 \supset \dots$$

of nonempty closed sets of X. Because X is compact, there is a point $x \in \bigcap \bar{V}_n$, by Theorem 5.9. The point x cannot equal x_n for any n, since x belongs to $\bar{V}_n$ and x_n does not. □

Corollary 6.6. *Every closed interval in* R *is uncountable.*

§3-6

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. Show that a connected metric space having more than one point is uncountable.

3. Let X be a metric space with metric d; let A be a subset of X. Recall that the *diameter* of A is defined by the equation

$$\operatorname{diam} A = \operatorname{lub}\{d(a_1, a_2) \mid a_1, a_2 \in A\},$$

if it exists. Show that if A is compact, then diam A exists and equals $d(a_1, a_2)$ for some $a_1, a_2 \in A$.

4. Let X be a metric space with metric d; let $A \subset X$.
   (a) If $x \in X$, define $d(x, A) = \operatorname{glb}\{d(x, a) \mid a \in A\}$. Show that the map carrying x into $d(x, A)$ is a continuous map of X into R.
   (b) Is it true that for some $a \in A$, we have $d(x, A) = d(x, a)$? Is this true if A is closed? If A is compact?
   (c) Is it true that $x \in \bar{A}$ if and only if $d(x, A) = 0$?
   (d) Define the *ϵ-neighborhood of A in X* to be the union

$$U(A, \epsilon) = \bigcup_{a \in A} B_d(a, \epsilon).$$

   Show that $U(A, \epsilon)$ equals the set

$$\{x \mid d(x, A) < \epsilon\}.$$

   (e) Show that if A and B are disjoint closed subsets of X and B is compact, then for some ϵ, the ϵ-neighborhoods of A and B are disjoint.
   (f) Does (e) hold if B is not compact?

5. Let X be a compact Hausdorff space. Show that if $\{A_n\}$ is a countable collection of closed sets in X, each of which has empty interior in X, then there is a point of X which is not in any set A_n. [*Hint:* Imitate the proof of Theorem 6.5.]

   This is a special case of the *Baire category theorem*, which we shall study in Chapter 7.

*6. Let A_0 be the closed interval $[0, 1]$ in 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

$$C = \bigcap_{n \in Z_+} A_n$$

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 every point of C is a limit point of C.
   (e) Conclude that C is uncountable.

3-7 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. One application is the third theorem of calculus quoted at the beginning of the chapter, the uniform continuity theorem, which we shall prove in suitably generalized form.

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 7.1. 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 that A has no limit point. Then A contains all its limit points and is therefore closed. Being a closed subset of a compact space, it is compact. For each a in A, we can choose a neighborhood U_a of a such that U_a does not intersect $A - \{a\}$, since a is not a limit point of A. The set A is covered by the open sets U_a; being compact, it can be covered by finitely many, say n, of them. Since each set U_a contains only one point of A, the set A itself contains n points.

The example following describes a space that is limit point compact but not compact. □

EXAMPLE 1. Consider the minimal uncountable well-ordered set S_Ω, in the order topology. The space S_Ω is not compact, for it is not closed in $\bar{S}_\Omega$. 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 already know $\bar{S}_\Omega$ is not metrizable, for it does not satisfy the "sequence lemma" (Example 3 of §2-10). Once we prove that compactness and limit point compactness are equivalent in metrizable spaces, it follows that S_Ω is not metrizable either, even though it does satisfy the sequence lemma. For S_Ω is limit point compact but not compact.

In order to prove the uniform continuity theorem, we need first to prove a classical lemma concerning open coverings of metric spaces. Recall that the *diameter* of a bounded subset A of a metric space (X, d) is the number

$$\text{lub}\,\{d(a_1, a_2) \mid a_1, a_2 \in A\}.$$

Lemma 7.2 (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. Step 1. Because X is compact, it is necessarily limit point compact. We shall prove that limit point compactness of X in turn implies that every infinite sequence (x_n) in X has a convergent subsequence. That is, there is an increasing sequence

$$n_1 < n_2 < \cdots < n_i < \cdots$$

of positive integers such that the sequence

$$x_{n_1}, x_{n_2}, \ldots, x_{n_i}, \ldots$$

converges.

(If every sequence in a space Y has a convergent subsequence, we say that Y is **sequentially compact.**)

Given the sequence (x_n), let us consider the set

$$A = \{x_n \mid n \in Z_+\}.$$

First, suppose that the set A is finite. In this case, we assert that there is a point x such that $x = x_n$ for infinitely many values of n. [To prove this assertion, let $f: Z_+ \to A$ be the indexing function defined by $f(n) = x_n$. Then since Z_+ is the union of the finite collection of sets $f^{-1}(x)$, as x ranges over A, at least one of the sets $f^{-1}(x)$ must be infinite.] Then the sequence (x_n) has a subsequence that is *constant*, and therefore converges automatically.

Second, suppose that 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 $x_{n_1}, x_{n_2}, \ldots$ converges to x.

Step 2. Now we show that if X is sequentially compact, then every open covering $\mathcal{A}$ of X has a Lebesgue number δ.

We shall prove the contrapositive: Suppose there is no $\delta > 0$ such that every set of diameter less than δ lies in at least one element of $\mathcal{A}$. Then X is not sequentially compact.

So let us assume there is no such δ. This means that for each $\delta > 0$, there exists a subset of X having diameter less than δ which does not lie inside any element of $\mathcal{A}$. In particular, for each $n \in Z_+$, we can choose a set C_n having diameter less than $1/n$ which is not contained in any element of $\mathcal{A}$. Choose, for each n, a point x_n of C_n. We assert that the sequence (x_n) has no convergent subsequence.

Suppose that (x_n) had a convergent subsequence (x_{n_i}), converging to x, say. Now x belongs to some element A of $\mathcal{A}$, and because A is open, there is an $\epsilon > 0$ such that $B(x, \epsilon) \subset A$. Choose i large enough that

$$d(x_{n_i}, x) < \epsilon/2 \quad \text{and} \quad 1/n_i < \epsilon/2.$$

Now C_{n_i} lies in the $1/n_i$ neighborhood of x_{n_i}: it follows that

$$C_{n_i} \subset B(x, \epsilon).$$

Then $C_{n_i} \subset A$, contradicting the choice of the sets C_n. □

The preceding proof is a typical proof of nonconstructive nature—it shows δ exists without giving any hint of how to find it.

Now we prove the third of the theorems of calculus stated at the beginning of the chapter, in suitably generalized form.

Theorem 7.3 (***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. That is, given* $\epsilon > 0$, *there exists a* $\delta > 0$ *such that for any two points* x_1, x_2 *of* X,

$$d_X(x_1, x_2) < \delta \implies d_Y(f(x_1), f(x_2)) < \epsilon.$$

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. □

§ 3-7

Now we prove that compactness and limit point compactness are equivalent in metrizable spaces.

Theorem 7.4. 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 proved above that (1) $\Rightarrow$ (2) $\Rightarrow$ (3); see the proofs of Theorem 7.1 and Lemma 7.2. We have also proved that sequential compactness implies that every open covering of X has a Lebesgue number. We now complete the proof by showing that sequential compactness of X implies compactness of X. So assume that X is sequentially compact.

Step 1. First we show that for every $\epsilon > 0$, there exists a finite covering of X by ϵ-balls. And once again, we prove the contrapositive: If for some $\epsilon > 0$, X cannot be covered by finitely many ϵ-balls, then X is not sequentially compact.

So suppose 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 $x_1, \ldots, x_n$, 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 $i = 1, \ldots, n$. 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.

Step 2. Now we prove X compact. Let $\mathcal{A}$ be an open covering of X. Since X is sequentially compact, the covering $\mathcal{A}$ has a Lebesgue number δ. Using Step 1, choose a finite covering of X by balls of radius $\delta/3$. Each of these balls has diameter at most $2\delta/3$, so we can choose for each of these balls an element of $\mathcal{A}$ containing it. We thus obtain a finite subcollection of $\mathcal{A}$ that covers X. □

Exercises

1. Let X be a two-point space in the indiscrete topology. Show that $X \times Z_+$ is limit point compact but not compact.
2. Show that $[0, 1]^\omega$ is not limit point compact in either the box or the uniform topologies.

3. Show that the set [0, 1] is not limit point compact as a subspace of R_l.

4. A space X is said to be **countably compact** if every countable open covering of X contains a finite subcollection covering X.
   (a) Show that countable compactness implies limit point compactness.
   (b) Prove the converse if X is Hausdorff. [*Hint:* If no finite subcollection of $\{U_n\}$ covers X, choose $x_n \notin U_1 \cup \cdots \cup U_n$.]

5. Let X be limit point compact.
   (a) If $f: X \longrightarrow Y$ is continuous, is $f(X)$ necessarily limit point compact?
   (b) If A is a closed subset of X, is A necessarily limit point compact?
   (c) If X is a subspace of Z, is X necessarily a closed set in Z?
   (d) Repeat (a)–(c) under the additional assumption that the spaces involved are Hausdorff.
   *(e) 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.

6. Let X, Y, and Z be metric spaces.
   (a) If $f: X \longrightarrow Y$ and $g: Y \longrightarrow Z$ are uniformly continuous, is $g \circ f: X \longrightarrow Z$ necessarily uniformly continuous?
   (b) If $f: X \longrightarrow Y$ is a homeomorphism and f is uniformly continuous, is f^{-1} necessarily uniformly continuous?

7. Let (X, d) be a compact metric space. Let $f: X \longrightarrow X$ be a continuous map.
   (a) We say that f is a **contraction** if there is a number $\alpha < 1$ such that
   $$d(f(x), f(y)) \leq \alpha d(x, y)$$
   for all points $x, y \in X$. Show that if f is a contraction, there is a unique point x of X such that $f(x) = x$. [*Hint:* Consider $\bigcap f^n(X)$.]
   (b) We say that f is an **isometry** if $d(f(x), f(y)) = d(x, y)$ for all points $x, y \in X$. Show that if f is an isometry, then f is surjective. [*Hint:* If $x \notin f(X)$, let the ϵ-neighborhood of x be disjoint from $f(X)$. Define $x_1 = x$ and, in general, $x_{n+1} = f(x_n)$. Show that $d(x_n, x_m) \geq \epsilon$ for $n \neq m$.]
   (c) Give examples to show these theorems fail if X is not compact.

*8. Show that $[0, 1]^\omega$ is compact in the product topology.

*3-8 *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, called its *one-point compactification.*

Definition. A space X is said to be **locally compact at x** if there is some compact subset 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.**

†This section will be assumed in §5-3, in Chapter 7, and in §8-12.

§ 3-8

Note that a compact space is automatically locally compact.

EXAMPLE 1. The real line R is locally compact. The point x lies in some interval (a, b), which in turn is contained in the compact set $[a, b]$. The subspace Q of rational numbers is not locally compact, as you can check.

EXAMPLE 2. The space 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 set $[a_1, b_1] \times \cdots \times [a_n, b_n]$. The space R^ω is not locally compact; *none* of its basis elements are contained in compact sets. For if

$$B = (a_1, b_1) \times \cdots \times (a_n, b_n) \times R \times \cdots \times R \times \cdots$$

were contained in a compact set, then its closure

$$\bar{B} = [a_1, b_1] \times \cdots \times [a_n, b_n] \times R \times \cdots$$

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 problem in Chapter 5 when we study compactifications in general.

Definition. Let X be a locally compact Hausdorff space. Take some object outside X, denoted 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 in Y to be all sets of the following types:

(1) U, where U is an open subset of X,

(2) $Y - C$, where C is a compact subset of X.

The space Y is called the **one-point compactification** 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:

$$U_1 \cap U_2 \qquad \text{is of type (1).}$$
$$(Y - C_1) \cap (Y - C_2) = Y - (C_1 \cup C_2) \qquad \text{is of type (2).}$$
$$U_1 \cap (Y - C_1) = U_1 \cap (X - C_1) \qquad \text{is of type (1),}$$

because C_1 is closed in X. Similarly, one checks that the union of any collection of open sets is open:

$$\bigcup U_\alpha = U \qquad \text{is of type (1).}$$
$$\bigcup (Y - C_\beta) = Y - (\bigcap C_\beta) = Y - C \qquad \text{is of type (2).}$$
$$(\bigcup U_\alpha) \cup (\bigcup (Y - C_\beta)) = U \cup (Y - C) = Y - (C - U),$$

which is of type (2) because $C - U$ is a closed subset of C and therefore compact.

The basic properties of the one-point compactification are given in the following theorem.

Theorem 8.1. *Let* X *be a locally compact Hausdorff space which is not compact; let* Y *be the one-point compactification of* X. *Then* Y *is a compact Hausdorff space;* X *is a subspace of* Y; *the set* Y − X *consists of a single point; and* $\bar{\mathrm{X}} = \mathrm{Y}$.

Proof. First we show that X is a subspace of Y and $\bar{X} = Y$. Given any open set of Y, its intersection with X is open in X, since $U \cap X = U$ and $(Y - C) \cap X = X - C$, both of which are open in X. Conversely, any set open in X is a set of type (1) and therefore open in Y. Since X is not compact, each open set $Y - C$ containing the point ∞ intersects X. Therefore, ∞ is a limit point of X, so that $\bar{X} = Y$.

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 in 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. □

EXAMPLE 4. The one-point compactification of the real line R is homeomorphic with the circle, as you may readily check. Similarly, the one-point compactification of R^2 is homeomorphic to the sphere S^2. If R^2 is looked at as the space C of complex numbers, then $C \cup \{\infty\}$ is called the *Riemann sphere,* or the *extended complex plane.*

§ 3-8

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.

Lemma 8.2. Let X *be a Hausdorff space. Then* X *is locally compact at* x *if and only if for every neighborhood* U *of* x, *there is a neighborhood* V *of* x *such that* $\bar{\mathrm{V}}$ *is compact and* $\bar{\mathrm{V}} \subset \mathrm{U}$.

Proof. It is clear that this new formulation implies local compactness; the set $C = \bar{V}$ is the desired compact set containing a neighborhood of x. Conversely, suppose that C is a compact set containing a neighborhood of x. Let U be an arbitrary neighborhood of x. Let $A = C - U$; being closed in C, the set A is compact. Apply Lemma 5.4 to choose disjoint open sets W and W' about x and A, respectively. See Figure 17. Let $V = W \cap (\mathrm{Int}\, C)$; recall that Int C is the union of all open sets contained in C. Then V is a neighborhood of x. Since X is Hausdorff, C is closed in X; therefore, $\bar{V} \subset C$, so that $\bar{V}$ is compact. Because V is contained in W, which is disjoint from W', the set $\bar{V}$ cannot intersect A. Therefore, $\bar{V} \subset C - A$, so that $\bar{V} \subset U$. □

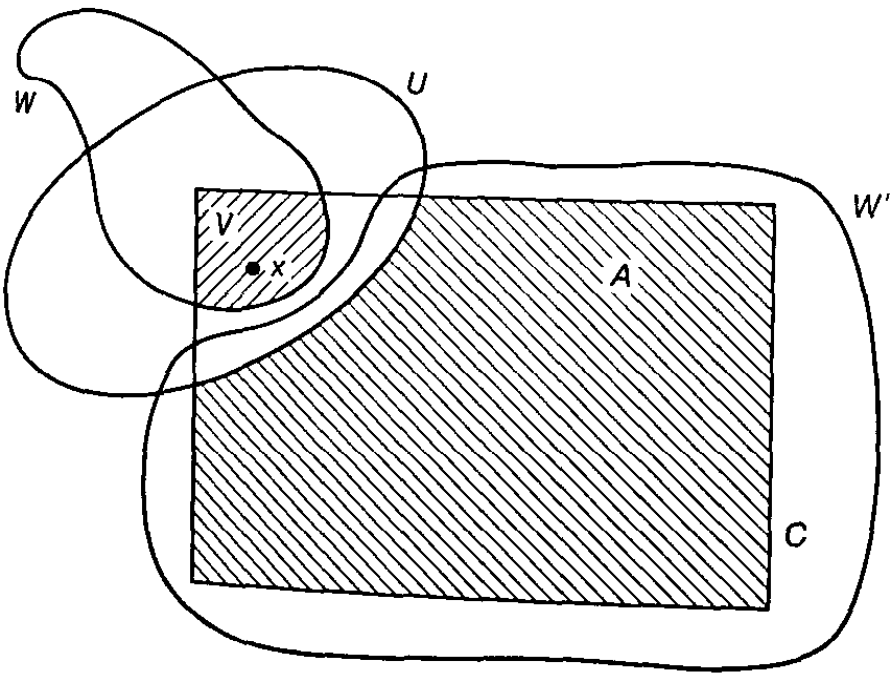


Figure 17

A somewhat simpler proof of this lemma is available if X is locally compact at *every* point. One uses the fact that the complement of U in the one-point compactification Y of X is compact; one chooses disjoint neighborhoods V and W of x and $Y - U$, respectively. Then the closure $\bar{V}$ of V in Y is compact and is contained in U.

Corollary 8.3. Let X *be locally compact Hausdorff; let* Y *be a subspace of* X. *If* Y *is closed in* X *or open in* X, *then* Y *is locally compact.*

Proof. Suppose that Y is closed in X. Given $y \in Y$, let C be a compact set in X containing the neighborhood U of y in X. Then $C \cap Y$ is closed in C and thus compact, and it contains the neighborhood $U \cap Y$ of y in Y. (We have not used the Hausdorff condition here.)

Suppose that Y is open in X. Given $y \in Y$, we apply the previous lemma to choose a neighborhood V of y in X such that $\bar{V}$ is compact and $\bar{V} \subset Y$. Then $C = \bar{V}$ is a compact set in Y containing the neighborhood V of y in Y. □

Corollary 8.4. A space X *is homeomorphic to an open subset of a compact Hausdorff space if and only if* X *is locally compact Hausdorff.*

Proof. This follows from Theorem 8.1 and Corollary 8.3. □

Exercises

1. Show that the rationals Q are not locally compact.
2. (a) Show that if $\coprod X_\alpha$ is locally compact, then each X_α is locally compact and X_α is compact for all but finitely many values of α. (Assume that each X_α is nonempty.)
   (b) Prove the converse, assuming the Tychonoff theorem.
3. Let X be a locally compact space. If $f: X \longrightarrow Y$ is continuous, is the space $f(X)$ necessarily 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. Show that the conditions in Theorem 8.1 characterize the one-point compactification up to homeomorphism.
6. Show that the one-point compactification of 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 Z_+ is homeomorphic with the subspace $\{0\} \cup \{1/n \mid n \in Z_+\}$ of R.

*9. Show that the one-point compactification of R^2 is homeomorphic with the two-sphere S^2.

10. Show that if G is a locally compact topological group and H is a subgroup, then G/H is locally compact.

*11. (a) Prove the following:

Lemma. If $p: X \longrightarrow Y$ 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 $U_i \times \bar{V} \subset \pi^{-1}(A)$, 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 $\pi(U \times V)$ is a neighborhood of $\pi(x \times y)$ contained in A.]

(b) Prove:

Theorem. Let $p : A \longrightarrow B$ and $q : C \longrightarrow D$ be quotient maps. If B and C are locally compact Hausdorff spaces, then $p \times q : A \times C \longrightarrow B \times D$ 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 metric 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 $\preceq$.
   (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 $x_\alpha \longrightarrow x$) if for

each neighborhood U of x, there exists $\alpha \in J$ such that

$$\alpha \leq \beta \Longrightarrow x_\beta \in U.$$

Show that these definitions reduce to familiar ones when $J = Z_+$.

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 $\Longrightarrow$, take as index set the collection of all neighborhoods of x, partially ordered by reverse inclusion.]

7. *Theorem. Let $f: X \longrightarrow Y$. 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 $f: J \longrightarrow X$ be a net in X; let $f(\alpha) = x_\alpha$. If K is a directed set and $g: K \longrightarrow J$ is a function such that

(i) $i \leq j \Longrightarrow g(i) \leq g(j)$,

(ii) $g(K)$ is cofinal in J,

then the composite function $f \circ g: K \longrightarrow X$ 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 $\Longrightarrow$, 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 $(\alpha, U) \leq (\beta, V)$ if $\alpha \leq \beta$ 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 $\Longrightarrow$, let $B_\alpha = \{x_\beta \mid \alpha \leq \beta\}$ and show that $\{B_\alpha\}$ satisfies the finite intersection condition. To prove $\Longleftarrow$, let $\mathcal{A}$ be a collection of closed sets satisfying the finite intersection condition, 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.

4. Countability and Separation Axioms

The concepts we are going to introduce now, unlike compactness and connectedness, do not arise naturally from the study of calculus and analysis. They arise instead from a deeper study of topology itself. Such problems as imbedding a given space in a metric space or in a compact Hausdorff space are basically problems of topology rather than analysis. These particular problems have solutions that involve the countability and separation axioms.

We have already introduced the first countability axiom; it arose in connection with our study of convergent sequences in §2-10. And we have also studied one of the separation axioms—the Hausdorff axiom. In this chapter we shall introduce other, and stronger, axioms like these and explore some of their consequences. Our basic goal is to prove the *Urysohn metrization theorem*. It says that if a topological space X satisfies a certain countability axiom (the second) and a certain separation axiom (the regularity axiom), then X can be imbedded in a metric space and is thus metrizable. This classical theorem may be considered the culmination of Part I of the book.

Another imbedding theorem, important to geometers, appears in the last section of the chapter. Given a space which is a compact *manifold* (the higher-dimensional analogue of a surface), we show that it can be imbedded in some finite-dimensional euclidean space.

4-1 *The Countability Axioms*

Recall the definition we gave in §2-10:

Definition. A space X is said to have a **countable basis at** x if there is a countable collection $\mathcal{B}$ of neighborhoods of x such that each neighborhood of x contains at least one of the elements of $\mathcal{B}$. A space that has a countable basis at each of its points is said to satisfy the **first countability axiom.**

We have already noted that every metrizable space satisfies this axiom; see §2-10.

The most useful fact concerning spaces that satisfy this axiom is the fact that in such a space, convergent sequences are adequate to detect limit points of sets and to check continuity of functions. We have noted this before; now we state it formally as a theorem:

Theorem 1.1. *Let* X *be a space satisfying the first countability axiom.*

(a) *The point* x *belongs to the closure* $\bar{A}$ *of the subset* A *of* X *if and only if there is a sequence of points of* A *converging to* x.

(b) *The function* $f : X \to Y$ *is continuous if and only if for every convergent sequence* (x_n) *in* X, *converging to* x, *say, the sequence* $(f(x_n))$ *converges to* $f(x)$.

The proof is a direct generalization of the proof given in §2-10 under the hypothesis of metrizability, so it will not be repeated here.

Of much greater importance than the first countability axiom is the following:

Definition. A topological space X is said to satisfy the **second countability axiom** if X has a countable basis for its topology.

Obviously the second axiom implies the first: If $\mathcal{B}$ is a countable basis for the topology of X, then the subset of $\mathcal{B}$ consisting of those basis elements containing the point x is a countable basis at x. The second axiom is, in fact, much stronger than the first; it is so strong that not even every metric space satisfies it.

Why then is the property interesting? Well, for one thing, many familiar spaces do have this property. For another, it is a crucial tool used in proving the Urysohn metrization theorem, as we shall see.

EXAMPLE 1. The real line R has a countable basis—the collection of all open intervals (a, b) with rational end points. Likewise, R^n has a countable basis—the collection of all products of intervals having rational end points. Even R^ω has a countable basis—the collection of all products $\prod_{n \in Z_+} U_n$, where U_n is an open interval with rational end points for finitely many values of n, and $U_n = R$ for all other values of n.

§4-1

EXAMPLE 2. In the uniform topology, R^ω satisfies the first countability axiom (being metrizable). But it does not satisfy the second. Consider the uncountable subset C of R^ω consisting of all sequences of 0's and 1's. If $\mathcal{B}$ is a basis for the uniform topology on R, we can choose, for each $x \in C$, an element B_x of $\mathcal{B}$ containing x and lying in the ball of radius 1 centered at x. If x and y are distinct points of C, then $B_x \neq B_y$; for $\bar{\rho}(x, y) = 1$, so that $y \notin B_x$. We conclude that $\mathcal{B}$ is uncountable.

Both countability axioms are well behaved with respect to the operations of taking subspaces or countable products:

Theorem 1.2. A subspace of a first-countable space is first-countable, and a countable product of first-countable spaces is first-countable. A subspace of a second-countable space is second-countable, and a countable product of second-countable spaces is second-countable.

Proof. Consider the second countability axiom. If $\mathcal{B}$ is a countable basis for X, then $\{B \cap A \mid B \in \mathcal{B}\}$ is a countable basis for the subspace A of X. If $\mathcal{B}_i$ is a countable basis for the space X_i, then the collection of all products $\prod U_i$, where $U_i \in \mathcal{B}_i$ for finitely many values of i and $U_i = X_i$ for all other values of i, is a countable basis for $\prod X_i$.

The proof for the first countability axiom is similar. □

Two consequences of the second countability axiom that will be useful to us later are given in the following theorem. First, a definition:

Definition. A subset A of a space X is said to be **dense** in X if $\bar{A} = X$.

Theorem 1.3. Suppose that X *has a countable basis. Then:*
(a) *Every open covering of* X *contains a countable subcollection covering* X.
(b) *There exists a countable subset of* X *which is dense in* X.

Proof. Let $\{B_n\}$ be a countable basis for X.

(a) Let $\mathcal{A}$ be an open covering of X. For each positive integer n for which it is possible, choose an element A_n of $\mathcal{A}$ containing the basis element B_n. The collection $\mathcal{A}'$ of the sets A_n is countable, since it is indexed with a subset of the positive integers. Furthermore, it covers X: Given a point $x \in X$, we can choose an element A of $\mathcal{A}$ containing x. Since A is open, there is a basis element B_n such that $x \in B_n \subset A$. Since B_n lies in an element of $\mathcal{A}$, the set A_n is defined for the index n; since A_n contains B_n, it contains x. Thus $\mathcal{A}'$ is a countable subcollection of $\mathcal{A}$ that covers X.

(b) From each nonempty basis element B_n, choose a point x_n. The set $D = \{x_n \mid n \in Z_+\}$ is dense in X; given $x \in X$, every basis element about x intersects the set D. □

The two properties listed in Theorem 1.3 are sometimes taken as alternative countability axioms. A space for which every open covering contains a countable subcovering is usually called a **Lindelöf space**. A space having

a countable dense subset is often said to be **separable** (an unfortunate choice of terminology).† Weaker in general than the second countability axiom, each of these properties is equivalent to the second countability axiom when the space is metrizable (see Exercise 7). They are less important than the second countability axiom, but you should be aware of their existence, for they are sometimes useful. It is often easier, for instance, to show that a space X has a countable dense subset than it is to show that X has a countable basis. If the space is metrizable (as it usually is in analysis), solving the easier problem solves the harder problem as well.

We shall not use these properties to prove any theorems, but one of them —the Lindelöf condition—will be useful in dealing with some examples. Neither of them is as well behaved as one might wish under the operations of taking subspaces and cartesian products, as we shall see in the examples that follow. (See also Exercise 10.)

EXAMPLE 3. *The space R_l satisfies all the countability axioms but the second.*

Given $x \in R_l$, the set of all basis elements of the form $[x, x + 1/n)$ is a countable basis at x. And it is easy to see that the rational numbers are dense in R_l.

To see that R_l has no countable basis, let $\mathcal{B}$ be a basis for R_l. Choose for each x, an element B_x of $\mathcal{B}$ such that $x \in B_x \subset [x, x + 1)$. If $x \neq y$, then $B_x \neq B_y$, since $x = \text{glb } B_x$ and $y = \text{glb } B_y$. Therefore, $\mathcal{B}$ must be uncountable.

To show that R_l is Lindelöf requires more work. It will suffice to show that every open covering of R_l by basis elements contains a countable subcollection covering R_l. (You can check this.) So let

$$\mathcal{A} = \{[a_\alpha, b_\alpha)\}_{\alpha \in J}$$

be a covering of R by basis elements for the lower limit topology. We wish to find a countable subcollection that covers R. Let C be the set

$$C = \bigcup_{\alpha \in J} (a_\alpha, b_\alpha),$$

considered as a subspace of R. Then C satisfies the second countability axiom. Because the collection $\{(a_\alpha, b_\alpha)\}$ consists of sets open in C, it must contain a countable subcollection that covers C, consisting, say, of the elements (a_α, b_α) for $\alpha = \alpha_1, \alpha_2, \ldots$. Then the collection

$$\mathcal{A}' = \{[a_\alpha, b_\alpha) \mid \alpha = \alpha_1, \alpha_2, \ldots\}$$

also covers C.

We assert that the set $R - C$ is countable. From this our result follows: Choose for each point of $R - C$ an element of $\mathcal{A}$ containing it; adjoining these elements to $\mathcal{A}'$, one obtains a countable subcollection of $\mathcal{A}$ that covers all of R_l.

So let x be a point of $R - C$. Necessarily $x = a_\alpha$ for some $\alpha \in J$. Choose q_x to be a rational number belonging to the interval (a_α, b_α); because this inter-

†This is a good example of how a word can be overused. We have already defined what we mean by a *separation* of a space; and we shall discuss the *separation* axioms shortly.

§4-1

val is contained in C, so is the interval (x, q_x). It follows that the function $x \to q_x$ is an injection of $R - C$ into the set Q of rational numbers, so that $R - C$ is countable. [For if x and y are two points of $R - C$ and $x < y$, then necessarily $q_x < q_y$, since otherwise y would belong to (x, q_x), although y is in $R - C$ and (x, q_x) is contained in C.]

EXAMPLE 4. *The product of two Lindelöf spaces need not be Lindelöf.* For the space R_l is Lindelöf, and we shall show that the space R_l^2 is not.

The space R_l^2 is an extremely useful example in topology called the Sorgenfrey plane. It has as basis all sets of the form $[a, b) \times [c, d)$ in the plane. Consider the subspace

$$L = \{x \times (-x) \mid x \in R_l\}$$

of R_l^2. It is easy to check that L is closed in R_l^2. Let us cover R_l^2 by the open set $R_l^2 - L$ and by all basis elements of the form

$$[a, b) \times [-a, d).$$

Each of these basis elements intersects L in at most one point. Since L is uncountable, no countable subcollection covers R_l^2. See Figure 1.

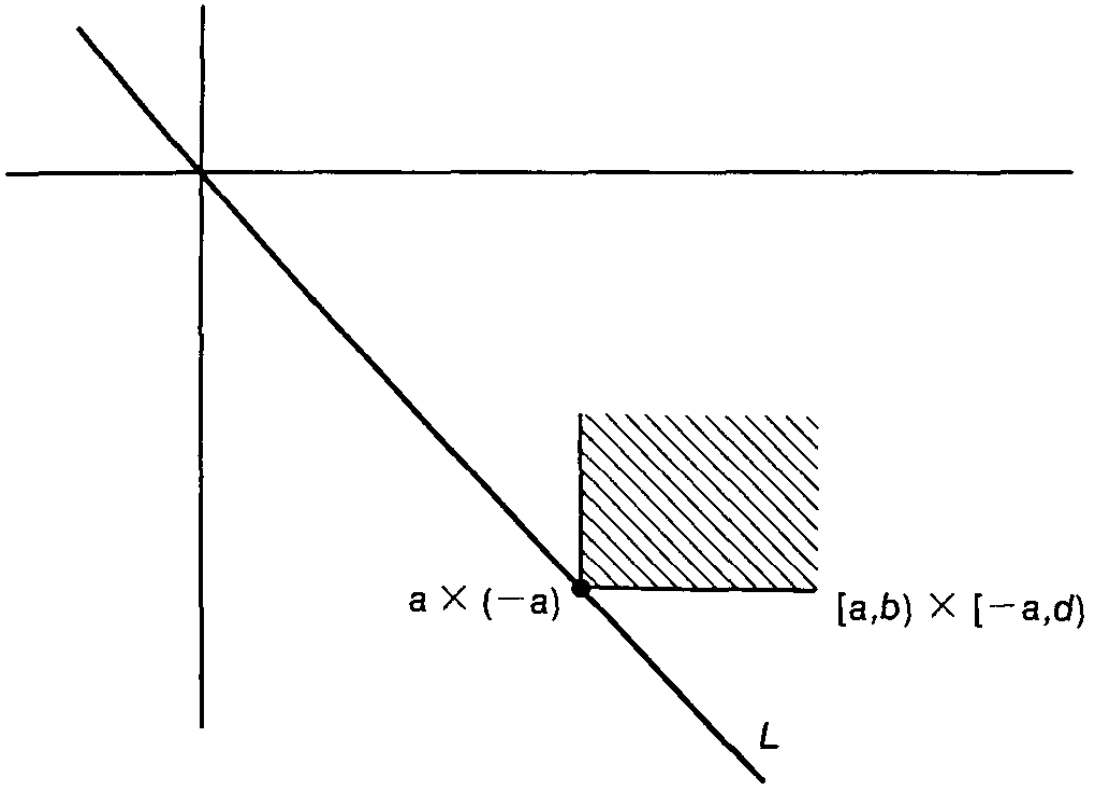


Figure 1

EXAMPLE 5. *A subspace of a space having a countable dense subset need not have a countable dense subset.* It is easy to see that the set of points having rational coordinates is dense in R_l^2. But the subspace L has the discrete topology; being uncountable, it cannot have a countable dense subset.

EXAMPLE 6. *The space S_Ω satisfies the first countability axiom, but has none of the other countability properties.*

Given a in S_Ω, let b be the immediate successor of a. The collection of all intervals whose end points are less than or equal to b is countable, so S_Ω has a countable basis at a.

Clearly, S_Ω has no countable dense subset, since any countable subset of S_Ω has an upper bound in S_Ω. Similarly, S_Ω is not Lindelöf: Consider the collection of open sets

$$\mathcal{A} = \{[a_0, b) \mid b \in S_\Omega\}$$

covering S_Ω, where a_0 is the smallest element of S_Ω. If $\mathcal{A}'$ is a countable subcollection of $\mathcal{A}$, the right-hand end points of the intervals in $\mathcal{A}'$ have an upper bound b in S_Ω; then b is contained in no element of $\mathcal{A}'$. Therefore, $\mathcal{A}'$ does not cover S_Ω, so S_Ω is not Lindelöf. As a consequence, S_Ω cannot have a countable basis.

Note that *the space $\bar{S}_\Omega$ is Lindelöf* (being compact), *and the subspace S_Ω is not.*

Exercises

1. (a) A *G_δ set* in a space X is a set A that equals a countable intersection of open sets of X. Show that in a first-countable Hausdorff space, every one-point set is a G_δ set.

   (b) There is a familiar space in which every one-point set is a G_δ set, which nevertheless does not satisfy the first countability axiom. What is it?

2. Show that every compact metrizable space X has a countable basis. [*Hint:* Let $\mathcal{A}_n$ be a finite covering of X by $1/n$-balls.]

3. Show that if X has a countable basis, every collection of disjoint open sets in X is countable.

4. Let X have a countable basis; let A be an uncountable subset of X.
   (a) Show that the subspace topology on A is not the discrete topology.
   (b) Show that A contains at least one of its limit points.
   (c) Show that A contains uncountably many of its limit points.

5. Show that if X has a countable basis $\{B_n\}$, then every basis $\mathcal{C}$ for X contains a countable basis for X. [*Hint:* For every pair of indices n, m for which it is possible, choose $C_{n,m} \in \mathcal{C}$ such that $B_n \subset C_{n,m} \subset B_m$. Let $\mathcal{C}'$ be the collection of the sets $C_{n,m}$.]

6. Show that in a Hausdorff space with a countable basis, all four varieties of compactness (compactness, limit point compactness, sequential compactness, countable compactness) are equivalent.

7. Let X be metrizable.
   (a) Show that if X has a countable dense subset, X has a countable basis.
   (b) Show that if X is Lindelöf, X has a countable basis. [*Hint:* Compare Exercise 2.]

§4-2

8. Show that R_l and $I \times I$ in the dictionary order topology are not metrizable.
9. Which countability axioms are satisfied by R^ω in the uniform topology?
10. Show that if X is a countable product of spaces having countable dense subsets, then X also has a countable dense subset.
11. Show that if X is Lindelöf and Y is compact, then $X \times Y$ is Lindelöf.
12. Consider the set $\mathcal{C}(I, R)$ of all continuous functions $f: I \longrightarrow R$, where $I = [0, 1]$, in the metric

$$\rho(f, g) = \text{lub } \{|f(x) - g(x)|\}.$$

Show that this space has a countable dense subset, and therefore has a countable basis. [*Hint:* Consider the set A of all continuous functions whose graphs are made up of finitely many line segments having rational end points.]

13. Show that the product space I^I, where $I = [0, 1]$, has a countable dense subset.
14. Show that the space ℓ^2, in the ℓ^2-metric, has a countable dense subset and therefore has a countable basis. (See Exercise 9 of §2-9.)

4-2 The Separation Axioms

In this section, we introduce three separation axioms and explore some of their properties. One you have already seen—the Hausdorff axiom. The others are similar but stronger. As always when we introduce new concepts, we shall examine the relationship between these axioms and the concepts introduced earlier in the book.

Recall that a space X is said to be *Hausdorff* if for each pair x, y of distinct points of X, there exist disjoint open sets containing x and y, respectively.

Definition. Suppose that one-point sets are closed in X. Then X is said to be regular if for each pair consisting of a point x and a closed set B disjoint from x, there exist disjoint open sets containing x and B, respectively. The space X is said to be **normal** if for each pair A, B of disjoint closed sets of X, there exist disjoint open sets containing A and B, respectively.

It is clear that a regular space is Hausdorff, and that a normal space is regular. (We need to include the condition that one-point sets be closed as part of the definition of regularity and normality in order for this to be the case. A two-point space in the indiscrete topology satisfies the other part of the definitions of regularity and normality, even though it is not Hausdorff.) For examples showing the regularity axiom stronger than the Hausdorff axiom, and normality stronger than regularity, see Examples 1 to 3.

These axioms are called separation axioms for the reason that they involve "separating" certain kinds of sets from one another by disjoint open sets. We have used the word "separation" before, of course, when we studied

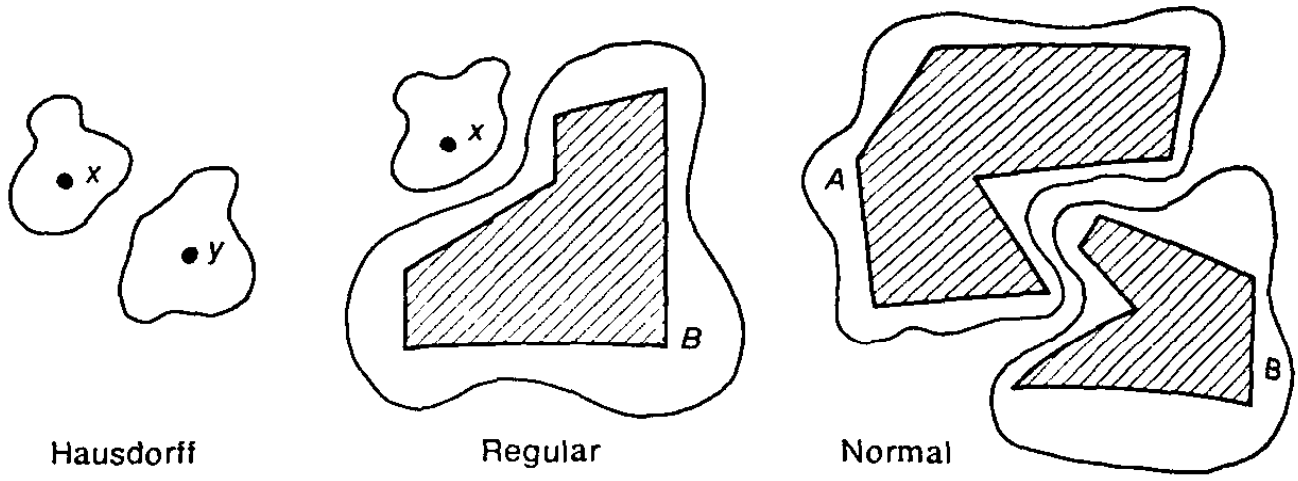


Figure 2

connected spaces. But in that case, we were trying to find disjoint open sets *whose union was the entire space*. The present situation is quite different, because the open sets need not satisfy this condition.

The three separation axioms are illustrated in Figure 2.

There are other ways to formulate the separation axioms. One formulation that is sometimes useful is given in the following lemma:

Lemma 2.1. *Let* X *be a topological space. Let one-point sets in* X *be closed.*

(a) X *is regular if and only if given a point* x *of* X *and a neighborhood* U *of* x, *there is a neighborhood* V *of* x *such that* $\bar{\mathrm{V}} \subset \mathrm{U}$.

(b) X *is normal if and only if given a closed set* A *and an open set* U *containing* A, *there is an open set* V *containing* A *such that* $\bar{\mathrm{V}} \subset \mathrm{U}$.

Proof. (a) Suppose that X is regular, and suppose that the point x and the neighborhood U of x are given. Let $B = X - U$; then B is a closed set. By hypothesis, there exist disjoint open sets V and W containing x and B, respectively. The set $\bar{V}$ is disjoint from B, since if $y \in B$, the set W is a neighborhood of y disjoint from V. Therefore, $\bar{V} \subset U$, as desired.

To prove the converse, suppose the point x and the closed set B not containing x are given. Let $U = X - B$. By hypothesis, there is a neighborhood V of x such that $\bar{V} \subset U$. The open sets V and $X - \bar{V}$ are disjoint open sets containing x and B, respectively. Thus X is regular.

(b) This proof uses exactly the same argument; one just replaces the point x by the set A throughout. □

Now we seek to relate the separation axioms with the concepts previously introduced. First we study the various kinds of topologies that were introduced in Chapter 2 and see what separation axioms they satisfy. And we prove two theorems which show that in the presence of compactness, or of

a countable basis, weak separation properties become strong. This gives us a fairly long list of theorems to remember, but luckily most of the proofs are not too difficult.

Theorem 2.2. (a) *A subspace of a Hausdorff space is Hausdorff; a product of Hausdorff spaces is Hausdorff.*

(b) *A subspace of a regular space is regular; a product of regular spaces is regular.*

(c) *A subspace of a normal space need not be normal; a product of normal spaces need not be normal* (!)

Proof. (a) Let X be Hausdorff. Let x and y be two points of the subspace Y of X. If U and V are disjoint neighborhoods in X of x and y, respectively, then $U \cap Y$ and $V \cap Y$ are disjoint neighborhoods of x and y in Y.

Let $\{X_\alpha\}$ be a family of Hausdorff spaces. Let $\mathbf{x} = (x_\alpha)$ and $\mathbf{y} = (y_\alpha)$ be distinct points of the product space $\prod X_\alpha$. Since $\mathbf{x} \neq \mathbf{y}$, there is some index β such that $x_\beta \neq y_\beta$. Choose disjoint open sets U and V in X_β containing x_β and y_β, respectively. Then the sets $\pi_\beta^{-1}(U)$ and $\pi_\beta^{-1}(V)$ are disjoint open sets in $\prod X_\alpha$ containing $\mathbf{x}$ and $\mathbf{y}$, respectively.

(b) Let Y be a subspace of the regular space X. Since Y is Hausdorff, one-point sets are closed in Y. Let x be a point of Y and let B be a closed subset of Y disjoint from x. Now $\bar{B} \cap Y = B$, where $\bar{B}$ denotes the closure of B in X. Therefore, $x \notin \bar{B}$, so, using regularity of X, we can choose disjoint open sets U and V of X containing x and $\bar{B}$, respectively. Then $U \cap Y$ and $V \cap Y$ are disjoint open sets in Y containing x and B, respectively.

Let $\{X_\alpha\}$ be a family of regular spaces; let $X = \prod X_\alpha$. By (a), X is Hausdorff, so that one-point sets are closed in X. We use the preceding lemma to prove regularity of X. Let $\mathbf{x} = (x_\alpha)$ be a point of X and let U be a neighborhood of $\mathbf{x}$ in X. Choose a basis element $\prod U_\alpha$ about $\mathbf{x}$ contained in U. Choose, for each α, a neighborhood V_α of x_α in X_α such that $\bar{V}_\alpha \subset U_\alpha$; if it happens that $U_\alpha = X_\alpha$, choose $V_\alpha = X_\alpha$. Then $V = \prod V_\alpha$ is a neighborhood of x in X. We assert that $\bar{V} = \prod \bar{V}_\alpha$. It follows that $\bar{V} \subset \prod U_\alpha \subset U$, so that X is regular.

To prove the assertion, we show that if $A_\alpha \subset X_\alpha$ for each α, and if $A = \prod A_\alpha$, then $\bar{A} = \prod \bar{A}_\alpha$. (This was an exercise in §2-8.) Suppose that $\mathbf{y} = (y_\alpha)$ is in $\prod \bar{A}_\alpha$. Let $U = \prod U_\alpha$ be a basis element containing $\mathbf{y}$. Since $y_\alpha \in \bar{A}_\alpha$, the open set U_α must intersect A_α, so we can choose a point $z_\alpha \in U_\alpha \cap A_\alpha$, for each α. Then U intersects A in the point $\mathbf{z} = (z_\alpha)$. Thus $\mathbf{y}$ is in $\bar{A}$.

Conversely, suppose that $\mathbf{y}$ is in $\bar{A}$. We show that for any given index β, we have $y_\beta \in \bar{A}_\beta$. Let U_β be a neighborhood of y_β. Then $\pi_\beta^{-1}(U_\beta)$ is a neighborhood of $\mathbf{y}$, so that it intersects A in some point $\mathbf{z}$. Then U_β intersects $\pi_\beta(A) = A_\beta$ in the point $\pi_\beta(\mathbf{z})$. Thus y_β is in $\bar{A}_\beta$.

(c) Finding an example of a subspace of a normal space that is not nor-

mal is rather difficult. So is the problem of finding a product of normal spaces that is not normal. There happens to be a single space that will suffice for both purposes; it is given in Example 2. □

The following three theorems give three very important sets of hypotheses under which normality of a space is assured.

Theorem 2.3. *Every metrizable space is normal.*

Proof. Let X be a metrizable space with metric d. Let A and B be disjoint closed subsets of X. For each $a \in A$, choose ϵ_a so that the ball $B(a, \epsilon_a)$ does not intersect B. Similarly, for each b in B, choose ϵ_b so that the ball $B(b, \epsilon_b)$ does not intersect A. Define

$$U = \bigcup_{a \in A} B(a, \epsilon_a/2) \quad \text{and} \quad V = \bigcup_{b \in B} B(b, \epsilon_b/2).$$

Then U and V are open sets containing A and B, respectively; we assert they are disjoint. For if $z \in U \cap V$, then

$$z \in B(a, \epsilon_a/2) \cap B(b, \epsilon_b/2)$$

for some $a \in A$ and some $b \in B$. The triangle inequality applies to show that $d(a, b) < (\epsilon_a + \epsilon_b)/2$. If $\epsilon_a \leq \epsilon_b$, then $d(a, b) < \epsilon_b$, so that the ball $B(b, \epsilon_b)$ contains the point a. If $\epsilon_b \leq \epsilon_a$, then $d(a, b) < \epsilon_a$, so that the ball $B(a, \epsilon_a)$ contains the point b. Neither situation is possible. □

Theorem 2.4. *Every compact Hausdorff space is normal.*

Proof. Let X be a compact Hausdorff space. We have already essentially proved that X is regular. For if x is a point of X and B is a closed set in X not containing x, then B is compact, so that Lemma 5.4 of Chapter 3 applies to show there exist disjoint open sets about x and B, respectively.

Essentially the same argument as given in that lemma can be used to show that X is normal: Given disjoint closed sets A and B in X, choose, for each point a of A, disjoint open sets U_a and V_a containing a and B, respectively. (Here we use regularity of X.) The collection $\{U_a\}$ covers A; because A is compact, A may be covered by finitely many sets $U_{a_1}, \ldots, U_{a_n}$. Then

$$U = U_{a_1} \cup \cdots \cup U_{a_n} \quad \text{and} \quad V = V_{a_1} \cap \cdots \cap V_{a_n}$$

are disjoint open sets containing A and B, respectively. □

Theorem 2.5. *Every regular space with a countable basis is normal.*

Proof. Let X be a regular space with a countable basis $\mathcal{B}$. Let A and B be disjoint closed subsets of X. Each point x of A has a neighborhood U not intersecting B. Using regularity, choose a neighborhood V of x whose closure lies in U; finally, choose an element of $\mathcal{B}$ containing x and contained in V. By choosing such a basis element for each x in A, we construct a countable covering of A by open sets whose closures do not intersect B. Since

this covering of A is countable, we can index it with the positive integers; let us denote it by $\{U_n\}$.

Similarly, choose a countable collection $\{V_n\}$ of open sets covering B, such that each set $\bar{V}_n$ is disjoint from A.

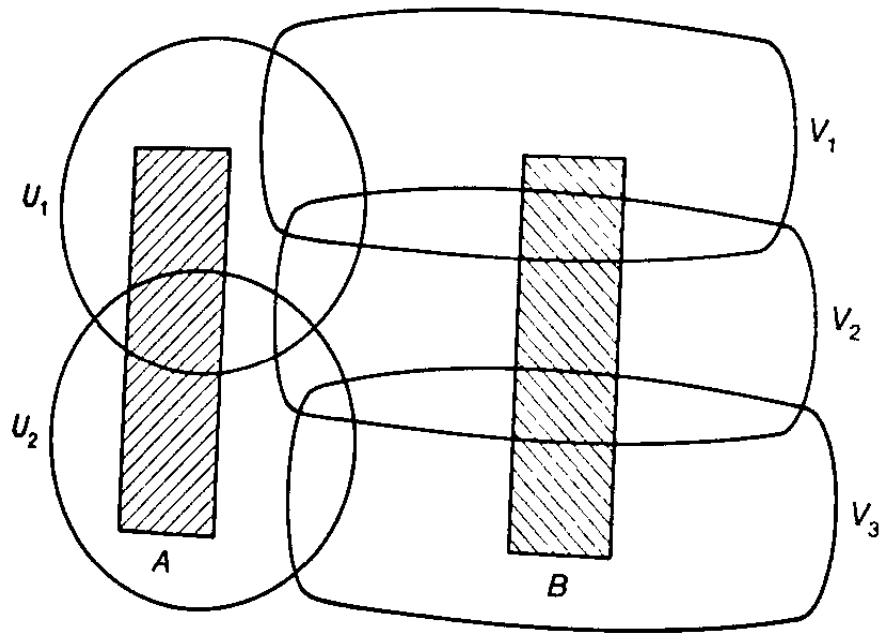
The sets $U = \bigcup U_n$ and $V = \bigcup V_n$ are open sets containing A and B, respectively, but they need not be disjoint. We perform the following simple trick to construct two open sets that *are* disjoint. Given n, define

$$U'_n = U_n - \bigcup_{i=1}^{n} \bar{V}_i \quad \text{and} \quad V'_n = V_n - \bigcup_{i=1}^{n} \bar{U}_i.$$

See Figure 3. Note that each set U'_n is open, being the difference of an open set U_n and a closed set $\bigcup_{i=1}^{n} \bar{V}_i$. Similarly, each set V'_n is open. The collection $\{U'_n\}$ covers A, because each x in A belongs to U_n for some n, and x belongs to *none* of the sets $\bar{V}_i$. Similarly, the collection $\{V'_n\}$ covers B.

Finally, the open sets

$$U' = \bigcup_{n \in Z_+} U'_n \quad \text{and} \quad V' = \bigcup_{n \in Z_+} V'_n$$



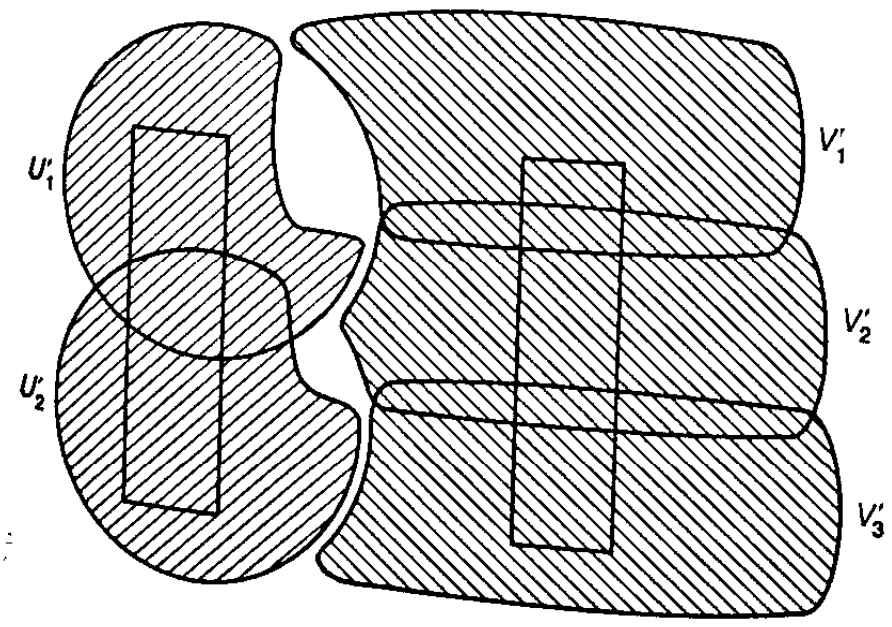


Figure 3

are disjoint. For if $x \in U' \cap V'$, then $x \in U'_j \cap V'_k$ for some j and k. Suppose that $j \leq k$. It follows from the definition of U'_j that $x \in U_j$; and since $j \leq k$ it follows from the definition of V'_k that $x \notin \bar{U}_j$. A similar contradiction arises if $j \geq k$. □

We shall use the following theorem in dealing with some examples:

Theorem 2.6. Every well-ordered set X *is normal in the order topology.*

It is, in fact, true that *every* order topology is normal (see Example 39 of [S-S]); but we shall not have occasion to use this stronger result.

Proof. Let X be a well-ordered set. We assert that every interval of the form $(x, y]$ is open in X: If X has a largest element and y is that element, $(x, y]$ is just a basis element about y. If y is not the largest element of X, then $(x, y]$ equals the open set (x, y'), where y' is the immediate successor of y.

Now let A and B be disjoint closed sets in X; assume for the moment that neither A nor B contains the smallest element a_0 of X. For each $a \in A$, there exists a basis element about a disjoint from B; it contains some interval of the form $(x, a]$. (Here is where we use the fact that a is not the smallest element of X.) Choose, for each $a \in A$, such an interval $(x_a, a]$ disjoint from B. Similarly, for each $b \in B$, choose an interval $(y_b, b]$ disjoint from A. The sets

$$U = \bigcup_{a \in A} (x_a, a] \quad \text{and} \quad V = \bigcup_{b \in B} (y_b, b]$$

are open sets containing A and B, respectively; we assert they are disjoint. For suppose that $z \in U \cap V$. Then $z \in (x_a, a] \cap (y_b, b]$ for some $a \in A$ and some $b \in B$. Assume that $a < b$. Then if $a \leq y_b$, the two intervals are disjoint, while if $a > y_b$, we have $a \in (y_b, b]$, contrary to the fact that $(y_b, b]$ is disjoint from A. A similar contradiction occurs if $b < a$.

Finally, assume that A and B are disjoint closed sets in X, and A contains the smallest element a_0 of X. The set $\{a_0\}$ is both open and closed in X. By the result of the preceding paragraph, there exist disjoint open sets U and V containing the closed sets $A - \{a_0\}$ and B, respectively. Then $U \cup \{a_0\}$ and V are disjoint open sets containing A and B, respectively. □

EXAMPLE 1. Here is an example showing the regularity axiom stronger than the Hausdorff axiom. Let K be the subset

$$K = \{1/n \mid n \in Z_+\}$$

of the real line R. Define a topology for R by taking as basis all sets of the following types:

(1) Every open interval (a, b).

(2) Every set $(a, b) - K$.

It is easy to check that this is a basis for a topology on R; the intersection of

§4-2

two basis elements is always another basis element or empty. The space is Hausdorff, because any two distinct points of R have disjoint open intervals about them.

But it is not regular. The set K is closed in this topology, and it does not contain the point 0. Suppose that there exist disjoint open sets U and V containing 0 and K, respectively. Choose a basis element containing 0 and lying in U. It must be a basis element $(a, b) - K$ of type (2), since each basis element of type (1) containing 0 intersects K. Then choose n large enough that $1/n \in (a, b)$. Choose a basis element about $1/n$ contained in V; it must be a basis element (c, d) of type (1). Finally, choose z so that $z < 1/n$ and $z > \max\{c, 1/n + 1\}$. Then z belongs to both U and V, so they are not disjoint. See Figure 4.

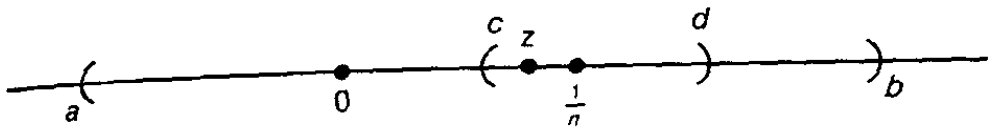


Figure 4

EXAMPLE 2. *The product space $S_\Omega \times \bar{S}_\Omega$ is not normal.*†

Consider the well-ordered set $\bar{S}_\Omega$, in the order topology, and consider the subset S_Ω, in the subspace topology (which is the same as the order topology). Both spaces are normal, by Theorem 2.6. We shall show that the product space $S_\Omega \times \bar{S}_\Omega$ is not normal.

This example serves three purposes. First, it shows that a regular space need not be normal, for $S_\Omega \times \bar{S}_\Omega$ is a product of regular spaces and therefore regular. Second, it shows that a subspace of a normal space need not be normal, for $S_\Omega \times \bar{S}_\Omega$ is a subspace of $\bar{S}_\Omega \times \bar{S}_\Omega$, which is a compact Hausdorff space and therefore normal. Third, it shows that the product of two normal spaces need not be normal.

First consider the space $\bar{S}_\Omega \times \bar{S}_\Omega$, and its "diagonal" $\Delta = \{x \times x \mid x \in \bar{S}_\Omega\}$. Because $\bar{S}_\Omega$ is Hausdorff, Δ is closed in $\bar{S}_\Omega \times \bar{S}_\Omega$: If U and V are disjoint neighborhoods of x and y, respectively, then $U \times V$ is a neighborhood of $x \times y$ that does not intersect Δ.

Therefore, in the subspace $S_\Omega \times \bar{S}_\Omega$, the set

$$A = \Delta \cap (S_\Omega \times \bar{S}_\Omega) = \Delta - \{\Omega \times \Omega\}$$

is closed. Likewise, the set

$$B = S_\Omega \times \{\Omega\}$$

is closed in $S_\Omega \times \bar{S}_\Omega$, being a "slice" of this product space. See Figure 5. The sets A and B are disjoint. We shall assume there exist disjoint open sets U and V of $S_\Omega \times \bar{S}_\Omega$ containing A and B, respectively, and derive a contradiction.

Given $x \in S_\Omega$, consider the vertical slice $x \times \bar{S}_\Omega$. We assert that there is some point β with $x < \beta < \Omega$ such that $x \times \beta$ lies outside U. For if U contained all points $x \times \beta$ for $x < \beta < \Omega$, then the top point $x \times \Omega$ of the

†Kelley [K] attributes this example to J. Dieudonné and A. P. Morse independently.

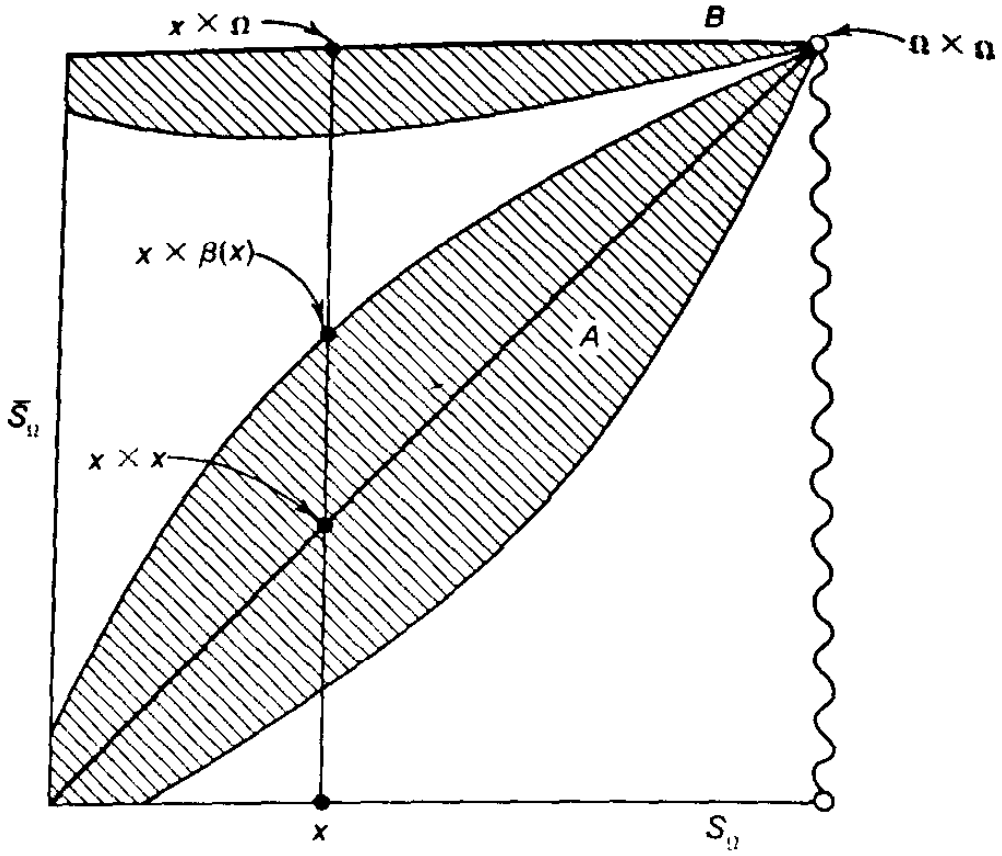


Figure 5

slice would be a limit point of U, which it is not because V is an open set disjoint from U containing this top point.

Choose $\beta(x)$ to be such a point: just to be definite, let $\beta(x)$ be the *smallest* element of S_Ω such that $x < \beta(x) < \Omega$ and $x \times \beta(x)$ lies outside U. Define a sequence of points of S_Ω as follows: Let x_1 be any point of S_Ω. Let $x_2 = \beta(x_1)$, and in general, $x_{n+1} = \beta(x_n)$. We have

$$x_1 < x_2 < \ldots,$$

because $\beta(x) > x$ for all x. The set $\{x_n\}$ is countable and therefore has an upper bound in S_Ω; let $b \in S_\Omega$ be its least upper bound. Because the sequence is increasing, it must converge to its least upper bound; thus $x_n \longrightarrow b$. But $\beta(x_n) = x_{n+1}$, so that $\beta(x_n) \longrightarrow b$ also. Then

$$x_n \times \beta(x_n) \longrightarrow b \times b$$

in the product space. See Figure 6. Now we have a contradiction, for the point $b \times b$ lies in the set A, which is contained in the open set U; and U contains *none* of the points $x_n \times \beta(x_n)$.

EXAMPLE 3. *The space R_l is normal, but the space R_l^2 is not.*

This example serves two purposes. It shows that a regular space R_l^2 need not be normal, and it shows that the product of two normal spaces need not be normal.

It is easy to see that R_l is normal. Suppose that A and B are disjoint closed sets in R_l. For each point a of A choose a basis element $[a, x_a)$ not intersecting

§4-2

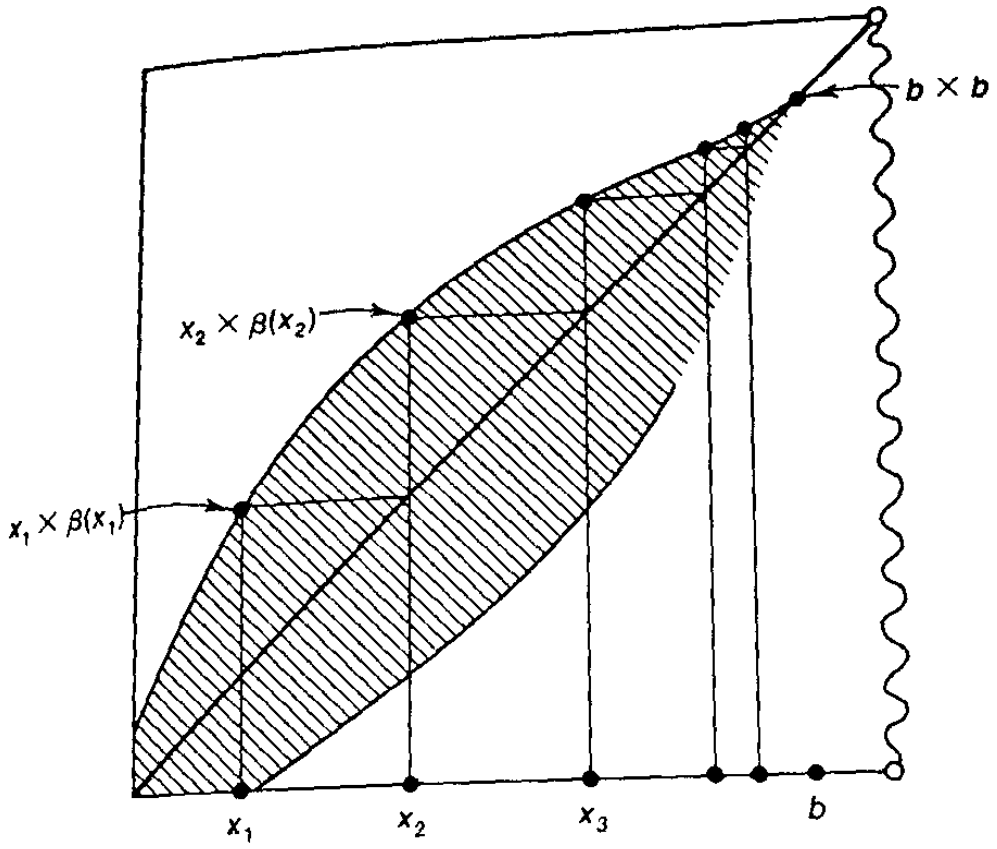


Figure 6

B; and for each point b of B choose a basis element $[b, x_b)$ not intersecting A. The open sets

$$U = \bigcup_{a \in A} [a, x_a) \quad \text{and} \quad V = \bigcup_{b \in B} [b, x_b)$$

are disjoint open sets about A and B, respectively.

It is much harder to show that R_l^2 is not normal. Let L be the line

$$L = \{x \times (-x) \mid x \in R_l\},$$

as usual. Then L is closed in R_l^2, and L has the discrete topology as a subspace of R_l^2. Therefore every subset A of L, being closed in L, is closed in R_l^2. If R_l^2 is normal, this means that for every nonempty proper subset A of L, one can choose disjoint open sets of R_l^2 containing the closed sets A and $L - A$ of R_l^2, respectively.

Using a simple "cardinality argument," one can prove that this is not possible; the proof is outlined in Exercise 14 below.

One specific subset of L for which it proves impossible to pick such open sets is the subset A consisting of points having irrational coordinates. The proof that this particular choice will work requires a strong tool: the special case of the Baire category theorem we gave as an exercise in §3-6. Assuming that exercise, we prove that this particular set will work.

So let A be the subset of L consisting of points having irrational coordinates. Suppose that U and V are disjoint open sets in R_l^2 containing A and $L - A$, respectively. For each point $x \times (-x)$ of A, there is a positive integer n such that the basis element

$$[x, x + 1/n) \times [-x, -x + 1/n)$$

about $x \times (-x)$ lies in U. Let K_n denote the set consisting of those irrational numbers x in $[0, 1]$ for which this basis element lies in U. Then $\bigcup K_n$ equals the set of all irrational numbers in $[0, 1]$.

Let $\bar{K}_n$ denote the closure of K_n in $[0, 1]$; then since $\bigcup \bar{K}_n$ contains $\bigcup K_n$, it contains all irrational numbers in $[0, 1]$. Therefore, we can write $[0, 1]$ as the countable union of the closed sets $\bar{K}_n$ (for $n \in Z_+$) and the one-point sets $\{q\}$ (for q a rational number in $[0, 1]$). By Exercise 5 of §3-6, at least one of these sets must have nonempty interior in $[0, 1]$. It cannot be one of the one-point sets $\{q\}$. Therefore, there is an n such that the set $\bar{K}_n$ contains an open interval (a, b) of R.

By definition, for each c in K_n the basis element

$$[c, c + 1/n) \times [-c, -c + 1/n)$$

for R_l^2 lies in U. We assert that the union of these basis elements, for $c \in K_n$, contains the entire "parallelogram region"

$$P = \{x \times y \mid a < x < b \text{ and } -x < y < -x + 1/n\},$$

so that P lies in U. See Figure 7.

To prove this fact, let $x_0 \times y_0$ be a point of P. Using the fact that $K_n \cap (a, b)$

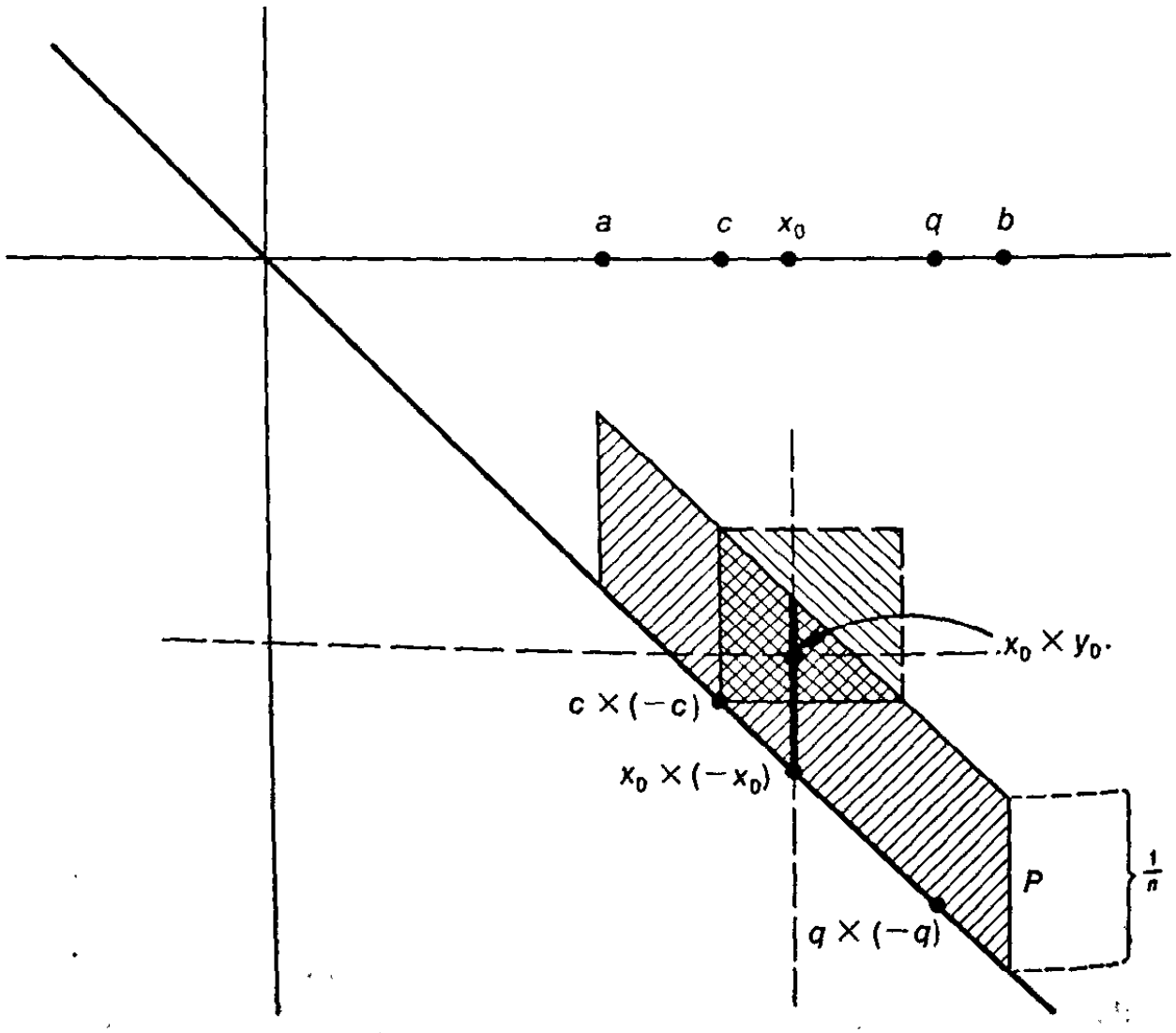


Figure 7

§4-2

is dense in (a, b), we can choose a point c of $K_n \cap (a, b)$ such that $c \times (-c)$ lies beneath the line $y = y_0$ and to the left of the line $x = x_0$. It is then simple to show that $x_0 \times y_0$ lies in the basis element $[c, c + 1/n) \times [-c, -c + 1/n)$. [Formally, choose c to be a point of K_n in the interval $(\max\{a, -y_0\}, x_0)$. Then we have the inequalities

$$c < x_0 \quad \text{and} \quad -c < y_0 < -x_0 + 1/n.$$

These incqualities imply that $c < x_0 < c + 1/n$ and $-c < y_0 < -c + 1/n$.]

The rest is easy. Take any rational number q of the interval (a, b). It is clear that the point $q \times (-q)$ is a limit point of the parallelogram region P, in the topology of R_l^2. Then because $P \subset U$, this point is a limit point of U. This contradicts the fact that V is an open set in R_l^2 containing $q \times (-q)$ and disjoint from U.

EXAMPLE 4. *If J is uncountable, the product space R^J is not normal.* The proof is fairly difficult; we leave it as a challenging exercise (see Exercise 15).

This example serves three purposes. It shows that a regular space R^J need not be normal. It shows that a subspace of a normal space need not be normal, for R^J is homeomorphic to the subspace $(0, 1)^J$ of $[0, 1]^J$, which (assuming the Tychonoff theorem) is compact Hausdorff and therefore normal. And it shows that an uncountable product of normal spaces need not be normal. It leaves unsettled the question as to whether a finite or a countable product of normal spaces might be normal. For that we need one of the two preceding examples.

Exercises

1. Show that if X is regular, every pair of points of X have neighborhoods whose closures are disjoint.
2. Show that if X is normal, every pair of disjoint closed sets have neighborhoods whose closures are disjoint.
3. Show that a closed subspace of a normal space is normal.
4. Show that every order topology is regular.
5. Show that if $\prod X_\alpha$ is Hausdorff, or regular, or normal, then so is X_α. (Assume that each X_α is nonempty.)
6. Let X and X' denote a single set under two topologies $\mathcal{T}$ and $\mathcal{T}'$, respectively; assume that $\mathcal{T}' \supset \mathcal{T}$. If one of the spaces is Hausdorff (or regular, or normal), what does that imply about the other?
7. Show that every locally compact Hausdorff space is regular.
8. Show that every regular Lindelöf space is normal.
9. Let $f, g : X \to Y$ be continuous; assume that Y is Hausdorff. Show that $\{x \mid f(x) = g(x)\}$ is closed in X.
10. Is R^ω normal in the product topology? In the uniform topology?

It is not known whether R^ω is normal in the box topology. Mary-Ellen Rudin has shown that the answer is affirmative if one assumes the continuum hypothesis [*R*]. In fact, she shows it satisfies a stronger condition called *paracompactness.*

11. A space X is said to be **completely normal** if every subspace of X is normal. Show that X is completely normal if and only if for every pair A, B of separated sets in X (that is, sets such that $\bar{A} \cap B = \varnothing$ and $A \cap \bar{B} = \varnothing$), there exist disjoint open sets containing them. [*Hint:* If X is completely normal, consider $X - (\bar{A} \cap \bar{B})$.]

12. Which of the following spaces are completely normal? Justify your answers.
(a) A subspace of a completely normal space.
(b) The product of two completely normal spaces.
(c) A well-ordered set in the order topology.
(d) A metrizable space.
(e) A compact Hausdorff space.
(f) A regular space with a countable basis.
(g) The space R_l.

13. Let $p: X \to Y$ be a quotient map. Show that if p is closed and X is normal, then Y is normal. (Classically, a partition X^* is said to be *upper semicontinuous* if the quotient map $p: X \to X^*$ is closed.)

***14.** Show that R_l^2 is not normal, as follows: Suppose that for each nonempty proper subset A of L, there exist disjoint open sets U_A and V_A of R_l^2 containing A and $L - A$, respectively. Let D be the set of points of R_l^2 having rational coordinates. Let $\mathcal{P}(L)$ and $\mathcal{P}(D)$ denote the collections of all subsets of L and D, respectively. Define $\theta: \mathcal{P}(L) \to \mathcal{P}(D)$ as follows:

$$\theta(A) = U_A \cap D \quad \text{if } A \neq \varnothing \text{ and } A \neq L,$$
$$\theta(\varnothing) = \varnothing,$$
$$\theta(L) = D.$$

(a) Show that θ is injective.
(b) Construct an injective map $\mathcal{P}(D) \to L$.
(c) Derive a contradiction.

***15.** *Theorem. If J is uncountable, then R^J is not normal.*

Proof. (This proof is due to A. H. Stone, as adapted in [S-S].) It will suffice to show that $(Z_+)^J$ is not normal, since $(Z_+)^J$ is closed in R^J. Let X denote the space $(Z_+)^J$; use functional notation for the elements of X.

(a) Define $P_1 \subset X$ to be the set of those functions $\mathbf{x}: J \to Z_+$ such that for each $i \neq 1$ the set $\mathbf{x}^{-1}(i)$ contains at most one element. Define P_2 to be the set of those $\mathbf{x}$ such that for each $i \neq 2$, the set $\mathbf{x}^{-1}(i)$ contains at most one element. Show that P_1 and P_2 are disjoint, and are closed in X.

(b) Assume that U and V are disjoint open sets containing P_1 and P_2, respectively. Show that you can choose a sequence

$$\alpha_1, \alpha_2, \ldots$$

of distinct elements of J, and a sequence of integers

$$n_0 = 0 < n_1 < n_2 < \cdots,$$

§4-3

such that for each positive integer i, the set U_i defined below is contained in U. Here U_i is the set of all $\mathbf{x}$ such that

$$\mathbf{x}(\alpha_j) = \begin{cases} j & \text{for } 1 \leq j \leq n_{i-1}, \\ 1 & \text{for } n_{i-1} < j \leq n_i. \end{cases}$$

(c) Now let $A = \{\alpha_j | j \in Z_+\}$. Show you can choose a finite subset B of J such that the set V_B defined below is contained in V. Here V_B is the set of all those $\mathbf{x}$ such that

$$\mathbf{x}(\alpha_j) = j \quad \text{for } \alpha_j \in B \cap A,$$
$$\mathbf{x}(\alpha) = 2 \quad \text{for } \alpha \in B - A.$$

(d) Show that for some i, $U_i \cap V_B \neq \emptyset$.

16. Is every regular topological group normal?

4-3 The Urysohn Lemma

Now we come to the first deep theorem of the book, a theorem that is commonly called the "Urysohn lemma." It asserts the existence of certain real-valued continuous functions on a normal space X. It is the crucial tool used in proving a number of important theorems. One of these, the Tietze extension theorem, we prove in this section. Another, the Urysohn metrization theorem, we prove in the next section. And yet another, an imbedding theorem for manifolds, appears in the last section of the chapter.

Why do we call the Urysohn lemma a "deep" theorem? Because its proof involves a really original idea, which the previous proofs did not. Perhaps we can explain what we mean this way: By and large, one would expect that if one went through this book and deleted all the proofs we have given up to now and then handed the book to a bright student who had not studied topology, that student ought to be able to go through the book and work out the proofs by himself. (It would take a good deal of time and effort, of course; and one would not expect him to handle the trickier examples.) But the Urysohn lemma is on a different level. It would take considerably more originality than most of us possess to prove this lemma unless we were given copious hints!

Theorem 3.1 (*Urysohn lemma*). *Let* X *be a normal space; let* A *and* B *be disjoint closed subsets of* X. *Let* [a, b] *be a closed interval in the real line. Then there exists a continuous map*

$$f : X \longrightarrow [a, b]$$

such that $f(x) = a$ *for every* x *in* A, *and* $f(x) = b$ *for every* x *in* B.

Proof. We need only consider the case where the interval in question is the interval [0, 1]; the general case follows from that one.

The first step of the proof is to construct, using normality, a certain family

U_p of open sets of X, indexed by the rational numbers. Then one uses these sets to define the continuous function f.

Step 1. Let P be the set of all rational numbers in the interval $[0, 1]$.† We shall define, for each p in P, an open set U_p of X, in such a way that whenever $p < q$, we have

$$\bar{U}_p \subset U_q.$$

Thus the sets U_p will be simply ordered by inclusion in the same way their subscripts are ordered by the usual ordering in the real line.

Because P is countable, we can use induction to define the sets U_p (or rather, the principle of recursive definition). Arrange the elements of P in an infinite sequence in some way; for convenience, let us suppose that the numbers 1 and 0 are the first two elements of the sequence.

Now define the sets U_p as follows: First, define $U_1 = X - B$. Second, since A is a closed set contained in the open set U_1, we may by normality of X choose an open set U_0 such that

$$A \subset U_0 \quad \text{and} \quad \bar{U}_0 \subset U_1.$$

In general, let P_n denote the set consisting of the first n rational numbers in the sequence. Suppose that U_p is defined for all rational numbers p belonging to the set P_n, satisfying the condition

$$(*) \qquad p < q \Longrightarrow \bar{U}_p \subset U_q.$$

Let r denote the next rational number in the sequence; we wish to define U_r.

Consider the set $P_{n+1} = P_n \cup \{r\}$. It is a finite subset of the interval $[0, 1]$, and, as such, it has a simple ordering derived from the usual order relation $<$ on the real line. In a finite simply ordered set, every element (other than the smallest and the largest) has an immediate predecessor and an immediate successor. (See Theorem 10.1 of Chapter 1.) The number 0 is the smallest element, and 1 is the largest element, of the simply ordered set P_{n+1}, and r is neither 0 nor 1. So r has an immediate predecessor p in P_{n+1} and an immediate successor q in P_{n+1}. The sets U_p and U_q are already defined, and $\bar{U}_p \subset U_q$ by the induction hypothesis. Using normality of X, we can find an open set U_r of X such that

$$\bar{U}_p \subset U_r \quad \text{and} \quad \bar{U}_r \subset U_q.$$

We assert that $(*)$ now holds for every pair of elements of P_{n+1}. If both elements lie in P_n, $(*)$ holds by the induction hypothesis. If one of them is r and the other is a point s of P_n, then either $s \leq p$, in which case

$$\bar{U}_s \subset \bar{U}_p \subset U_r,$$

†Actually, any countable dense subset of $[0, 1]$ will do, providing it contains the points 0 and 1.

§4-3

or $s \geq q$, in which case

$$\bar{U}_r \subset U_q \subset U_s.$$

Thus for every pair of elements of P_n, relation (*) holds.

By induction, we have U_p defined for all $p \in P$.

To illustrate, let us suppose we started with the standard way of arranging the elements of P in an infinite sequence:

$$P = \{1, 0, \tfrac{1}{2}, \tfrac{1}{3}, \tfrac{2}{3}, \tfrac{1}{4}, \tfrac{3}{4}, \tfrac{1}{5}, \tfrac{2}{5}, \tfrac{3}{5}, \ldots\}$$

After defining U_0 and U_1, we would define $U_{1/2}$ so that $\bar{U}_0 \subset U_{1/2}$ and $\bar{U}_{1/2} \subset U_1$. Then we would fit in $U_{1/3}$ between U_0 and $U_{1/2}$; and $U_{2/3}$ between $U_{1/2}$ and U_1. And so on. At the eighth step of the proof we would have the situation pictured in Figure 8. And the ninth step would consist of choosing an open set $U_{2/5}$ to fit in between $U_{1/3}$ and $U_{1/2}$. And so on.

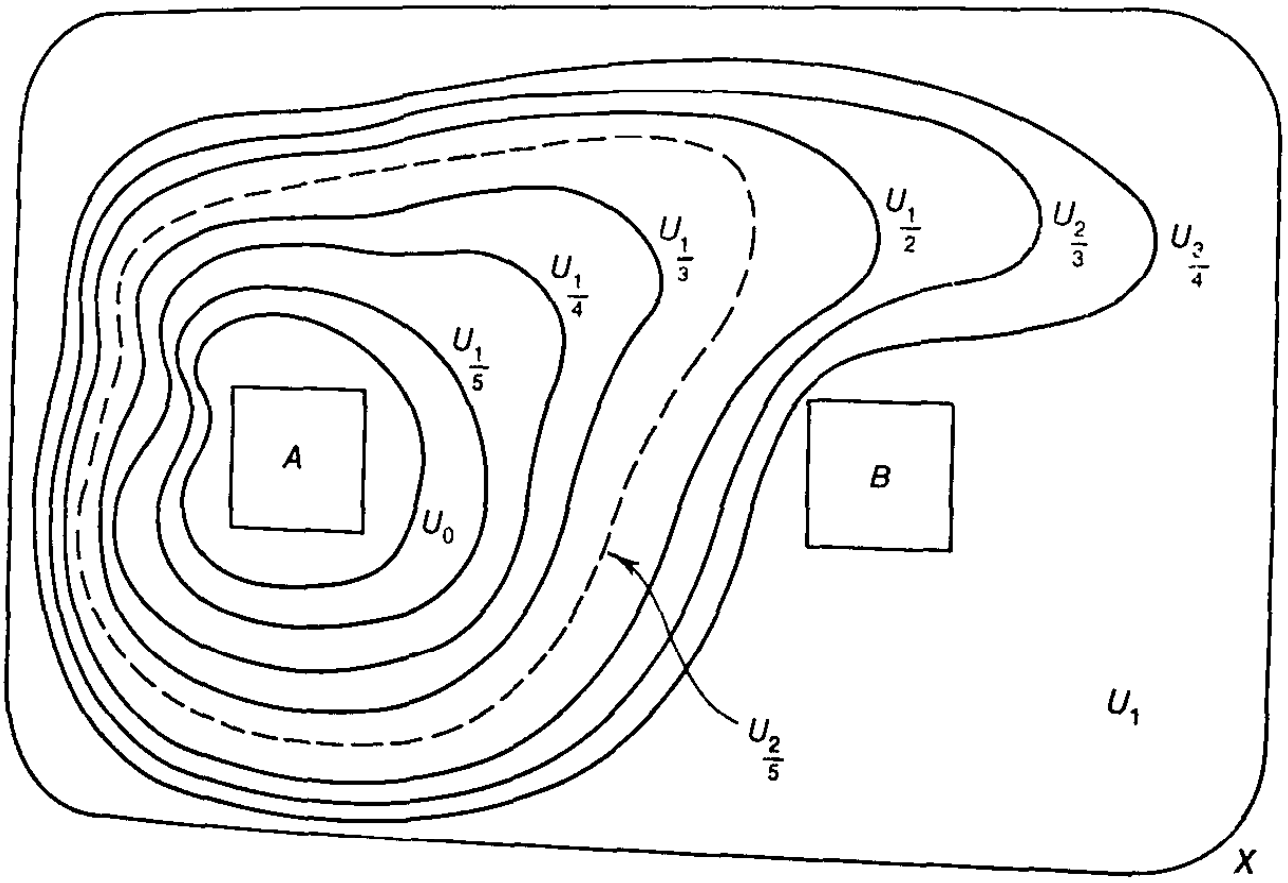


Figure 8

Step 2. Now we have defined U_p for all rational numbers p in the interval $[0, 1]$. We extend this definition to all rational numbers p in R by defining

$$U_p = \varnothing \quad \text{if } p < 0,$$

$$U_p = X \quad \text{if } p > 1.$$

It is still true that for any pair of rational numbers p and q,

$$p < q \Longrightarrow \bar{U}_p \subset U_q,$$

as you can check.

Step 3. Given a point x of X, let us define $Q(x)$ to be the set of those rational numbers p such that the corresponding open sets U_p contain x:

$$Q(x) = \{p \mid x \in U_p\}.$$

This set contains no number less than 0, since no x is in U_p for $p < 0$. And it contains every number greater than 1, since every x is in U_p for $p > 1$. Therefore, $Q(x)$ is bounded below, and its greatest lower bound is a point of the interval [0, 1]. Define

$$f(x) = \text{glb } Q(x) = \text{glb } \{p \mid x \in U_p\}.$$

Step 4. We show that f is the desired function. If $x \in A$, then $x \in U_p$ for every $p \geq 0$, so that $Q(x)$ equals the set of all nonnegative rationals, and $f(x) = \text{glb } Q(x) = 0$. Similarly, if $x \in B$, then $x \in U_p$ for no $p \leq 1$, so that $Q(x)$ consists of all rational numbers greater than 1, and $f(x) = 1$.

All this is easy. The only hard part is to show that f is continuous. For this purpose, we first prove the following elementary facts:

(1) $x \in \bar{U}_r \Rightarrow f(x) \leq r$.

(2) $x \notin U_r \Rightarrow f(x) \geq r$.

To prove (1), note that if $x \in \bar{U}_r$, then $x \in U_s$ for every $s > r$. Therefore, $Q(x)$ contains all rational numbers greater than r, so that by definition we have

$$f(x) = \text{glb } Q(x) \leq r.$$

To prove (2), note that if $x \notin U_r$, then x is not in U_s for any $s < r$. Therefore, $Q(x)$ contains no rational numbers less than r, so that

$$f(x) = \text{glb } Q(x) \geq r.$$

Now we prove continuity of f. Given a point x_0 of X and an open interval (c, d) in R containing the point $f(x_0)$, we wish to find a neighborhood U of x_0 such that $f(U) \subset (c, d)$. Choose rational numbers p and q such that

$$c < p < f(x_0) < q < d.$$

We assert that the open set

$$U = U_q - \bar{U}_p$$

is the desired neighborhood of x_0. See Figure 9.

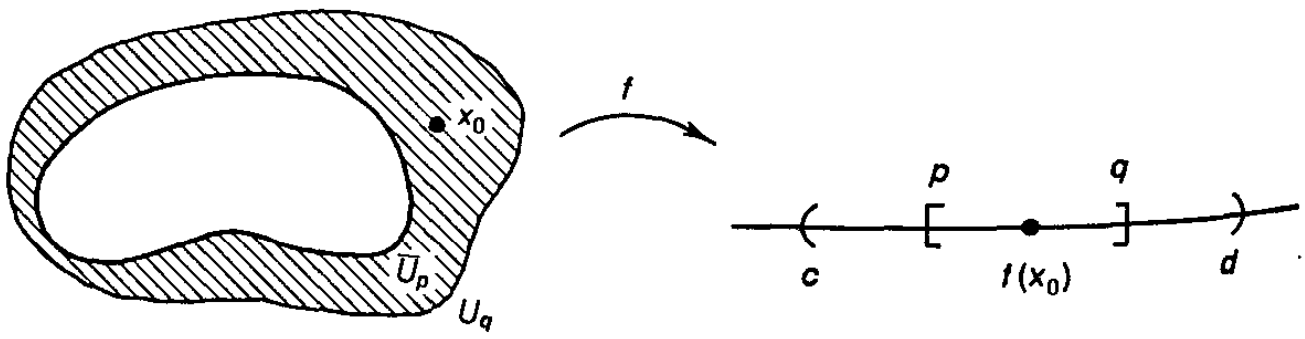


Figure 9

§4-3

First, we show that $x_0 \in U$. Necessarily $x_0 \in U_q$, because $x_0 \notin U_q$ implies by (2) that $f(x_0) \geq q$. Also, $x_0 \notin \bar{U}_p$, because $x_0 \in \bar{U}_p$ implies by (1) that $f(x) \leq p$. Hence $x_0 \in U$.

Second, we show that $f(U) \subset (c, d)$. Let $x \in U$. Then $x \in U_q \subset \bar{U}_q$, so that $f(x) \leq q$, by (1). And $x \notin \bar{U}_p$, so that $x \notin U_p$ and $f(x) \geq p$, by (2). Thus $f(x) \in [p, q] \subset (c, d)$, as desired. □

Definition. If A and B are two subsets of the topological space X, and if there is a continuous function $f: X \to [0, 1]$ such that $f(A) = \{0\}$ and $f(B) = \{1\}$, we say that A and B **can be separated by a continuous function.**

The Urysohn lemma says that if every pair of disjoint closed sets in X can be separated by disjoint open sets, then each such pair can be separated by a continuous function. The converse is trivial, for if $f: X \to [0, 1]$ is the function, then $f^{-1}([0, \frac{1}{2}))$ and $f^{-1}((\frac{1}{2}, 1])$ are disjoint open sets containing A and B, respectively.

This fact leads to a question that may already have occurred to you: Why cannot the proof of the Urysohn lemma be generalized to show that in a regular space, where you can separate points from closed sets by disjoint open sets, you can also separate points from closed sets by continuous functions?

At first glance, it seems that the proof of the Urysohn lemma should go through. You take a point a and a closed set B not containing a, and you begin the proof just as before by defining $U_1 = X - B$ and choosing U_0 to be an open set about a whose closure is contained in U_1 (using regularity of X). But at the very next step of the proof, you run into difficulty. Suppose that p is the next rational number in the sequence after 0 and 1. You want to find an open set U_p such that $\bar{U}_0 \subset U_p$ and $\bar{U}_p \subset U_1$. For this regularity is not enough.

Requiring that one be able to separate a point from a closed set by a continuous function is, in fact, a stronger condition than requiring that one can separate them by disjoint open sets. We make this requirement into a new separation axiom, which we shall study further in Chapter 5: A space X is said to be **completely regular** if one-point sets are closed in X, and if for every point a of X and every closed set B not containing a, there exists a continuous function $f: X \to [0, 1]$ such that $f(a) = 0$ and $f(B) = \{1\}$.

Let us make another comment on the Urysohn lemma: Note that the Urysohn lemma does not assert the existence of a continuous function f such that $f(x) = 0$ *only* for $x \in A$, and $f(x) = 1$ *only* for $x \in B$. One cannot in general choose f so that this condition is satisfied; see Exercise 5.

One immediate corollary of the Urysohn lemma is the useful theorem called the *Tietze extension theorem*. It states that any continuous function

f mapping a closed subset of a normal space into the reals may be extended to a continuous map of the entire space into the reals. Although we shall not use the Tietze theorem in this book (except in the exercises), it is very important in the applications of topology.

Theorem 3.2 (Tietze extension theorem). *Let* X *be a normal space; let* A *be a closed subset of* X.

(a) *Any continuous map of* A *into the closed interval* [a, b] *of* R *may be extended to a continuous map of all of* X *into* [a, b].

(b) *Any continuous map of* A *into the reals* R *may be extended to a continuous map of all of* X *into* R.

Proof. The idea of the proof is to construct a sequence of continuous functions s_n defined on the entire space X, such that the sequence s_n converges uniformly, and such that the restriction of s_n to A approximates f more and more closely as n becomes large. Then the limit function will be continuous, and its restriction to A will equal f.

Step 1. The first step is to construct a particular function g defined on all of X such that g is not too large, and such that g approximates f on the set A to a fair degree of accuracy.

To be more precise, let us take the case $f : A \to [-r, r]$. We assert that there exists a continuous function $g : X \to R$ such that

$$|g(x)| \leq \tfrac{1}{3}r \quad \text{for all } x \in X,$$
$$|g(a) - f(a)| \leq \tfrac{2}{3}r \quad \text{for all } a \in A.$$

The function g is constructed as follows:

Divide the interval $[-r, r]$ into three equal intervals of length $\frac{2}{3}r$:

$$I_1 = [-r, -\tfrac{1}{3}r], \qquad I_2 = [-\tfrac{1}{3}r, \tfrac{1}{3}r], \qquad I_3 = [\tfrac{1}{3}r, r].$$

Let B and C be the subsets

$$B = f^{-1}(I_1) \quad \text{and} \quad C = f^{-1}(I_3)$$

of A. Because f is continuous, B and C are closed disjoint subsets of A. Therefore, they are closed in X. By the Urysohn lemma, there exists a continuous function

$$g : X \to [-\tfrac{1}{3}r, \tfrac{1}{3}r]$$

having the property that $g(x) = -\frac{1}{3}r$ for each x in B, and $g(x) = \frac{1}{3}r$ for each x in C. See Figure 10.

Then $|g(x)| \leq \frac{1}{3}r$ for all x. We assert that for each a in A,

$$|g(a) - f(a)| \leq \tfrac{2}{3}r.$$

There are three cases. If $a \in B$, then both $f(a)$ and $g(a)$ belong to I_1. If $a \in C$, then $f(a)$ and $g(a)$ are in I_3. And if $a \notin B \cup C$, then $f(a)$ and $g(a)$ are in I_2. In each case, $|g(a) - f(a)| \leq \frac{2}{3}r$.

§4-3

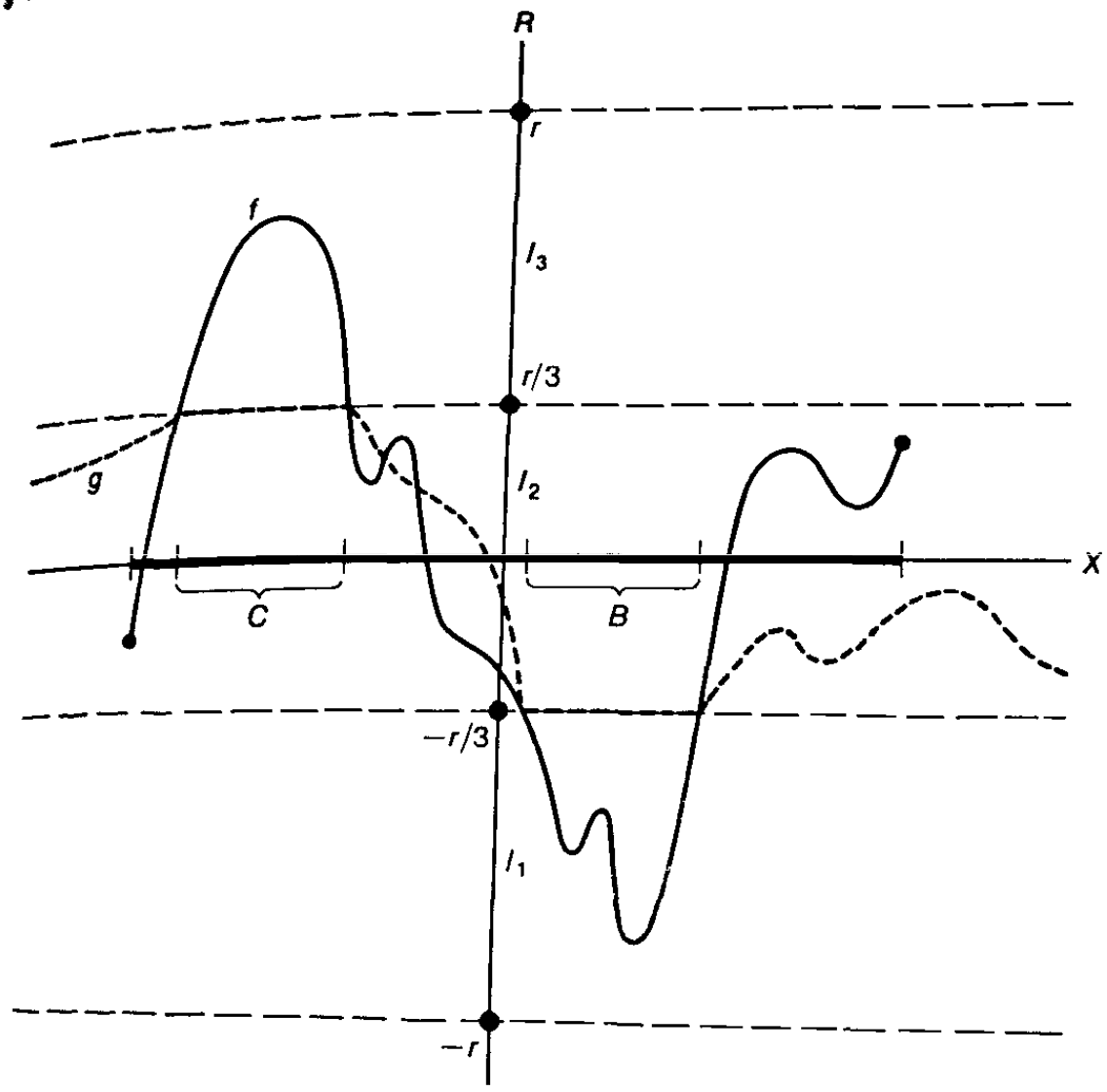


Figure 10

Step 2. We now prove the Tietze theorem in the case $f : A \to [-1, 1]$. Part (a) of the theorem, where f maps A into an arbitrary closed interval $[a, b]$, follows at once.

First, we apply Step 1 to the function $f : A \to [-1, 1]$. Here $r = 1$. We obtain a real-valued function g_1 defined on all of X such that

$$|g_1(x)| \leq \tfrac{1}{3} \quad \text{for } x \in X,$$
$$|f(a) - g_1(a)| \leq \tfrac{2}{3} \quad \text{for } a \in A.$$

We then consider the function $f - g_1$. Letting $r = \frac{2}{3}$, we see that $f - g_1$ maps A into the interval $[-r, r]$; applying Step 1, we obtain a real-valued function g_2 defined on all of X such that

$$|g_2(x)| \leq \tfrac{1}{3}(\tfrac{2}{3}) \quad \text{for } x \in X,$$
$$|f(a) - g_1(a) - g_2(a)| \leq (\tfrac{2}{3})^2 \quad \text{for } a \in A.$$

Then we apply Step 1 to the function $f - g_1 - g_2$. And so on.

At the general step, we have real-valued functions $g_1, \ldots, g_n$ defined on

all of X such that

$$|f(a) - g_1(a) - \cdots - g_n(a)| \leq (\tfrac{2}{3})^n$$

for $a \in A$. Applying Step 1 to the function $f - g_1 - \cdots - g_n$, with $r = (\frac{2}{3})^n$, we obtain a real-valued function g_{n+1} defined on all of X such that

$$|g_{n+1}(x)| \leq \tfrac{1}{3}(\tfrac{2}{3})^n \quad \text{for } x \in X,$$

$$|f(a) - g_1(a) - \cdots - g_{n+1}(a)| \leq (\tfrac{2}{3})^{n+1} \quad \text{for } a \in A.$$

By induction, the functions g_n are defined for all n.

We now define

$$g(x) = \textstyle\sum_{n=1}^{\infty} g_n(x)$$

for all x in X. Of course, we have to know that this infinite series converges. But that follows from the comparison theorem of calculus; it converges by comparison with the geometric series

$$\tfrac{1}{3} \textstyle\sum_{n=1}^{\infty} (\frac{2}{3})^{n-1}$$

Since this geometric series converges to 1 (as you can check), it follows that $|g(x)| \leq 1$, so that g maps X into $[-1, 1]$, as desired.

We assert that $g(a) = f(a)$ for $a \in A$. Let $s_n(x) = \sum_{i=1}^{n} g_i(x)$, the nth partial sum of the series. Then $g(x)$ is by definition the limit of the infinite sequence $s_n(x)$ of partial sums. Since

$$|f(a) - \textstyle\sum_{i=1}^{n} g_i(a)| = |f(a) - s_n(a)| \leq (\tfrac{2}{3})^n$$

for all a in A, it follows that $s_n(a) \longrightarrow f(a)$ for all $a \in A$. Therefore, we have $f(a) = g(a)$ for $a \in A$.

To show that g is continuous, we have to show that the sequence s_n converges to g uniformly. This fact follows at once from the "Weierstrass M-test" of analysis. Without assuming this result, one can simply note that if $k > n$, then

$$\begin{aligned} |s_k(x) - s_n(x)| &= |\textstyle\sum_{i=n+1}^{k} g_i(x)| \\ &\leq \tfrac{1}{3} \textstyle\sum_{i=n+1}^{k} (\frac{2}{3})^{i-1} \\ &< \tfrac{1}{3} \textstyle\sum_{i=n+1}^{\infty} (\frac{2}{3})^{i-1} = (\frac{2}{3})^n. \end{aligned}$$

Holding n fixed and letting $k \longrightarrow \infty$, we see that

$$|g(x) - s_n(x)| \leq (\tfrac{2}{3})^n$$

for all $x \in X$. Hence s_n converges to g uniformly.

Step 3. We now prove part (b) of the theorem, in which f maps A into R. We can replace R by the open interval $(-1, 1)$, since this interval is homeomorphic to R.

So let f be a continuous map from A into $(-1, 1)$. The half of the Tietze theorem already proved shows that we can extend f to a continuous map $g : X \longrightarrow [-1, 1]$ mapping X into the *closed* interval. How can we find a map h carrying X into the *open* interval?

§ 4-3

Given g, let us define a subset D of X by the equation

$$D = g^{-1}(\{-1\}) \cup g^{-1}(\{1\}).$$

Since g is continuous, D is a closed subset of X. Because $g(A) = f(A)$, which is contained in $(-1, 1)$, the set A is disjoint from D. By the Urysohn lemma, there is a continuous function $\phi: X \to [0, 1]$ such that $\phi(D) = \{0\}$ and $\phi(A) = \{1\}$. Define

$$h(x) = \phi(x)g(x).$$

Then h is continuous, being the product of two continuous functions. Also, h is an extension of f, since for a in A,

$$h(a) = \phi(a)g(a) = 1 \cdot g(a) = f(a).$$

Finally, h maps all of X into the open interval $(-1, 1)$. For if $x \in D$, then $h(x) = 0 \cdot g(x) = 0$. And if $x \notin D$, then $|g(x)| < 1$; it follows at once that $|h(x)| \leq 1 \cdot |g(x)| < 1$. □

EXAMPLE 1. The hypothesis that the sets involved are closed is needed in the Urysohn lemma and the Tietze theorem. For example, the sets

$$A = (0, 1) \quad \text{and} \quad B = (1, 2)$$

are disjoint subsets of the normal space R, but there is no continuous function $f: R \to [0, 1]$ that carries A into 0 and B into 1. Where could f carry the point 1 of $\bar{A} \cap \bar{B}$?

Exercises

1. Examine the proof of the Urysohn lemma, and show that for given r,

$$f^{-1}(r) = \bigcap_{p>r} U_p - \bigcup_{q<r} U_q,$$

p, q rational.

2. Show that the Tietze extension theorem implies the Urysohn lemma.

3. (a) Show that a connected normal space having more than one point is uncountable.
   (b) Show that a connected regular space having more than one point is uncountable.† [*Hint:* Any countable space is Lindelöf.]

4. Let X be a regular space with a countable basis; let U be open in X.
   (a) Show that U equals a countable union of closed sets of X.
   (b) Show there is a continuous function $f: X \to [0, 1]$ such that $f(x) > 0$ for $x \in U$ and $f(x) = 0$ for $x \notin U$.

5. Recall that A is a "G_δ set" in X if A is the intersection of a countable collection of open sets of X.

   Theorem (Strong form of the Urysohn lemma). Let A and B be closed disjoint

†Surprisingly enough, there does exist a connected *Hausdorff* space that is countably infinite. See Example 75 of [S-S].

subsets of the normal space X. There exists a continuous function $f: X \to [0, 1]$ such that $f^{-1}(\{0\}) = A$ and $f(B) = \{1\}$ if and only if A is a G_δ set in X.

6. Let X be metrizable. Show that the following are equivalent:
(i) X is bounded under every metric that gives the topology of X.
(ii) Every continuous function $\phi : X \to R$ is bounded.
(iii) X is compact.
[*Hint:* If ϕ is not bounded, map X into $X \times R$ by the map $x \to x \times \phi(x)$. If (x_n) is a sequence of distinct points having no convergent subsequence, find a continuous function ϕ with $\phi(x_n) = n$.]

7. Let Z be a topological space. If Y is a subspace of Z, we say that Y is a retract of Z if there is a continuous map $r : Z \to Y$ such that $r(y) = y$ for each $y \in Y$.
(a) Show that if Z is Hausdorff and Y is a retract of Z, then Y is closed in Z.
(b) Let A be a two-point set in R^2. Show that A is not a retract of R^2.
(c) Let S^1 be the unit circle in R^2; show that S^1 is a retract of $R^2 - \{0\}$, where 0 is the origin. Can you conjecture whether or not S^1 is a retract of R^2?

8. A space Y is said to have the universal extension property if for each triple consisting of a normal space X, a closed subset A of X, and a continuous function $f: A \to Y$, there exists an extension of f to a continuous map of X into Y.
(a) Show that R^J has the universal extension property.
(b) Show that if Y is homeomorphic to a retract of R^J, then Y has the universal extension property.

9. Let $X_1 \subset X_2 \subset \cdots$ be a sequence of spaces, where X_i is a closed subspace of X_{i+1} for each i. Let $X = \bigcup X_i$. Suppose that we define a subset U of X to be open in X if $U \cap X_i$ is open in X_i for each i. The topology we obtain is called the topology on X coherent with the spaces X_i.
(a) Show that X_i is a subspace of X in this topology.
(b) Show that if each space X_i is normal, so is the space X. [*Hint:* Use the Tietze theorem.]

***10.** The subspace $\bar{R}^n$ of R^ω, consisting of all sequences $(x_1, x_2, \ldots)$ such that $x_i = 0$ for $i > n$, has a standard topology. Suppose that we give the set $R^\infty = \bigcup \bar{R}^n$ the topology coherent with the subspaces $\bar{R}^n$. Compare this topology on R^∞ with the topology it inherits from the box topology on R^ω.

4-4 The Urysohn Metrization Theorem

Now we come to the major goal of this chapter, a theorem that gives us conditions under which a topological space is metrizable. The proof weaves together a number of strands from previous parts of the book; it uses results on metric topologies from Chapter 2 as well as facts concerning the countability and separation axioms proved in the present chapter. The basic construction used in the proof is a simple one, but very useful. You will see it several times more in this book, in various guises.

There are two versions of the proof, and since each has useful generali-

zations that will appear subsequently, we present both of them here. The first version is the one we shall use in Chapter 5 to prove an imbedding theorem for completely regular spaces. The second version will be generalized in Chapter 6 when we prove the Nagata–Smirnov metrization theorem.

***Theorem 4.1* (*Urysohn metrization theorem*).** *Every regular space* X *with a countable basis is metrizable.*

Proof. We shall prove that X is metrizable by imbedding X in a metrizable space Y; that is, by showing X homeomorphic with a subspace of Y. The two versions of the proof differ in the choice of the metrizable space Y. In the first version, Y is the space R^ω in the product topology, a space that we have previously proved to be metrizable (Theorem 9.5 of Chapter 2). In the second version, the space Y is also R^ω, but this time in the topology given by the uniform metric $\bar{\rho}$ (see §2-9). In each case, it turns out that our proof actually imbeds X in $[0, 1]^\omega$. Let $\{B_n\}$ be a countable basis for X.

Step 1. We prove the following: *There exists a countable collection of continuous functions $f_n : X \to [0, 1]$ having the property that given a point x_0 of X and given a neighborhood U of x_0, there exists an index n such that f_n is positive at x_0 and vanishes outside U.*

The construction of the functions f_n is simple. We apply Exercise 4 of the preceding section to choose, for each basis element B_n, a continuous function $f_n : X \to [0, 1]$ such that $f(x) > 0$ for $x \in B_n$ and $f(x) = 0$ for $x \notin B_n$. The collection $\{f_n\}$ satisfies our requirement. Given x_0 and given a neighborhood U of x_0, we can choose a basis element B_n with $x_0 \in B_n \subset U$; then the function f_n will be positive at x_0 and vanish outside U.

Another way to construct the collection $\{f_n\}$, which does not rely on the exercise cited, is the following: For each pair n, m of indices for which $\bar{B}_n \subset B_m$, apply the Urysohn lemma to choose a continuous function $g_{n,m} : X \to [0, 1]$ such that $g_{n,m}(\bar{B}_n) = \{1\}$ and $g_{n,m}(X - B_m) = \{0\}$. Then the collection $\{g_{n,m}\}$ satisfies our requirement: Given $x \in U$, one can choose a basis element B_m such that $x \in B_m \subset U$. Using regularity, one can then choose B_n so that $x_0 \in B_n$ and $\bar{B}_n \subset B_m$. Then the function $g_{n,m}$ is defined, and $g_{n,m}$ is positive at x_0 and vanishes outside U. Because this collection is indexed by a subset of $Z_+ \times Z_+$, it is countable; therefore it can be reindexed with the positive integers, giving us the desired indexed family $\{f_n\}$.

Step 2 (*First version of the proof*). Given the functions f_n of Step 1, take R^ω in the product topology and define a map $F : X \to R^\omega$ by the rule

$$F(x) = (f_1(x), f_2(x), \ldots).$$

We assert that F is an imbedding.

First, F is continuous because R^ω has the product topology and each f_n

is continuous. Second, F is injective because given $x \neq y$, we know there is an index n such that $f_n(x) > 0$ and $f_n(y) = 0$; therefore, $F(x) \neq F(y)$.

Finally, we must prove that F is a homeomorphism of X onto its image, the subspace $Z = F(X)$ of R^ω. We know that F defines a continuous bijection of X with Z, so we need only show that for each open set U in X, the set $F(U)$ is open in Z. Let z_0 be a point of $F(U)$. We shall find an open set W of Z such that

$$z_0 \in W \subset F(U).$$

Let x_0 be the point of U such that $F(x_0) = z_0$. Choose an index N for which $f_N(x_0) > 0$ and $f_N(X - U) = \{0\}$. Take the open ray $(0, +\infty)$ in R, and let V be the open set

$$V = \pi_N^{-1}((0, +\infty))$$

of R^ω. Let $W = V \cap Z$; then W is open in Z, by definition of the subspace topology. See Figure 11. We assert that $z_0 \in W \subset F(U)$. First, $z_0 \in W$

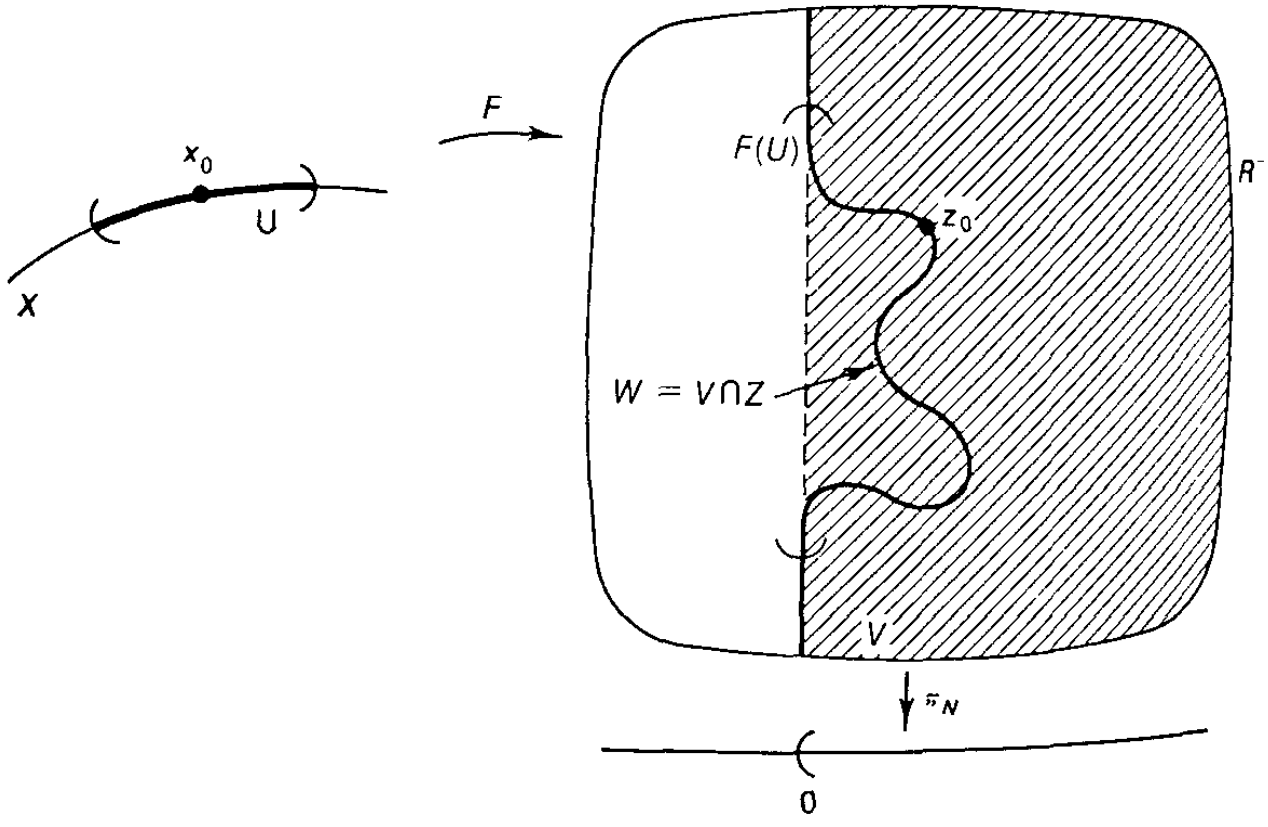


Figure 11

because

$$\pi_N(z_0) = \pi_N(F(x_0)) = f_N(x_0) > 0.$$

Second, $W \subset F(U)$. For if $z \in W$, then $z = F(x)$ for some $x \in X$, and $\pi_N(z) \in (0, +\infty)$. Since $\pi_N(z) = \pi_N(F(x)) = f_N(x)$, and f_N vanishes outside U, we must have x in U. Then $z = F(x)$ is in $F(U)$, as desired.

Thus F is an imbedding of X in R^ω.

Step 3 (*Second version of the proof*). In this version, we imbed X in the metric space $(R^\omega, \bar{\rho})$. Actually we imbed X in the subspace $[0, 1]^\omega$, on which $\bar{\rho}$ equals the metric

$$\rho(\mathbf{x}, \mathbf{y}) = \text{lub } \{|x_i - y_i|\}.$$

We use the countable collection of functions $f_n : X \to [0, 1]$ constructed in Step 1. But now we impose the additional condition that $f_n(x) \leq 1/n$ for all x. (This condition is easy to satisfy; we can just divide each function f_n by n.)

Define $F : X \to [0, 1]^\omega$ by the equation

$$F(x) = (f_1(x), f_2(x), \ldots)$$

as before. We assert that F is now an imbedding relative to the metric ρ on $[0, 1]^\omega$. We know from Step 2 that F is injective. Furthermore, we know that if we use the *product* topology on $[0, 1]^\omega$, the map F carries open sets of X onto open sets of $Z = F(X)$. This statement remains true if one passes to the finer (larger) topology on $[0, 1]^\omega$ induced by the metric ρ.

The one thing left to do is to prove that F is continuous. This does not follow from the fact that each component function is continuous, for we are not using the product topology on R^ω now. Here is where the assumption $f_n(x) \leq 1/n$ comes in.

Let x_0 be a point of X, and let $\epsilon > 0$. To prove continuity, we need to find a neighborhood U of x_0 such that

$$x \in U \Longrightarrow \rho(F(x), F(x_0)) < \epsilon.$$

First choose N large enough that $1/N \leq \epsilon/2$. Then for each $n = 1, \ldots, N$, use the continuity of f_n to choose a neighborhood U_n of x_0 such that

$$|f_n(x) - f_n(x_0)| \leq \epsilon/2$$

for $x \in U_n$. Let $U = U_1 \cap \cdots \cap U_N$; we show that U is the desired neighborhood of x_0. Let $x \in U$. If $n \leq N$,

$$|f_n(x) - f_n(x_0)| \leq \epsilon/2$$

by choice of U. And if $n > N$, then

$$|f_n(x) - f_n(x_0)| < 1/N \leq \epsilon/2$$

because f_n maps X into $[0, 1/n]$. Therefore for all $x \in U$,

$$\rho(F(x), F(x_0)) \leq \epsilon/2 < \epsilon,$$

as desired. □

In Step 2 of the preceding proof, we actually proved something stronger than the result stated there. For later use, we state it here:

Theorem 4.2 (Imbedding theorem). Let X *be Hausdorff. Suppose that* $\{f_\alpha\}_{\alpha \in J}$ *is a collection of continuous functions* $f_\alpha : X \to \mathbb{R}$ *satisfying the requirement that for each point* x_0 *of* X *and each neighborhood* U *of* x_0*, there is an index* α *such that* f_α *is positive at* x_0 *and vanishes outside* U*. Then the function* $F : X \to \mathbb{R}^J$ *defined by*

$$F(x) = (f_\alpha(x))_{\alpha \in J}$$

is an imbedding of X *in* $\mathbb{R}^J$.

The proof is almost a copy of Step 2 of the preceding proof; one merely replaces n by α and R^ω by R^J throughout.

A collection $\{f_\alpha\}$ of continuous functions satisfying the hypotheses of this theorem is said to **separate points from closed sets** in X.

Exercises

1. Give an example showing that a Hausdorff space with a countable basis need not be metrizable.
2. Give an example showing that a space can be completely normal, and satisfy the first countability axiom, the Lindelöf condition, and have a countable dense subset, and still not be metrizable.
3. Let X be a compact Hausdorff space. Is it true that if X has a countable basis, then X is metrizable? What about the converse?
4. Let X be a locally compact Hausdorff space.
   (a) Is it true that if X has a countable basis, then X is metrizable? What about the converse?
   (b) Let Y be the one-point compactification of X. Is it true that if X has a countable basis, then Y is metrizable? What about the converse?
5. A space X is **locally metrizable** if each point x of X has a neighborhood that is metrizable in the subspace topology. Show that a compact Hausdorff space X is metrizable if it is locally metrizable. [*Hint:* Show that X is a finite union of open subspaces, each of which has a countable basis.]
6. Show that a regular Lindelöf space is metrizable if it is locally metrizable. [*Hint:* A closed subspace of a Lindelöf space is Lindelöf.] Regularity is essential; where do you use it in the proof?
7. Show that if Y is a normal space with basis $\mathcal{B}$, then Y can be imbedded in $[0, 1]^J$, where J is some subset of $\mathcal{B} \times \mathcal{B}$.
8. Check the details of the proof of Theorem 4.2.
9. Show that if $\{f_\alpha\}$ is a family of real-valued continuous functions on X that separates points from closed sets, then the topology of X is the coarsest (smallest) topology relative to which all the functions f_α are continuous.
10. Show that if X is completely regular, then X can be imbedded in $[0, 1]^J$ for some J.

11. Let Y be a normal space. Then Y is said to be an **absolute retract** (AR) if whenever one has an imbedding $h: Y \to Z$ of Y into a normal space Z, such that $h(Y)$ is closed in Z, then $h(Y)$ is a retract of Z. Show that if Y is compact, the following statements are equivalent:

(i) Y is homeomorphic to a retract of $[0, 1]^J$ for some J.
(ii) Y is homeomorphic to a retract of R^J for some J.
(iii) Y has the universal extension property.
(iv) Y is an absolute retract.

In fact, show that (i) $\Rightarrow$ (ii) $\Rightarrow$ (iii) $\Rightarrow$ (iv) without assuming Y compact, and (iv) $\Rightarrow$ (i) if Y is compact. [*Hint:* Assume the Tychonoff theorem, so you know that $[0, 1]^J$ is normal.]

12. (a) Show the logarithmic spiral

$$C = \{0 \times 0\} \cup \{e^t \cos t \times e^t \sin t \mid t \in R\}$$

is a retract of R^2. Can you define a specific retraction $r: R^2 \to C$?

(b) Show that the "knotted x-axis" K of Figure 12 is a retract of R^3.

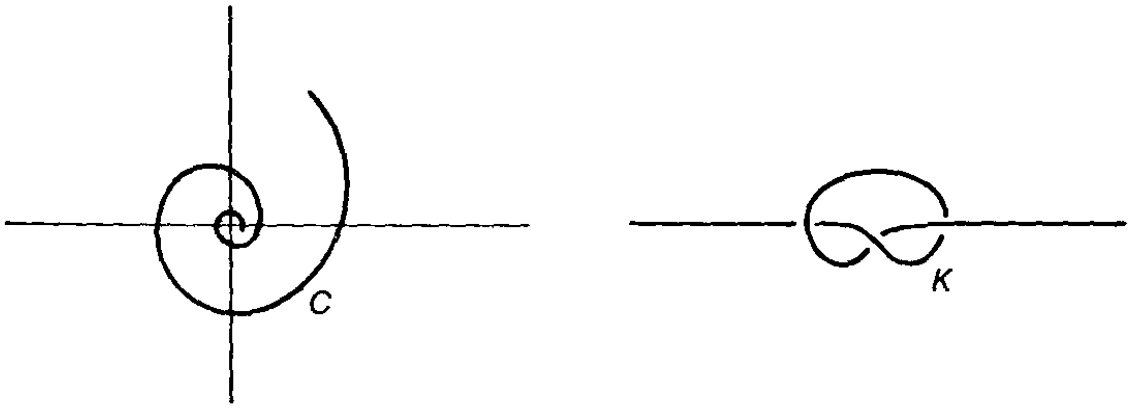


Figure 12

***13.** Prove the following:

Theorem. Let Y be a normal space. Then Y is an absolute retract if and only if Y has the universal extension property.

[*Hint:* If X and Y are disjoint normal spaces, A is closed in X, and $f: A \to Y$ is a continuous map, define the **adjunction space** Z_f to be the quotient space obtained from $X \cup Y$ by identifying each point a of A with the point $f(a)$ and with all the points of $f^{-1}(\{f(a)\})$. Using the Tietze theorem, show that Z_f is normal.]

***14.** A subspace Y of a space Z is called a **neighborhood retract** of Z if Y is a retract of some open set of Z. A normal space Y is called an **absolute neighborhood retract** (ANR) if it is a neighborhood retract of every normal space Z in which it is imbedded as a closed subspace. We say that Y has the **universal neighborhood extension property** if every continuous map $f: A \to Y$, where A is a closed subset of the normal space X, has a continuous extension to an open set in X about A. Generalize Exercises 11 and 13 to neighborhood retracts.

*4-5 Partitions of Unity†

We have shown that every regular space with a countable basis can be imbedded in the "infinite-dimensional" euclidean space R^ω. It is natural to ask under what conditions a space X can be imbedded in some finite-dimensional euclidean space R^N. One answer to this question is given below; we show that every compact manifold can be imbedded in R^N for some N. A more general answer will be obtained in Chapter 7, when we study dimension theory.

The proof uses a certain family of functions $\{\phi_\alpha\}$ on X that form what is called a "partition of unity." Such families have proved to be very useful tools in analysis, geometry, and topology. We shall consider only the case of finite partitions of unity (which is all we need) and leave the general case to the exercises.

First, we need some terminology.

If $\phi: X \longrightarrow R$, then the support of ϕ is defined to be the closure of the set $\phi^{-1}(R - \{0\})$. Thus if x lies outside the support of ϕ, there is some neighborhood of x on which ϕ vanishes.

Definition. Let $\{U_1, \ldots, U_n\}$ be a finite indexed open covering of the space X. An indexed family of continuous functions

$$\phi_i: X \longrightarrow [0, 1] \quad \text{for} \quad i = 1, \ldots, n,$$

is said to be a partition of unity dominated by $\{U_i\}$ if:

(1) (support ϕ_i) $\subset U_i$ for each i.

(2) $\sum_{i=1}^{n} \phi_i(x) = 1$ for each x.

Theorem 5.1 (Existence of finite partitions of unity). *Let $\{U_1, \ldots, U_n\}$ be a finite open covering of the normal space X. Then there exists a partition of unity dominated by $\{U_i\}$.*

Proof. Step 1. First, we prove that one can "shrink" the covering $\{U_i\}$ to an open covering $\{V_1, \ldots, V_n\}$ of X such that $\bar{V}_i \subset U_i$ for each i.

We proceed by induction. First, note that the set

$$A = X - (U_2 \cup \cdots \cup U_n)$$

is a closed subset of X. Because $\{U_1, \ldots, U_n\}$ covers X, the set A is contained in the open set U_1. Using normality, choose an open set V_1 containing A such that $\bar{V}_1 \subset U_1$. Then the collection $\{V_1, U_2, \ldots, U_n\}$ covers X.

In general, given open sets $V_1, \ldots, V_{k-1}$ such that the collection

$$\{V_1, \ldots, V_{k-1}, U_k, U_{k+1}, \ldots, U_n\}$$

covers X, let

†This section will be used when we study dimension theory in §7-9.

$$A = X - (V_1 \cup \cdots \cup V_{k-1}) - (U_{k+1} \cup \cdots \cup U_n).$$

Then A is a closed subset of X which is contained in the open set U_k. Choose V_k to be an open set containing A such that $\bar{V}_k \subset U_k$. Then

$$\{V_1, \ldots, V_{k-1}, V_k, U_{k+1}, \ldots, U_n\}$$

covers X. At the nth step of the induction, our result is proved.

Step 2. Now we prove the theorem. Given the open covering $\{U_1, \ldots, U_n\}$ of X, choose an open covering $\{V_1, \ldots, V_n\}$ of X such that $\bar{V}_i \subset U_i$ for each i. Then choose an open covering $\{W_1, \ldots, W_n\}$ of X such that $\bar{W}_i \subset V_i$ for each i. Using the Urysohn lemma, choose for each i a continuous function

$$\psi_i : X \longrightarrow [0, 1]$$

such that $\psi_i(\bar{W}_i) = \{1\}$ and $\psi_i(X - V_i) = \{0\}$. Since $\psi_i^{-1}(R - \{0\})$ is contained in V_i, we have

$$(\text{support } \psi_i) \subset \bar{V}_i \subset U_i.$$

Because the collection $\{W_i\}$ covers X, the sum $\Psi(x) = \sum_{i=1}^n \psi_i(x)$ is positive for each x. Therefore, we may define, for each j,

$$\phi_j(x) = \frac{\psi_j(x)}{\Psi(x)}.$$

It is easy to check that the functions $\phi_1, \ldots, \phi_n$ form the desired partition of unity. □

Definition. An m-manifold is a Hausdorff space X with a countable basis such that each point x of X has a neighborhood that is homeomorphic with an open subset of R^m.

A 1-manifold is often called a **curve**, and a 2-manifold is called a **surface**. Manifolds form a very important class of spaces; they are much studied in differential geometry and algebraic topology.

We shall prove that if X is a compact manifold, then X can be imbedded in a finite-dimensional euclidean space. The theorem holds without the assumption of compactness, but the proof is a good deal harder.

Theorem 5.2. *If* X *is a compact m-manifold, then* X *can be imbedded in* R^{N} *for some positive integer* N.

Proof. Cover X by finitely many open sets $\{U_1, \ldots, U_n\}$, each of which may be imbedded in R^m. Choose imbeddings $g_i : U_i \longrightarrow R^m$ for each i. Being compact and Hausdorff, X is normal. Let $\phi_1, \ldots, \phi_n$ be a partition of unity dominated by $\{U_i\}$; let $A_i = \text{support } \phi_i$. For each $i = 1, \ldots, n$, define a function $h_i : X \longrightarrow R^m$ by the rule

$$h_i(x) = \begin{cases} \phi_i(x) \cdot g_i(x) & \text{for } x \in U_i, \\ \mathbf{0} = (0, \ldots, 0) & \text{for } x \in X - A_i. \end{cases}$$

[Here $\phi_i(x)$ is a real number c and $g_i(x)$ is a point $\mathbf{y} = (y_1, \ldots, y_m)$ of R^m; the product $c \cdot \mathbf{y}$ denotes of course the point $(cy_1, \ldots, cy_m)$ of R^m.] The function h_i is well defined because the two definitions of h_i agree on the intersection of their domains, and h_i is continuous because its restrictions to the open sets U_i and $X - A_i$ are continuous.

Now define

$$F: X \longrightarrow (\underbrace{R \times \cdots \times R}_{n \text{ times}} \times \underbrace{R^m \times \cdots \times R^m}_{n \text{ times}})$$

by the rule

$$F(x) = (\phi_1(x), \ldots, \phi_n(x), h_1(x), \ldots, h_n(x)).$$

Clearly F is continuous. To prove that F is an imbedding we need only to show that F is injective (because X is compact). Suppose that $F(x) = F(y)$. Then $\phi_i(x) = \phi_i(y)$ and $h_i(x) = h_i(y)$ for all i. Now $\phi_i(x) > 0$ for some i [since $\sum \phi_i(x) = 1$]. Therefore, $\phi_i(y) > 0$ also, so that $x, y \in U_i$. Then

$$\phi_i(x) \cdot g_i(x) = h_i(x) = h_i(y) = \phi_i(y) \cdot g_i(y).$$

Because $\phi_i(x) = \phi_i(y) > 0$, we conclude that $g_i(x) = g_i(y)$. But $g_i: U_i \to R^n$ is injective, so that $x = y$, as desired. □

In many applications of partitions of unity, such as the one just given, all one needs to know is that the sum $\sum \phi_i(x)$ is positive for each x. In others, however, one needs the stronger condition that $\sum \phi_i(x) = 1$. See §7-9.

Exercises

1. Prove that every manifold is regular and hence metrizable. Where do you use the Hausdorff condition?
2. Let X be a compact Hausdorff space. Suppose that for each $x \in X$, there is a neighborhood U of x and a positive integer k such that U can be imbedded in R^k. Show that X can be imbedded in R^N for some positive integer N.
3. An indexed collection $\{A_\alpha\}$ of subsets of X is said to be a **point-finite indexed family** if each $x \in X$ belongs to A_α for only finitely many values of α.

   Lemma (The shrinking lemma). Let X be a normal space; let $\{U_\alpha\}_{\alpha \in J}$ be a point-finite indexed open covering of X. Then there exists an indexed open covering $\{V_\alpha\}_{\alpha \in J}$ of X such that $\bar{V}_\alpha \subset U_\alpha$ for each α.

   (a) Prove the lemma in the case $J = Z_+$, by induction. Where do you use the fact that $\{U_\alpha\}$ is point-finite?

   *(b) Use transfinite induction to prove the lemma for an arbitrary well-ordered index set J.
4. Let X be a space. An indexed family $\{A_\alpha\}_{\alpha \in J}$ of subsets of X is said to be a **locally finite indexed family** if each point of X has a neighborhood that intersects A_α for only finitely many values of α.

Let $\{U_\alpha\}_{\alpha \in J}$ be an indexed open covering of X. A family of continuous functions

$$\phi_\alpha : X \longrightarrow [0, 1],$$

indexed by $\alpha \in J$, is said to be a **partition of unity** dominated by $\{U_\alpha\}$ if

(i) (support ϕ_α) $\subset U_\alpha$ for each $\alpha \in J$,

(ii) $\{\text{support } \phi_\alpha\}_{\alpha \in J}$ is locally finite,

(iii) $\Sigma_{\alpha \in J} \phi_\alpha(x) = 1$ for each x.

[The sum makes sense because for given x, we have $\phi_\alpha(x) \neq 0$ for only finitely many values of α.] Assuming the shrinking lemma, prove the following:

Theorem. If X is a normal space and if $\{U_\alpha\}_{\alpha \in J}$ is a locally finite indexed family of open sets covering X, then there exists a partition of unity dominated by $\{U_\alpha\}$.

5. If $\{U_\alpha\}_{\alpha \in J}$ is a collection of subsets of X, a collection $\{V_\beta\}_{\beta \in K}$ is said to **refine** $\{U_\alpha\}$ if for each set V_β, there is at least one set U_α containing it.

Theorem. Let X be a normal space, and let $\{U_\alpha\}_{\alpha \in J}$ be an indexed open covering of X. If there is an open covering $\{V_\beta\}_{\beta \in K}$ of X which refines $\{U_\alpha\}$ and is locally finite, then there exists a partition of unity dominated by $\{U_\alpha\}$.

[*Hint:* Choose $f: K \longrightarrow J$ so that $V_\beta \subset U_{f(\beta)}$. Let W_α be the union of those V_β for which $f(\beta) = \alpha$; show that $\{W_\alpha\}_{\alpha \in J}$ is locally finite.]

A Hausdorff space X is said to be **paracompact** if every open covering of X has a locally finite refinement that is an open covering of X. We shall study paracompact spaces in Chapter 6.

6. The Hausdorff condition is an essential part of the definition of a manifold; it is not implied by the other parts of the definition. Consider the following subset of R^2:

$$X = (R \times 0) \cup (\bar{R}_+ \times 1).$$

Topologize X by taking as basis all sets of the following types:

(i) $(a, b) \times 0$ for $a < b$,

(ii) $(a, b) \times 1$ for $0 \leq a < b$,

(iii) $((a, 0) \times 0) \cup ([0, b) \times 1)$ for $a < 0 < b$.

(a) Show that this is a basis.

(b) Show that each of the subspaces $R \times 0$ and $(R_- \times 0) \cup (\bar{R}_+ \times 1)$ of X is homeomorphic to R.

(c) Show that X has a countable basis, that finite point sets are closed in X, and that each point of X has a neighborhood homeomorphic with an open set in R.

(d) Show that X is not Hausdorff.

*Supplementary Exercises: Review of Part I

Consider the following properties a space may satisfy:

(1) connected

(2) path connected

(3) locally connected
(4) locally path connected
(5) compact
(6) limit point compact
(7) locally compact Hausdorff
(8) Hausdorff
(9) regular
(10) normal
(11) first-countable
(12) second-countable
(13) Lindelöf
(14) has a countable dense subset
(15) locally metrizable
(16) metrizable

1. For each of the following spaces, determine (if you can) which of these properties it satisfies. (Assume the Tychonoff theorem if you need it.)
(a) S_Ω
(b) $\bar{S}_\Omega$
(c) $S_\Omega \times \bar{S}_\Omega$
(d) $S_\Omega \times [0, 1)$ in the dictionary order
(e) $I \times I$ in the dictionary order, where $I = [0, 1]$
(f) R_l
(g) R_l^2
(h) R^ω in the box, uniform, and product topologies
(i) R^I in the product topology
(j) R in the finite complement topology

2. Which of these properties does a metric space necessarily have?

3. Which of these properties does a compact Hausdorff space have?

4. Which of these properties are preserved when one passes to a subspace? To a closed subspace? To an open subspace?

5. Which of these properties are preserved under finite products? Countable products? Arbitrary products?

6. Which of these properties are preserved by continuous maps?

7. After studying Chapters 5, 6, and 7, repeat Exercises 1–6 for the following properties:
(17) completely regular
(18) paracompact
(19) topologically complete

You should be able to answer all but one of the 336 questions involved in Exercises 1–6, and all but one of the 63 questions involved in Exercise 7. These two are unsolved; see the remark in Exercise 10 of § 4-2.

5. The Tychonoff Theorem

We now return to a problem we left unresolved in Chapter 3. We shall prove the Tychonoff theorem, to the effect that arbitrary products of compact spaces are compact. The proof makes use of the maximum principle of set theory (see §1-11).

The Tychonoff theorem is of great usefulness to analysts (less so to geometers). We apply it in §5-2 to obtain an interesting characterization of completely regular spaces. Another application appears in §5-3, where we construct the Stone–Čech compactification and explore its properties; in this section we assume §3-8, Local compactness.

5-1 The Tychonoff Theorem

Like the Urysohn lemma, the Tychonoff theorem is what we call a "deep" theorem. Its proof involves not one but several original ideas; it is anything but straightforward. We shall discuss the crucial ideas of the proof in some detail before turning to the proof itself.

In Chapter 3, we proved the product $X \times Y$ of two compact spaces to be compact. For that proof the open covering formulation of compactness was quite satisfactory. Given an open covering of $X \times Y$ by basis elements,

we covered each slice $x \times Y$ by finitely many of them, and proceeded from that to construct a finite covering of $X \times Y$.

Try as we may, we cannot make this approach work for an infinite product of compact spaces. So the first idea involved in the proof of the Tychonoff theorem is to abandon open coverings and to approach the problem using instead the closed set formulation of compactness. Actually what we use is the formulation expressed in Corollary 5.10 of Chapter 3 which states that X is compact if and only if for every collection $\mathcal{A}$ of subsets of X satisfying the f.i.c. (finite intersection condition), the intersection $\bigcap_{A \in \mathcal{A}} \bar{A}$ of the closures of the elements of $\mathcal{A}$ is nonempty.

To see how this approach might work, let us consider first the simplest possible case: the product of two compact spaces $X_1 \times X_2$. Suppose that $\mathcal{A}$ is a collection of subsets of $X_1 \times X_2$ satisfying the f.i.c. Consider the projection map $\pi_1 : X_1 \times X_2 \longrightarrow X_1$. The collection

$$\{\pi_1(A) \mid A \in \mathcal{A}\}$$

of subsets of X_1 also satisfies the f.i.c., since if $x \in A_1 \cap \cdots \cap A_n$, then $\pi_1(x) \in \pi_1(A_1) \cap \cdots \cap \pi_1(A_n)$, so the latter set is nonempty. Compactness of X_1 guarantees that the intersection of all the sets $\overline{\pi_1(A)}$ is nonempty. Let us choose a point x_1 belonging to this intersection. Similarly, let us choose a point x_2 belonging to all the sets $\overline{\pi_2(A)}$. The obvious conclusion we would like to draw is that the point $x_1 \times x_2$ lies in $\bigcap_{A \in \mathcal{A}} \bar{A}$, for then our theorem would be proved.

But that is unfortunately not true. Consider the following example, in which $X_1 = X_2 = [0, 1]$ and the collection $\mathcal{A}$ consists of all elliptical regions bounded by ellipses that have the points $p = (\frac{1}{3}, \frac{1}{3})$ and $q = (\frac{1}{2}, \frac{2}{3})$ as their foci. See Figure 1. Certainly $\mathcal{A}$ satisfies the f.i.c. Now let us pick a point x_1 in the intersection of the sets $\{\overline{\pi_1(A)} \mid A \in \mathcal{A}\}$. Any point of the interval

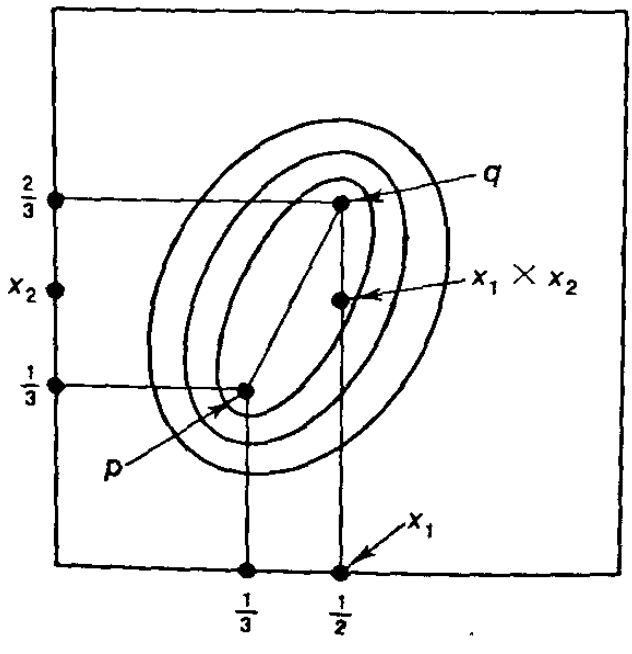


Figure 1

$[\frac{1}{3}, \frac{1}{2}]$ will do; suppose we choose $x_1 = \frac{1}{2}$. Similarly, choose a point x_2 in the intersection of the sets $\{\overline{\pi_2(A)} \mid A \in \mathcal{A}\}$. Any point of the interval $[\frac{1}{3}, \frac{2}{3}]$ will do; suppose we pick $x_2 = \frac{1}{2}$. This proves to be an unfortunate choice, for the point

$$x_1 \times x_2 = \tfrac{1}{2} \times \tfrac{1}{2}$$

does not lie in the intersection of the sets $\bar{A}$.

"Aha!" you say, "you made a bad choice. If after choosing $x_1 = \frac{1}{2}$ you had chosen $x_2 = \frac{2}{3}$, *then* you would have found a point in $\bigcap_{A \in \mathcal{A}} \bar{A}$." The difficulty with our tentative proof is that it gave us too much freedom in picking x_1 and x_2; it allowed us to make a "bad" choice instead of a "good" choice.

How can we alter the proof so as to avoid this difficulty?

This question leads to the second idea of the proof: Perhaps if we *expand* the collection $\mathcal{A}$ (retaining the f.i.c., of course), that expansion will restrict the choices of x_1 and x_2 sufficiently that we will be forced to make the "right" choice. To illustrate, suppose that in the previous example we expand the collection $\mathcal{A}$ to the collection $\mathcal{D}$ consisting of all elliptical regions bounded by ellipses that have $p = (\frac{1}{3}, \frac{1}{3})$ as one focus and any point of the line segment pq as the other focus. This collection is illustrated in Figure 2. The new collec-

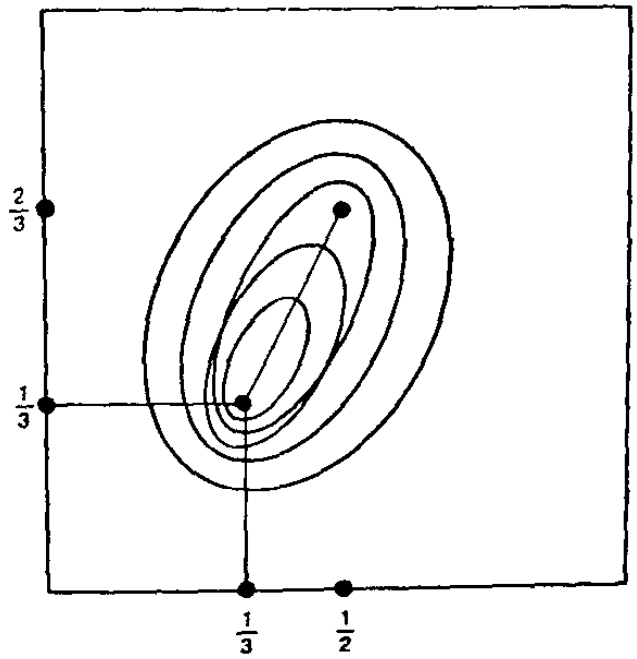


Figure 2

tion $\mathcal{D}$ still satisfies the f.i.c. But if you try to choose a point x_1 in

$$\bigcap_{D \in \mathcal{D}} \overline{\pi_1(D)},$$

the only possible choice for x_1 is $\frac{1}{3}$. Similarly, the only possible choice for x_2 is $\frac{1}{3}$. And $\frac{1}{3} \times \frac{1}{3}$ *does* belong to every set $\bar{D}$, and hence to every set $\bar{A}$. In other words, expanding the collection $\mathcal{A}$ to the collection $\mathcal{D}$ forces the proper choice on us.

Now of course in this example we chose $\mathcal{D}$ carefully so that the proof would work. What hope can we have for choosing $\mathcal{D}$ correctly in general? Here is the third idea of the proof: Why not simply choose $\mathcal{D}$ to be a collection that is "as large as possible"—so that no larger collection satisfies the f.i.c.—and see whether such a $\mathcal{D}$ will work? It is not at all obvious that such a collection $\mathcal{D}$ exists; to prove it, we must appeal to the maximum principle of set theory. But after we prove that $\mathcal{D}$ exists, we shall in fact be able to show that $\mathcal{D}$ is large enough to force the proper choices on us.

Now let us apply these ideas:

***Theorem 1.1** (**The Tychonoff theorem**). An arbitrary product of compact spaces is compact in the product topology.*

Proof. Step 1. Let X be a set; let $\mathcal{A}$ be a collection of subsets of X satisfying the f.i.c. We show that there is a collection $\mathcal{D}$ of subsets of X such that

(1) $\mathcal{D} \supset \mathcal{A}$.

(2) $\mathcal{D}$ satisfies the f.i.c.

(3) If $\mathcal{E} \supsetneq \mathcal{D}$, then $\mathcal{E}$ does not satisfy the f.i.c.

We often express conditions (2) and (3) by saying that $\mathcal{D}$ is "maximal with respect to the f.i.c."

Recall the maximum principle. It states that if $\prec$ is a strict partial order relation on the set A, and if B is a subset of A that is simply ordered by $\prec$, then there is a maximal simply ordered subset C of A containing B. We need a special case of this principle, the case where B is a one-element subset of A; such a subset is automatically simply ordered. Thus what we need is the following:

If A is a set with a strict partial order and if a is an element of A, then there is a maximal simply ordered subset C of A such that $a \in C$.

We shall apply this principle in the case where A is not a set, nor even a collection of sets, but rather a set whose elements are collections of sets. For purposes of this proof, we shall call such a set a "superset" and shall denote it by an outline letter. To summarize the notation:

c is an element of X,
C is a subset of X,
$\mathcal{C}$ is a collection of subsets of X,
$\mathbb{C}$ is a superset of collections of subsets of X.

We are given the collection $\mathcal{A}$ of subsets of X, satisfying the f.i.c. Let $\mathbb{A}$ denote the superset consisting of *all* collections of subsets of X that satisfy the f.i.c. Then $\mathcal{A} \in \mathbb{A}$. Let us use proper inclusion $\subsetneq$ as the strict partial order on $\mathbb{A}$. By the special case of the maximum principle just cited, there exists a maximal simply ordered subsuperset $\mathbb{C}$ of $\mathbb{A}$ such that $\mathcal{A} \in \mathbb{C}$.

Define $\mathcal{D}$ to be the union of the elements of $\mathbb{C}$;

$$\mathcal{D} = \bigcup_{\mathcal{C} \in \mathbb{C}} \mathcal{C}.$$

We assert that $\mathcal{D}$ is the collection we want.

First, note that $\mathcal{A} \subset \mathcal{D}$, because $\mathcal{A} \in \mathbb{C}$. Second, we show that $\mathcal{D}$ satisfies the f.i.c. Let $D_1, \ldots, D_n$ be elements of $\mathcal{D}$. Since $\mathcal{D}$ is the union of the elements of $\mathbb{C}$, for each i there is an element $\mathcal{C}_i$ of $\mathbb{C}$ such that $D_i \in \mathcal{C}_i$. Now $\mathbb{C}$ is simply ordered by proper inclusion. Therefore, the superset $\{\mathcal{C}_1, \ldots, \mathcal{C}_n\}$ is also simply ordered by proper inclusion. Being finite, it has a largest element; that is, there is an index k such that $\mathcal{C}_i \subset \mathcal{C}_k$ for $i = 1, \ldots, n$. In particular, the sets $D_1, \ldots, D_n$ all belong to $\mathcal{C}_k$. Since $\mathcal{C}_k$ satisfies the f.i.c. by hypothesis, $D_1 \cap \cdots \cap D_n$ is nonempty.

Third, we show $\mathcal{D}$ is maximal with respect to the f.i.c. Let $\mathcal{D} \subsetneq \mathcal{E}$ and suppose that $\mathcal{E}$ satisfies the f.i.c. Then we can form a new superset $\mathbb{C}'$ by adjoining the element $\mathcal{E}$ to $\mathbb{C}$;

$$\mathbb{C}' = \mathbb{C} \cup \{\mathcal{E}\}.$$

Now $\mathcal{C} \subset \mathcal{D}$ for every $\mathcal{C} \in \mathbb{C}$, and $\mathcal{D} \subsetneq \mathcal{E}$. Therefore, every two elements of $\mathbb{C}'$ are comparable under proper inclusion, so that $\mathbb{C}'$ is simply ordered. This contradicts the maximality of $\mathbb{C}$.

Step 2. Let $\mathcal{D}$ be a collection of subsets of X that is maximal with respect to the f.i.c. Then:

(a) *Any finite intersection of elements of* $\mathcal{D}$ *belongs to* $\mathcal{D}$.

Let B be the intersection of finitely many elements of $\mathcal{D}$. Define a collection $\mathcal{E}$ by adjoining B to $\mathcal{D}$, so that $\mathcal{E} = \mathcal{D} \cup \{B\}$. We show that $\mathcal{E}$ satisfies the f.i.c.; then maximality of $\mathcal{D}$ implies that $\mathcal{E} = \mathcal{D}$, whence $B \in \mathcal{D}$ as desired.

Take finitely many elements of $\mathcal{E}$. If none of them is the set B, then their intersection is nonempty because $\mathcal{D}$ satisfies the f.i.c. If one of them is the set B, then their intersection is of the form

$$D_1 \cap \cdots \cap D_m \cap B.$$

Since B equals a finite intersection of elements of $\mathcal{D}$, this set is nonempty.

(b) *If A is any subset of X that intersects every element of $\mathcal{D}$, then A belongs to $\mathcal{D}$.*

Given A, define $\mathcal{E} = \mathcal{D} \cup \{A\}$. We show that $\mathcal{E}$ satisfies the f.i.c., from which we conclude that A belongs to $\mathcal{D}$. Take finitely many elements of $\mathcal{E}$. If none of them is the set A, their intersection is automatically nonempty. Otherwise it is of the form

$$D_1 \cap \cdots \cap D_n \cap A.$$

Now $D_1 \cap \cdots \cap D_n$ belongs to $\mathcal{D}$, by (a); therefore, this intersection is nonempty, by hypothesis.

Step 3. Now we prove the Tychonoff theorem. Let

$$X = \prod_{\alpha \in J} X_\alpha,$$

where each space X_α is compact. Let $\mathcal{A}$ be a collection of subsets of X satis-

fying the f.i.c. We prove that the intersection

$$\bigcap_{A \in \mathcal{A}} \bar{A}$$

is nonempty. Compactness of X follows.

Applying Step 1, choose a collection $\mathcal{D}$ of subsets of X such that $\mathcal{D} \supset \mathcal{A}$ and $\mathcal{D}$ is maximal with respect to the f.i.c. It will suffice to prove that the intersection $\bigcap_{D \in \mathcal{D}} \bar{D}$ is nonempty.

Given $\alpha \in J$, let $\pi_\alpha : X \to X_\alpha$ be the projection map, as usual. Consider the collection

$$\{\pi_\alpha(D) \mid D \in \mathcal{D}\}$$

of subsets of X_α. This collection satisfies the f.i.c. because $\mathcal{D}$ does. By compactness of X_α, we can for each α choose a point x_α of X_α such that

$$x_\alpha \in \bigcap_{D \in \mathcal{D}} \overline{\pi_\alpha(D)}.$$

Let $\mathbf{x}$ be the point $(x_\alpha)_{\alpha \in J}$ of X. We shall show that $\mathbf{x} \in \bar{D}$ for every $D \in \mathcal{D}$; then our proof will be finished.

First we show that if $\pi_\beta^{-1}(U_\beta)$ is any subbasis element (for the product topology on X) containing $\mathbf{x}$, then $\pi_\beta^{-1}(U_\beta)$ intersects every element of $\mathcal{D}$: The set U_β is a neighborhood of x_β in X_β. Since $x_\beta \in \overline{\pi_\beta(D)}$ by definition, U_β intersects $\pi_\beta(D)$ in some point $\pi_\beta(\mathrm{y})$, where $\mathrm{y} \in D$. Then it follows that $\mathbf{y} \in \pi_\beta^{-1}(U_\beta) \cap D$.

It follows from (b) of Step 2 that every subbasis element containing $\mathbf{x}$ *belongs* to $\mathcal{D}$. And then it follows from (a) of Step 2 that every *basis* element containing $\mathbf{x}$ belongs to $\mathcal{D}$. Since $\mathcal{D}$ has the f.i.c., this means that every basis element containing $\mathbf{x}$ intersects every element of $\mathcal{D}$; whence $\mathbf{x} \in \bar{D}$ for every $D \in \mathcal{D}$, as desired. □

Exercises

1. Let X be a space. Let $\mathcal{D}$ be a maximal collection of subsets of X satisfying the f.i.c.
   (a) Show that $x \in \bar{D}$ for every $D \in \mathcal{D}$ if and only if every neighborhood of x belongs to $\mathcal{D}$. Which implication uses maximality of $\mathcal{D}$?
   (b) Let $D \in \mathcal{D}$. Show that if $A \supset D$, then $A \in \mathcal{D}$.
   (c) Show that if X is Hausdorff, there is at most one point belonging to $\bigcap_{D \in \mathcal{D}} \bar{D}$.
2. A collection $\mathcal{A}$ of subsets of X satisfies the **countable intersection condition** (c.i.c.) if every countable intersection of elements of $\mathcal{A}$ is nonempty. Show that X is a Lindelöf space if and only if for every collection $\mathcal{A}$ of subsets of X satisfying the c.i.c.,
$$\bigcap_{A \in \mathcal{A}} \bar{A} \neq \varnothing.$$
3. Consider the three statements:
   (i) If X is a set and $\mathcal{A}$ is a collection of subsets of X satisfying the c.i.c.,

then there is a collection $\mathcal{D}$ of subsets of X such that $\mathcal{D} \supset \mathcal{A}$ and $\mathcal{D}$ is maximal with respect to the c.i.c.

(ii) Suppose $\mathcal{D}$ is maximal with respect to the c.i.c. Then countable intersections of elements of $\mathcal{D}$ are in $\mathcal{D}$. And if A is a subset of X that intersects every element of $\mathcal{D}$, then necessarily $A \in \mathcal{D}$.

(iii) Products of Lindelöf spaces are Lindelöf.

(a) Prove (ii).

(b) Show that (i) and (ii) together imply (iii).

(c) Given a set X, show there exists a maximal collection of subsets of X satisfying the c.i.c. Why is this not adequate to prove (i)?

(d) The space R_l is Lindelöf and the space R_l^2 is not. (See Example 4 of §4-1). Therefore, (i) does not hold. Where does the proof of (i) break down?

4. Here is another theorem whose proof uses the maximum principle. Recall that if A is a space and if $x, y \in A$, we say that x and y belong to the same *quasicomponent* of A if there is no separation $A = C \cup D$ of A into two disjoint sets open in A such that $x \in C$ and $y \in D$.

(a) Let X be a compact Hausdorff space; let $x, y \in X$. Let $\mathcal{A}$ be a collection of closed sets of X such that for each $A \in \mathcal{A}$, the points x and y lie in the same quasicomponent of A. If $\mathcal{A}$ is simply ordered by proper inclusion, show that x and y lie in the same quasicomponent of

$$Y = \bigcap_{A \in \mathcal{A}} A.$$

[*Hint:* See Exercise 12 of §3-5.]

(b) Prove the following:

Theorem. If X is compact Hausdorff, the quasicomponents of X equal the components of X.

[*Hint:* If x and y lie in the same quasicomponent of X, let $\mathcal{A}$ be the collection of all closed sets A in X such that x and y lie in the same quasicomponent of A. Let $\mathcal{A}'$ be a maximal simply ordered subcollection and show that

$$\bigcap_{A \in \mathcal{A}'} A$$

is connected.]

(c) Prove the following:

Corollary. Let X be compact Hausdorff; let $x \in X$. The intersection of all those sets A containing x which are both open and closed in X equals the component of X containing x.

5-2 Completely Regular Spaces

We have already mentioned that the failure of the proof of the Urysohn lemma to generalize to regular spaces leads us to formulate a new separation axiom, the axiom of complete regularity. In this section we study some of the properties of spaces satisfying this axiom. One justification for our interest in this class of spaces comes from an imbedding theorem we shall prove, which shows that the class of completely regular spaces is identical

with a class of spaces we have already considered, the class of all subspaces of compact Hausdorff spaces.

Definition. A space X is **completely regular** if one-point sets are closed in X and if for each point x_0 and each closed set A not containing x_0, there is a continuous function $f: X \to [0, 1]$ such that $f(x_0) = 1$ and $f(A) = \{0\}$.

A normal space is completely regular, by the Urysohn lemma, and a completely regular space is regular, since given f, the sets $f^{-1}([0, \frac{1}{2}))$ and $f^{-1}((\frac{1}{2}, 1])$ are disjoint open sets about A and x_0, respectively. As a result, this new axiom fits in between regularity and normality in the list of separation axioms. Note that in the definition one could just as well require the function to map x_0 to 0, and A to $\{1\}$, for $g(x) = 1 - f(x)$ satisfies this condition.

Theorem 2.1. A subspace of a completely regular space is completely regular. A product of completely regular spaces is completely regular.

Proof. Let X be completely regular; let Y be a subspace of X. Let x_0 be a point of Y, and let A be a closed set of Y disjoint from x_0. Now $A = \bar{A} \cap Y$, where $\bar{A}$ denotes the closure of A in X. Therefore, $x_0 \notin \bar{A}$. Since X is completely regular, we can choose a continuous function $f: X \to [0, 1]$ such that $f(x_0) = 1$ and $f(\bar{A}) = \{0\}$. The restriction of f to Y is the desired continuous function on Y.

Let $X = \prod X_\alpha$ be a product of completely regular spaces. Let $\mathbf{b} = (b_\alpha)$ be a point of X and let A be a closed set of X disjoint from $\mathbf{b}$. Choose a basis element $\prod U_\alpha$ containing $\mathbf{b}$ that does not intersect A; then $U_\alpha = X_\alpha$ except for finitely many α, say $\alpha = \alpha_1, \ldots, \alpha_n$. Given $i = 1, \ldots, n$, choose a continuous function

$$f_i: X_{\alpha_i} \to [0, 1]$$

such that $f_i(b_{\alpha_i}) = 1$ and $f_i(X - U_{\alpha_i}) = \{0\}$. Let $\phi_i(\mathbf{x}) = f_i(\pi_{\alpha_i}(\mathbf{x}))$; then ϕ_i maps X continuously into R and vanishes outside $\pi_{\alpha_i}^{-1}(U_{\alpha_i})$. The product

$$f(\mathbf{x}) = \phi_1(\mathbf{x}) \cdot \phi_2(\mathbf{x}) \cdot \cdots \cdot \phi_n(\mathbf{x})$$

is the desired continuous function on X, for it equals 1 at $\mathbf{b}$ and vanishes outside $\prod U_\alpha$. □

EXAMPLE 1. The space $S_\Omega \times \bar{S}_\Omega$ is a space that is completely regular, since it is a subspace of the normal space $\bar{S}_\Omega \times \bar{S}_\Omega$. But it is not normal; see Example 2 of §4-2.

A space that is regular but not completely regular is much harder to find. Most of the examples that have been constructed for this purpose are difficult, and require considerable familiarity with cardinal numbers. Fairly recently, however, John Thomas [T] has constructed a much more elementary example, which we outline in Exercise 6.

The separation of disjoint closed sets by continuous real-valued functions is the crucial tool used in proving the Urysohn metrization theorem; one uses it to imbed the space X into the product space R^ω. In a completely regular space, we can separate points from closed sets by continuous functions, so it is natural to ask whether one can prove a similar theorem under the hypothesis of complete regularity. One can. It does not turn out to be a metrization theorem, but it *is* an interesting imbedding theorem:

Theorem 2.2. *If* X *is completely regular, then* X *can be imbedded in* $[0, 1]^J$ *for some* J.

Proof. Let $\{f_\alpha\}_{\alpha \in J}$ denote the set of all continuous functions from X into the interval $[0, 1]$, indexed by some index set J. Complete regularity of X guarantees that this collection separates points from closed sets in X. Therefore, by the Imbedding theorem of §4-4, the map

$$F(x) = (f_\alpha(x))_{\alpha \in J}$$

is an imbedding of X in $[0, 1]^J$. □

Corollary 2.3. *Let* X *be a space. The following are equivalent:*

(1) X *is completely regular.*

(2) X *is homeomorphic to a subspace of a compact Hausdorff space.*

(3) X *is homeomorphic to a subspace of a normal space.*

Exercises

1. Show that every locally compact Hausdorff space is completely regular.

2. Show that although R_l^2 is not normal, it is completely regular.

3. Let X be completely regular; let A and B be disjoint closed subsets of X. Show that if A is compact, there is a continuous function $f: X \to [0, 1]$ such that $f(A) = \{0\}$ and $f(B) = \{1\}$.

*4. Show that R^ω in the box topology is completely regular. (This fact is an immediate consequence of the following theorem, but give a direct proof.)

*5. Prove the following:

Theorem. Every Hausdorff topological group is completely regular.

Proof. Let V_0 be a neighborhood of the identity element e, in the topological group G. In general, choose V_n to be a neighborhood of e such that $V_n \cdot V_n \subset V_{n-1}$. Consider the set of all dyadic rationals p, that is, all rational numbers of the form $k/2^n$, with k and n integers. For each dyadic rational p in $(0, 1]$, define an open set $U(p)$ inductively as follows: $U(1) = V_0$ and $U(\frac{1}{2}) = V_1$. Given n, if $U(k/2^n)$ is defined for $0 < k/2^n \leq 1$, define

$$U(1/2^{n+1}) = V_{n+1},$$

$$U((2k + 1)/2^{n+1}) = V_{n+1} \cdot U(k/2^n)$$

for $0 < k < 2^n$. For $p \leq 0$, let $U(p) = \emptyset$; and for $p > 1$, let $U(p) = G$. Show that

$$V_n \cdot U(k/2^n) \subset U((k+1)/2^n)$$

for all k and n. Proceed as in the Urysohn lemma.

This exercise is adapted from [M-Z], to which the reader is referred for further results on topological groups.

*6. Define a set X as follows: For each even integer m, let L_m denote the line segment $m \times [-1, 0]$ in the plane. For each odd integer n and each integer $k \geq 2$, let $C_{n,k}$ denote the union of the line segments $(n + 1 - 1/k) \times [-1, 0]$ and $(n - 1 + 1/k) \times [-1, 0]$ and the semicircle

$$\{x \times y \mid (x - n)^2 + y^2 = (1 - 1/k)^2 \quad \text{and} \quad y \geq 0\}$$

in the plane. Let $p_{n,k}$ denote the topmost point $n \times (1 - 1/k)$ of this semicircle. Let X be the union of all the sets L_m and $C_{n,k}$, along with two extra points a and b. Topologize X by taking sets of the following four types as basis elements:

(i) The intersection of X with a horizontal open line segment that contains none of the points $p_{n,k}$.
(ii) A set formed from one of the sets $C_{n,k}$ by deleting finitely many points.
(iii) For each even integer m, the union of $\{a\}$ and the set of points $x \times y$ of X for which $x < m$.
(iv) For each even integer m, the union of $\{b\}$ and the set of points $x \times y$ of X for which $x > m$.

(a) Sketch X; show that these sets form a basis for a topology on X.
(b) Let f be a continuous real-valued function on X. Show that for any c, the set $f^{-1}(c)$ is a G_δ set in X. (This is true for any space X.) Conclude that the set $S_{n,k}$ consisting of those points p of $C_{n,k}$ for which $f(p) \neq f(p_{n,k})$ is countable. Choose $d \in [-1, 0]$ so that the line $y = d$ intersects none of the sets $S_{n,k}$. Show that for n odd,

$$f((n-1) \times d) = \lim_{k \to \infty} f(p_{n,k}) = f((n+1) \times d).$$

Conclude that $f(a) = f(b)$.
(c) Show that X is regular but not completely regular.

5-3 *The Stone–Čech Compactification*

We have already studied one way of compactifying a topological space X, the one-point compactification (§3-8); it is in some sense the minimal compactification of X. The Stone–Čech compactification of X is in some sense the maximal compactification of X. We introduct it here as an interesting application of the Tychonoff theorem. It has a number of applications in modern analysis, but these lie outside the scope of this book.

Definition. A **compactification** of a space X is a compact Hausdorff space Y containing X such that X is dense in Y (that is, such that $\bar{X} = Y$). Two compactifications Y_1 and Y_2 of X are said to be **equivalent** if there is a homeomorphism $h : Y_1 \to Y_2$ such that $h(x) = x$ for every $x \in X$.

In order for X to have a compactification, X must be completely regular. Conversely, every completely regular space has at least one compactification. One way of obtaining a compactification of X is as follows:

Let X be completely regular. Choose an imbedding $h : X \to Z$ of X in a compact Hausdorff space Z. Let X_0 denote the subspace $h(X)$ of Z, and let Y_0 denote its closure in Z. Then Y_0 is a compact Hausdorff space and $\bar{X}_0 = Y_0$; therefore, Y_0 is a compactification of X_0.

We now construct a space Y containing X such that the pair (X, Y) is homeomorphic to the pair (X_0, Y_0). Let us choose a set A disjoint from X that is in bijective correspondence with the set $Y_0 - X_0$ under some map $k : A \to Y_0 - X_0$. Define $Y = X \cup A$, and define a bijective correspondence $H : Y \to Y_0$ by the rule

$$H(x) = h(x) \quad \text{for } x \in X,$$
$$H(a) = k(a) \quad \text{for } a \in A.$$

Then topologize Y by declaring U to be open in Y if and only if $H(U)$ is open in Y_0. The map H is automatically a homeomorphism; and the space X is a subspace of Y because H equals the homeomorphism h when restricted to the subset X of Y.

The space Y is called the **compactification of X induced by the imbedding h.**

We summarize the preceding construction as follows: *If $h : X \to Z$ is an imbedding of X in the compact Hausdorff space Z, then h induces a compactification Y of X. It has the property that the imbedding h can be extended to an imbedding $H : Y \to Z$.*

In general, there are many different ways of compactifying a given space X. Consider for instance the following compactifications of the open interval $X = (0, 1)$:

EXAMPLE 1. Take the unit circle S^1 in R^2 and let $h : (0, 1) \to S^1$ be the map

$$h(t) = (\cos 2\pi t) \times (\sin 2\pi t).$$

The compactification induced by the imbedding h is equivalent to the one-point compactification of X.

EXAMPLE 2. Let Y be the space $[0, 1]$. Then Y is a compactification of X; it is obtained by "adding one point at each end of $(0, 1)$."

EXAMPLE 3. Consider the square $[-1, 1]^2$ in R^2 and let $h : (0, 1) \to [-1, 1]^2$ be the map

$$h(x) = x \times \sin(1/x).$$

The space $Y_0 = \overline{h(X)}$ is the topologist's sine curve (see Example 8 of §3-2). The imbedding h gives rise to a compactification of (0, 1) quite different from the other two. It is obtained by adding one point at the right-hand end of (0, 1), and an entire line segment of points at the left-hand end!

A basic problem that occurs in studying compactifications is the following:

If Y is a compactification of X, under what conditions can a continuous real-valued function f defined on X be extended continuously to Y?

The function f will have to be bounded if it is to be extendable, since its extension will carry the compact space Y into R and will thus be bounded. But boundedness is not enough, in general. Consider the following example:

EXAMPLE 4. Let $X = (0, 1)$. Consider the one-point compactification of X given in Example 1. A bounded continuous function $f: (0, 1) \longrightarrow R$ is extendable to this compactification if and only if the limits

$$\lim_{x \to 0+} f(x) \quad \text{and} \quad \lim_{x \to 1-} f(x)$$

exist and are equal.

For the "two-point compactification" of X considered in Example 2, the function f is extendable if and only if both these limits simply exist.

For the compactification of Example 3, extensions exist for a still broader class of functions. It is easy to see that f is extendable if both the above limits exist. But the function $f(x) = \sin(1/x)$ is also extendable to this compactification: Let H be the imbedding of Y in R^2 which equals h on the subspace X. Then the composite map

$$Y \xrightarrow{H} R \times R \xrightarrow{\pi_2} R$$

is the desired extension of f. For if $x \in X$, then $H(x) = h(x) = x \times \sin(1/x)$, so that $\pi_2(H(x)) = \sin(1/x)$, as desired.

There is something especially interesting about this last compactification. We constructed it by choosing an imbedding

$$h: (0, 1) \longrightarrow R^2$$

whose component functions were the functions x and $\sin(1/x)$. Then we found that both the functions x and $\sin(1/x)$ could be extended to the compactification. This suggests that if we have a whole *collection* of bounded continuous functions defined on (0, 1), we might use them as component functions of an imbedding of (0, 1) into R^J for some J, and thereby obtain a compactification for which every function in the collection is extendable.

This idea is the basic idea behind the Stone–Čech compactification. It is defined as follows:

Let X be a completely regular space. Let $\{f_\alpha\}_{\alpha \in J}$ be the collection of *all*

bounded continuous real-valued functions on X, indexed by some index set J. For each $\alpha \in J$, choose a closed interval I_α in R containing $f_\alpha(X)$. To be definite, choose

$$I_\alpha = [\text{glb} f_\alpha(X), \text{lub} f_\alpha(X)].$$

Then define $h : X \to \prod_{\alpha \in J} I_\alpha$ by the rule

$$h(x) = (f_\alpha(x))_{\alpha \in J}.$$

By the Tychonoff theorem, $\prod I_\alpha$ is compact. Because X is completely regular, the collection $\{f_\alpha\}$ separates points from closed sets in X. Therefore, by the Imbedding theorem of §4-4, h is an imbedding.

The compactification of X induced by h is called the **Stone-Čech compactification** of X. It is commonly denoted by $\beta(X)$.

Many mathematicians avoid using the arbitrary index set J when defining the Stone-Čech compactification. Instead, they use the collection $\mathcal{F}$ of all bounded continuous real-valued functions on X as its *own* index set, letting $J = \mathcal{F}$ and letting the indexing function $f: J \to \mathcal{F}$ be the identity function. This choice of indexing is not essential for the definition, however; and we feel it leads to notational confusion.

The crucial property of the Stone-Čech compactification is the following *extension property*:

Theorem 3.1. *Let* X *be completely regular; let* β(X) *be its Stone-Čech compactification. Then every bounded continuous real-valued function on* X *can be uniquely extended to a continuous real-valued function on* β(X).

Proof. The compactification $\beta(X)$ is induced by the imbedding $h: X \to \prod I_\alpha$ defined above. This means there is an imbedding $H : \beta(X) \to \prod I_\alpha$ that equals h when restricted to the subspace X of $\beta(X)$. Given a continuous bounded real-valued function on X, it equals f_β for some $\beta \in J$. Now if $\pi_\beta : \prod I_\alpha \to I_\beta$ is projection onto the βth coordinate, then the composite map $\pi_\beta \circ H: \beta(X) \to I_\beta$ is the desired extension of f_β. For if $x \in X$, we have

$$\pi_\beta(H(x)) = \pi_\beta(h(x)) = \pi_\beta((f_\alpha(x))_{\alpha \in J}) = f_\beta(x).$$

Uniqueness of the extension is a consequence of the following lemma. □

Lemma 3.2. *Let* A ⊂ X; *let* f : A → Z *be a continuous map of* A *into the Hausdorff space* Z. *There is at most one extension of* f *to a continuous function* g : Ā → Z.

Proof. This lemma was given as an exercise in §2-7; we give a proof here. Suppose that $g, g' : \bar{A} \to X$ are two different extensions of f; choose x so that $g(x) \neq g'(x)$. Let U and U' be disjoint neighborhoods of $g(x)$ and $g'(x)$, respectively. Choose a neighborhood V of x so that $g(V) \subset U$ and $g'(V) \subset$

U'. Now V intersects A in some point y; then $g(y) \in U$ and $g'(y) \in U'$. But since $y \in A$, we have $g(y) = f(y)$ and $g'(y) = f(y)$. This contradicts the fact that U and U' are disjoint. □

We now prove a theorem to the effect that the Stone–Čech compactification is essentially unique, and is characterized by the extension property.

Theorem 3.3. *Let* X *be completely regular. Let* Y_1 *and* Y_2 *be two compactifications of* X *having the extension property of Theorem* 3.1. *Then there is a homeomorphism* ϕ *of* Y_1 *onto* Y_2 *such that* $\phi(x) = x$ *for each* $x \in X$.

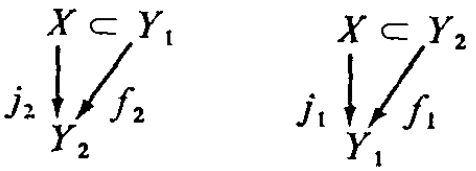
Proof. Step 1. We first prove the following fact: Suppose Y is a compactification of X having the extension property stated in Theorem 3.1. If Z is any compact Hausdorff space and $g : X \to Z$ is any continuous function, then g can be extended to a continuous function k mapping Y into Z.

To prove this fact, note that Z is completely regular, so that it can be imbedded in $[0, 1]^J$ for some J. So we may as well assume that $Z \subset [0, 1]^J$. Now consider $g : X \to Z \subset [0, 1]^J \subset R^J$. Each component function g_α of the map g is a continuous bounded real-valued function on X; by hypothesis, g_α can be extended to a continuous map k_α of Y into R. Define $k : Y \to R^J$ by setting $k(y) = (k_\alpha(y))_{\alpha \in J}$. The map k is continuous because R^J has the product topology. We assert that k actually maps Y into the subspace Z. For $g(X)$ is contained in Z, and $k(X) = g(X)$. Since Z is closed in R^J, it follows that $\overline{k(X)} \subset Z$. By continuity of k,

$$k(Y) = k(\bar{X}) \subset \overline{k(X)}.$$

Therefore, k maps Y into Z.

Step 2. Now we prove the theorem. Consider the inclusion mapping $j_2 : X \to Y_2$. It is a continuous map of X into the compact Hausdorff space Y_2. Because Y_1 has the extension property, we may, by Step 1, extend j_2 to a continuous map $f_2 : Y_1 \to Y_2$. Similarly, we may extend the inclusion map $j_1 : X \to Y_1$ to a continuous map $f_1 : Y_2 \to Y_1$ (because Y_2 has the extension property and Y_1 is compact Hausdorff).



The composite $f_1 \circ f_2 : Y_1 \to Y_1$ has the property that for every $x \in X$, one has $f_1(f_2(x)) = x$. Therefore, $f_1 \circ f_2$ is a continuous extension of the identity map $i_X : X \to X$. But the identity map of Y_1 is also a continuous extension of i_X. By uniqueness of extensions (Lemma 3.2), $f_1 \circ f_2$ must equal the identity map of Y_1. Similarly, $f_2 \circ f_1$ must equal the identity map of Y_2. Thus f_1 and f_2 are homeomorphisms. □

Exercises

1. Verify the statements made in Examples 1 and 4.

2. Show that the bounded continuous function $g : (0, 1) \longrightarrow R$ defined by $g(x) = \cos(1/x)$ cannot be extended to the compactification of Example 3. Define an imbedding $h : (0, 1) \longrightarrow R^3$ such that the functions x, $\sin(1/x)$, and $\cos(1/x)$ are all extendable to the compactification induced by h.

3. Under what conditions does a metrizable space have a metrizable compactification?

4. Let Y be an arbitrary compactification of X; let $\beta(X)$ be the Stone–Čech compactification. Show there is a continuous surjective closed map $g : \beta(X) \longrightarrow Y$ that equals the identity on X.

   [This exercise makes precise what we mean by saying that $\beta(X)$ is the "maximal" compactification of X. If you are familiar with quotient spaces, you will recognize that g is a quotient map. Thus every compactification of X is equivalent to a quotient space of $\beta(X)$.]

5. (a) Show that every continuous real-valued function defined on S_Ω is "eventually constant." [*Hint:* First prove that for each ϵ, there is an element α of S_Ω such that $|f(\beta) - f(\alpha)| < \epsilon$ for all $\beta > \alpha$. Then let $\epsilon = 1/n$ for $n \in Z_+$ and consider the corresponding points α_n.]
   (b) Show that the one-point compactification of S_Ω and the Stone–Čech compactification are equivalent.
   (c) Conclude that every compactification of S_Ω is equivalent to the one-point compactification.

6. Let X be completely regular. Show that X is connected if and only if $\beta(X)$ is connected. [*Hint:* If $X = A \cup B$ is a separation of X, let $f(x) = 0$ for $x \in A$ and $f(x) = 1$ for $x \in B$.]

7. Let X be discrete; consider $\beta(X)$.
   (a) Show that if $A \subset X$, then $\bar{A}$ and $\overline{X - A}$ are disjoint [closures taken in $\beta(X)$].
   (b) Show that if U is open in $\beta(X)$, then $\bar{U}$ is open in $\beta(X)$.
   (c) Show that $\beta(X)$ is totally disconnected.

8. Show that $\beta(Z_+)$ has cardinality at least as great as I^I, where $I = [0, 1]$. [*Hint:* I^I has a countable dense subset, by Exercise 13 of §4-1.]

9. Show that if $\beta(X) \neq X$, then $\beta(X)$ is not metrizable. [*Hint:* Show that if X is normal and y is a point of $\beta(X) - X$, then y is not the limit of a sequence of points of X.]

6. Metrization Theorems and Paracompactness

The Urysohn metrization theorem of Chapter 4 was the first step—a giant one—toward an answer to the question: When is a topological space metrizable? It gives conditions under which a space X is metrizable: that it be regular and have a countable basis. But mathematicians are never satisfied with a theorem if there is some hope of proving a stronger one. In the present case, one can hope to strengthen the theorem by finding conditions on X which are both necessary and sufficient for X to be metrizable, that is, conditions which are *equivalent* to metrizability.

We know that the regularity hypothesis in the Urysohn metrization theorem is a necessary one, but the countable basis condition is not. So the obvious thing to do is try to replace the countable basis condition by something weaker. Finding such a condition is a delicate task. The condition has to be strong enough to imply metrizability, and yet weak enough that all metrizable spaces satisfy it. In a situation like this, discovering the right hypothesis is more than half the battle.

The condition that was eventually formulated, by J. Nagata and Y. Smirnov independently, involves a new notion, that of local finiteness. We say that a collection $\mathcal{A}$ of subsets of a space X is *locally finite* if every point of X has a neighborhood that intersects only finitely many elements of $\mathcal{A}$.

Now one way of expressing the condition that the basis $\mathcal{B}$ is countable is to say that $\mathcal{B}$ can be expressed in the form

$$\mathcal{B} = \bigcup_{n \in Z_+} \mathcal{B}_n,$$

where each collection $\mathcal{B}_n$ is finite. This is an awkward way of saying that $\mathcal{B}$ is countable, but it suggests how to formulate a weaker version of it. The Nagata–Smirnov condition is to require that the basis $\mathcal{B}$ can be expressed in the form

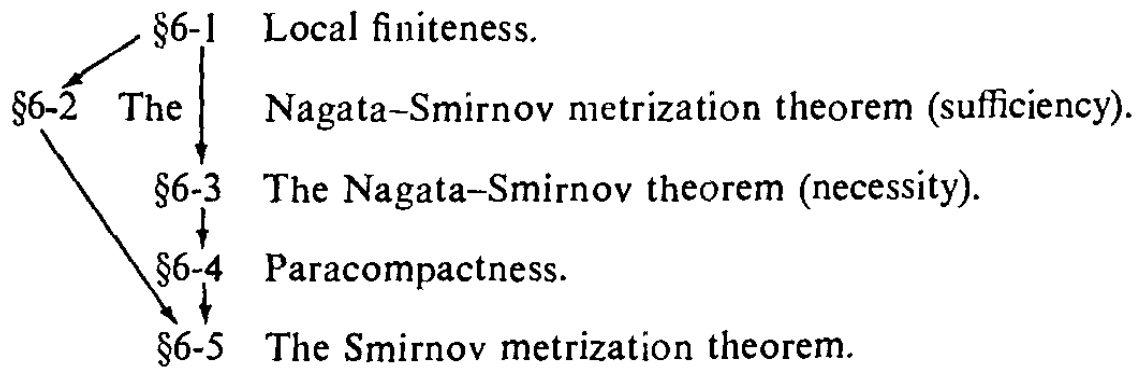
$$\mathcal{B} = \bigcup_{n \in Z_+} \mathcal{B}_n,$$

where each collection $\mathcal{B}_n$ is *locally* finite. We say that such a collection $\mathcal{B}$ is *countably locally finite*. Surprisingly enough, this condition, along with regularity, is both necessary and sufficient for metrizability of X.

In §6-1 we study the notion of local finiteness. In §6-2 we show that the Nagata–Smirnov condition for X, along with regularity, implies metrizability of X; the proof is an adaptation of the proof of the Urysohn metrization theorem. In §6-3 we prove that metrizability implies the Nagata–Smirnov condition; this proof uses the well-ordering theorem.

There is another concept in topology which involves the notion of local finiteness. It is a generalization of the concept of compactness called "paracompactness." Although of fairly recent origin, it has proved useful in many parts of mathematics. We introduce it here so that we can give another set of necessary and sufficient conditions for a space X to be metrizable. It turns out that X is metrizable if and only if it is both paracompact and locally metrizable. This we prove in §6-5.

Some of the sections of this chapter are independent of one another. The dependence among them is expressed in the following diagram:



6-1 Local Finiteness

In this section we prove some elementary properties of locally finite collections.

Definition. Let X be a topological space. A collection $\mathcal{A}$ of subsets of X is said to be **locally finite** if every point of X has a neighborhood that intersects only finitely many elements of $\mathcal{A}$.

EXAMPLE 1. The collection of intervals

$$\mathcal{A} = \{(n, n+2) \mid n \in Z\}$$

is locally finite in the topological space R, as you can check. So is the collection

$$\mathcal{B} = \{(n, 2n) \mid n \in Z_+\}.$$

On the other hand, the collection

$$\mathcal{C} = \{(0, 1/n) \mid n \in Z_+\}$$

is not locally finite in R, nor is the collection

$$\mathcal{D} = \{(1/(n+1), 1/n) \mid n \in Z_+\}.$$

Lemma 1.1. Let $\mathcal{A}$ be a locally finite collection of subsets of X. Then:

(a) *Any subcollection of $\mathcal{A}$ is locally finite.*

(b) *The collection $\mathcal{B} = \{\bar{A}\}_{A \in \mathcal{A}}$ of the closures of the elements of $\mathcal{A}$ is locally finite.*

(c) $\overline{\bigcup_{A \in \mathcal{A}} A} = \bigcup_{A \in \mathcal{A}} \bar{A}$.

Proof. Statement (a) is trivial. To prove (b), note that any open set U which intersects the set $\bar{A}$ necessarily intersects A. Therefore, if U is a neighborhood of x that intersects only finitely many elements A of $\mathcal{A}$, then U can intersect at most the same number of sets of the collection $\mathcal{B}$. (It might intersect fewer sets of $\mathcal{B}$, since $\bar{A}_1$ and $\bar{A}_2$ can be equal even though A_1 and A_2 are different.)

To prove (c), let Y denote the union of the elements of $\mathcal{A}$:

$$\bigcup_{A \in \mathcal{A}} A = Y.$$

In general, $\bigcup \bar{A} \subset \bar{Y}$; we prove the reverse inclusion, under the assumption of local finiteness. Let $x \in \bar{Y}$; let U be a neighborhood of x that intersects only finitely many elements of $\mathcal{A}$, say $A_1, \ldots, A_k$. We assert that x belongs to one of the sets $\bar{A}_1, \ldots, \bar{A}_k$, and hence belongs to $\bigcup \bar{A}$. For otherwise, the set $U - \bar{A}_1 - \cdots - \bar{A}_k$ would be a neighborhood of x that intersects no element of $\mathcal{A}$ and hence does not intersect Y, contrary to the assumption that $x \in \bar{Y}$. □

In the exercises of §2-7 and §4-5, we defined a concept of local finiteness for an indexed family of subsets of X. The indexed family $\{A_\alpha\}_{\alpha \in J}$ is said to be a *locally finite indexed family* if every $x \in X$ has a neighborhood that intersects A_α for only finitely many values of α. What is the relation between the two formulations of local finiteness? It is easy to see that $\{A_\alpha\}_{\alpha \in J}$ is a locally finite indexed family if and only if it is locally finite as a *collection* of sets and each nonempty subset A of X equals A_α for at most finitely many values of α.

We shall not be concerned with locally finite indexed families in this chapter, except in one or two exercises.

Definition. A collection $\mathcal{B}$ of subsets of X is said to be **countably locally finite** if $\mathcal{B}$ can be written as the countable union of collections $\mathcal{B}_n$, each of which is locally finite.

Most authors use the term "σ-locally finite" for this concept. The σ comes from measure theory and stands for the phrase "countable union of." Note that both a countable collection and a locally finite collection are countably locally finite.

Exercises

1. Check the statements in Example 1.
2. Check the relation stated above between local finiteness for a collection of sets and for an indexed family of sets.
3. Find a point-finite open covering $\mathcal{A}$ of R that is not locally finite. [The collection $\mathcal{A}$ is *point-finite* if each point of R lies in only finitely many elements of $\mathcal{A}$.]
4. Give an example of a collection of sets $\mathcal{A}$ that is not locally finite, such that the collection $\mathcal{B} = \{\bar{A} \mid A \in \mathcal{A}\}$ is locally finite.
5. Show that if X has a countable basis, a collection $\mathcal{A}$ of subsets of X is countably locally finite if and only if it is countable.
6. Take R^ω in the box topology. Find a collection of subsets of R^ω that is countably locally finite but is neither countable nor locally finite.
7. Many spaces have *countable* bases; but no Hausdorff space has a *locally finite* basis unless it is discrete Prove this fact.
8. Find a space that has a countably locally finite basis but does not have a countable basis.

6-2 The Nagata–Smirnov Metrization Theorem (sufficiency)

Now we prove that regularity and the Nagata–Smirnov condition suffice to prove a space metrizable.

The proof follows very closely the *second* proof we gave of the Urysohn metrization theorem. In this proof we constructed a map of the space X into R^ω that was an imbedding relative to the uniform metric $\bar{\rho}$ on R^ω. So let us review the major elements of that proof. The first step of the proof was to prove that every regular space with a countable basis is normal. The second step was to construct a countable collection $\{f_n\}$ of real-valued functions on X that separated points from closed sets. And the third step was to use the functions f_n to imbed X in the metric space $(R^\omega, \bar{\rho})$.

Each of these steps needs to be generalized in order to prove the Nagata–Smirnov theorem. First, we prove that a regular space X with a countably locally finite basis is normal. Second, we construct a certain collection of real-valued functions $\{f_\alpha\}$ on X. Third, we use these functions to imbed X in the metric space $(R^J, \bar{\rho})$ for some J.

Before we start, we need to recall a notion we have already introduced in the exercises, that of a G_δ set.

Definition. A subset A of a space X is called a **G_δ set** in X if it equals the intersection of a countable collection of open subsets of X.

EXAMPLE 1. Each open subset of X is a G_δ set, trivially. In a first-countable Hausdorff space, each one-point set is a G_δ set. The one-point subset $\{\Omega\}$ of $\bar{S}_\Omega$ is not a G_δ set, as you can check.

EXAMPLE 2. In a metric space X, each closed set is a G_δ set. Given $A \subset X$, define

$$U(A, \epsilon) = \bigcup_{x \in A} B(x, \epsilon).$$

If A is closed, you can check that

$$A = \bigcap_{n \in Z_+} U(A, 1/n)$$

Lemma 2.1. *Let* X *be a regular space with a basis* $\mathcal{B}$ *that is countably locally finite. Then* X *is normal, and every closed set in* X *is a* G_δ *set in* X.

Proof. Step 1. Let W be open in X. We show there is a countable collection $\{U_n\}$ of open sets of X such that

$$W = \bigcup U_n = \bigcup \bar{U}_n.$$

Since the basis $\mathcal{B}$ for X is countably locally finite, we can write $\mathcal{B} = \bigcup \mathcal{B}_n$, where each collection $\mathcal{B}_n$ is locally finite. Let $\mathcal{C}_n$ be the collection of those basis elements B such that $B \in \mathcal{B}_n$ and $\bar{B} \subset W$. Then $\mathcal{C}_n$ is locally finite, being a subcollection of $\mathcal{B}_n$. Define

$$U_n = \bigcup_{B \in \mathcal{C}_n} B.$$

Then U_n is an open set, and by Lemma 1.1,

$$\bar{U}_n = \bigcup_{B \in \mathcal{C}_n} \bar{B}.$$

Therefore, $\bar{U}_n \subset W$, so that

$$\bigcup U_n \subset \bigcup \bar{U}_n \subset W.$$

We assert that equality holds. Given $x \in W$, there is by regularity a basis element $B \in \mathcal{B}$ such that $x \in B$ and $\bar{B} \subset W$. Now $B \in \mathcal{B}_n$ for some n. Then $B \in \mathcal{C}_n$ by definition, so that $x \in U_n$. Thus $W \subset \bigcup U_n$, as desired.

Step 2. We show that every closed set C in X is a G_δ set in X. Given C, let $W = X - C$. By Step 1, there are sets U_n in X such that $W = \bigcup \bar{U}_n$. Then

$$C = \bigcap (X - \bar{U}_n),$$

so that C equals a countable intersection of open sets of X.

Step 3. Now we show X is normal. Let C and D be disjoint closed sets in X. Applying Step 1 to the open set $X - D$, we construct a countable collec-

tion $\{U_n\}$ of open sets such that $\bigcup U_n = \bigcup \bar{U}_n = X - D$. Then $\{U_n\}$ covers C and each set $\bar{U}_n$ is disjoint from D. Similarly, we construct a countable covering $\{V_n\}$ of D by open sets whose closures are disjoint from C.

Now we are back in the situation which arose in the proof that a regular space with a countable basis is normal (Theorem 2.5 of Chapter 4). We can repeat that proof *verbatim*. Define

$$U_n' = U_n - \bigcup_{i=1}^{n} \bar{V}_i \quad \text{and} \quad V_n' = V_n - \bigcup_{i=1}^{n} \bar{U}_i.$$

Then the sets

$$U' = \bigcup_{n \in Z_+} U_n' \quad \text{and} \quad V' = \bigcup_{n \in Z_+} V_n'$$

are disjoint open sets about C and D, respectively. □

Theorem 2.2. Let X *be a regular space with a basis* $\mathcal{B}$ *that is countably locally finite. Then* X *is metrizable.*

Proof. Step 1. First we show that if W is open in X, there is a continuous function $f: X \to [0, 1]$ such that $f(x) > 0$ for $x \in W$ and $f(x) = 0$ for $x \notin W$.

By the preceding lemma, each closed set of X is a countable intersection of open sets of X. Passing to complements, it follows that the open set W is a countable union of closed sets A_n of X. Using normality, choose for each positive integer n, a continuous function $f_n : X \to [0, 1]$ such that $f_n(A_n) = \{1\}$ and $f_n(X - W) = \{0\}$. Define $f(x) = \sum f_n(x)/2^n$. The series converges uniformly, by comparison with $\sum 1/2^n$, so that f is continuous. Also, f is positive on W and vanishes outside W.

Step 2. Let $\mathcal{B} = \bigcup \mathcal{B}_n$, where each collection $\mathcal{B}_n$ is locally finite. For each positive integer n, and each basis element $B \in \mathcal{B}_n$, choose a continuous function

$$f_{n,B}: X \longrightarrow [0, 1/n]$$

such that $f_{n,B}(x) > 0$ for $x \in B$ and $f_{n,B}(x) = 0$ for $x \notin B$. The collection $\{f_{n,B}\}$ separates points from closed sets in X: Given a point x_0 and a neighborhood U of x_0, there is a basis element B such that $x_0 \in B \subset U$. Then $B \in \mathcal{B}_n$ for some n, so that $f_{n,B}(x_0) > 0$ and $f_{n,B}$ vanishes outside U.

Let J be the subset of $Z_+ \times \mathcal{B}$ consisting of all pairs (n, B) such that B belongs to $\mathcal{B}_n$. Define

$$F: X \longrightarrow [0, 1]^J$$

by the equation

$$F(x) = (f_{n,B}(x))_{(n,B) \in J}.$$

Relative to the product topology on $[0, 1]^J$, the map F is an imbedding, by the Imbedding theorem of §4-4.

Of course, $[0, 1]^J$ is not metrizable in general, so we have not yet proved the metrization theorem.

Step 3. Now we give $[0, 1]^J$ the topology induced by the uniform metric $\bar{\rho}$ and show that F is an imbedding relative to this topology as well. The uniform topology is finer (larger) than the product topology. Therefore, relative to the uniform metric, the map F is injective and carries open sets of X onto open sets of the image space $Z = F(X)$. We must give a separate proof that F is continuous.

Note that on the subspace $[0, 1]^J$ of R^J, the uniform metric equals the metric

$$\rho((x_\alpha), (y_\alpha)) = \text{lub } \{|x_\alpha - y_\alpha|\}.$$

To prove continuity, we take a point x_0 of X and a number $\epsilon > 0$, and find a neighborhood W of x_0 such that

$$x \in W \Longrightarrow \rho(F(x), F(x_0)) < \epsilon.$$

Let n be fixed for the moment. Choose a neighborhood U_n of x_0 that intersects only finitely many elements of the collection $\mathcal{B}_n$. This means that as B ranges over $\mathcal{B}_n$, all but finitely many of the functions $f_{n,B}$ are identically equal to zero on U_n. Because each function $f_{n,B}$ is continuous, we can now choose a neighborhood V_n of x_0 contained in U_n on which each of the remaining functions $f_{n,B}$, for $B \in \mathcal{B}_n$, varies by at most $\epsilon/2$.

Choose such a neighborhood V_n of x_0 for each $n \in Z_+$. Then choose N so that $1/N \leq \epsilon/2$, and define $W = V_1 \cap \cdots \cap V_N$. We assert that W is the desired neighborhood of x_0. Let $x \in W$. If $n \leq N$, then

$$|f_{n,B}(x) - f_{n,B}(x_0)| \leq \epsilon/2$$

because the function $f_{n,B}$ either vanishes identically or varies by at most $\epsilon/2$ on W. If $n > N$, then

$$|f_{n,B}(x) - f_{n,B}(x_0)| \leq 1/n < \epsilon/2$$

because $f_{n,B}$ maps X into $[0, 1/n]$. Therefore,

$$\rho(F(x), F(x_0)) \leq \epsilon/2 < \epsilon,$$

as desired. □

Exercises

1. (a) Show that in a metric space, every closed set is a G_δ set.
   (b) Show that the irrationals are a G_δ set in R.

2. A subset W of X is said to be an "F_σ set" in X if W equals a countable union of closed sets of X. Show that W is an F_σ set in X if and only if $X - W$ is a G_δ set in X.

   [The terminology comes from the French. The "F" stands for "fermé," which means "closed," and the "σ" for "somme," which means "union."]

3. Let (X, d) be a metric space. Show that if U is open in X, then the function

$$f(x) = d(x, X - U)$$

is positive on U and vanishes outside U. (The function $d(x, A)$ is defined in Exercise 4 of §3-6.)

6-3 *The Nagata–Smirnov Theorem (necessity)*

We now prove that every metrizable space has a countably locally finite basis. This will complete the proof of the Nagata–Smirnov metrization theorem.

Definition. Let $\mathcal{A}$ be a collection of subsets of the space X. A collection $\mathcal{B}$ of subsets of X is said to be a **refinement** of $\mathcal{A}$ (or is said to **refine** $\mathcal{A}$) if for each element B of $\mathcal{B}$, there is an element A of $\mathcal{A}$ containing B. If the elements of $\mathcal{B}$ are open sets, we call $\mathcal{B}$ an **open refinement** of $\mathcal{A}$; if they are closed sets, we call $\mathcal{B}$ a **closed refinement**.

The first step in proving the necessity of the Nagata–Smirnov condition is the following lemma. It will be generalized in the next section.

Lemma 3.1. *Let* X *be a metrizable space. If* $\mathcal{A}$ *is an open covering of* X, *then there is a collection* $\mathcal{D}$ *of subsets of* X *such that:*

(1) $\mathcal{D}$ *is an open covering of* X.

(2) $\mathcal{D}$ *is a refinement of* $\mathcal{A}$.

(3) $\mathcal{D}$ *is countably locally finite.*

Proof. We shall need to use the well-ordering theorem in order to prove this theorem. Choose a well-ordering $<$ for the collection $\mathcal{A}$. Let us denote the elements of $\mathcal{A}$ generically by the letters $U, V, W, \ldots$.

Choose a metric for X. Let n be a positive integer, fixed for the moment. Given an element U of $\mathcal{A}$, let us define $S_n(U)$ to be the subset of U obtained by "shrinking" U a distance of $1/n$. More precisely, let

$$S_n(U) = \{x \mid B(x, 1/n) \subset U\}.$$

(It happens that $S_n(U)$ is a closed set, but that is not important for our purposes.) Now we use the well-ordering $<$ of $\mathcal{A}$ to pass to a still smaller set. For each U in $\mathcal{A}$, define

$$S_n'(U) = S_n(U) - \bigcup_{V<U} V.$$

The situation where $\mathcal{A}$ consists of the three sets $U < V < W$ is pictured in Figure 1. Just as the figure suggests, the sets we have formed are disjoint. Indeed, we assert that they are separated by a distance of at least $1/n$. That is,

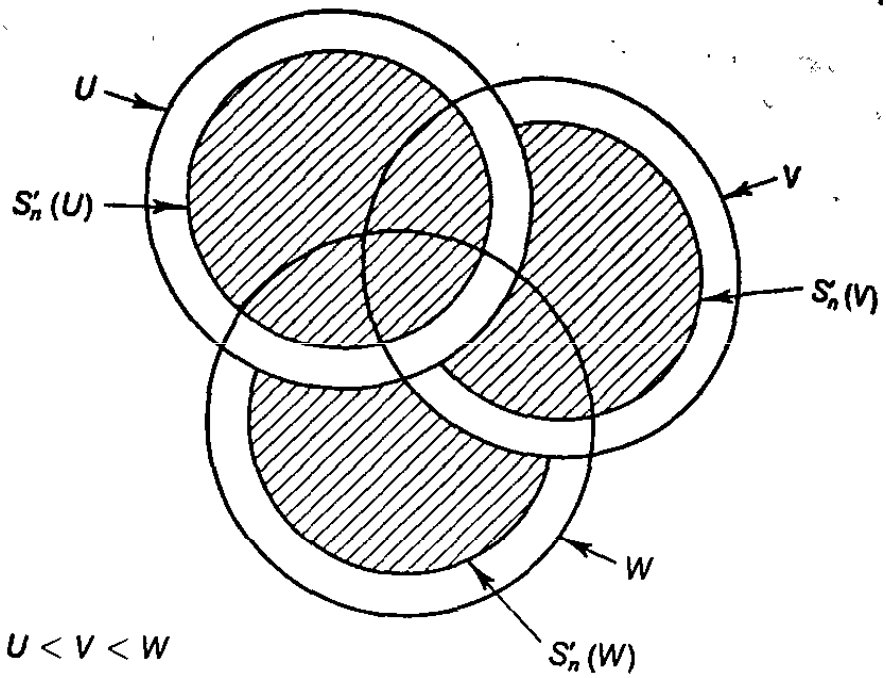


Figure 1 $U < V < W$

if V and W are distinct elements of $\mathcal{A}$, we assert that

$$(*) \qquad x \in S_n'(V) \text{ and } y \in S_n'(W) \implies d(x, y) \geq 1/n.$$

To prove this fact, assume the notation has been so chosen that $V < W$. Now $x \in S_n'(V)$ implies that $x \in S_n(V)$. And $y \in S_n'(W)$ implies by definition that $y \notin V$ (since $V < W$). Since $x \in S_n(V)$ and $y \notin V$, we must have $d(x, y) \geq 1/n$.

The sets $S_n'(U)$ are not yet the ones we want, for we do not know that they are open sets. (In fact, they are closed.) So let us expand each of them slightly to obtain an open set $E_n(U)$. Specifically, let $E_n(U)$ be the $1/3n$ neighborhood of $S_n'(U)$; that is,

$$E_n(U) = \bigcup \{B(x, 1/3n) \mid x \in S_n'(U)\}.$$

In the case $U < V < W$, we have the situation pictured in Figure 2. As the figure suggests, the sets we have formed are disjoint, and indeed they are separated by a distance of at least $1/3n$. That is, if V and W are distinct elements of $\mathcal{A}$, we assert that

$$x \in E_n(V) \text{ and } y \in E_n(W) \implies d(x, y) \geq 1/3n.$$

This follows at once from (*) and the triangle inequality. Note also that for each $V \in \mathcal{A}$, the set $E_n(V)$ is contained in V.

Now let us define

$$\mathcal{E}_n = \{E_n(U) \mid U \in \mathcal{A}\}.$$

We claim that $\mathcal{E}_n$ is a locally finite collection of open sets and that $\mathcal{E}_n$ refines $\mathcal{A}$. The fact that $\mathcal{E}_n$ refines $\mathcal{A}$ comes from the fact that $E_n(V) \subset V$ for each $V \in \mathcal{A}$. The fact that $\mathcal{E}_n$ is locally finite comes from the fact that for any x in X, the $1/6n$ neighborhood of x can intersect at most *one* element of $\mathcal{E}_n$.

Of course, the collection $\mathcal{E}_n$ will not cover X. (Figure 2 illustrates that fact.)

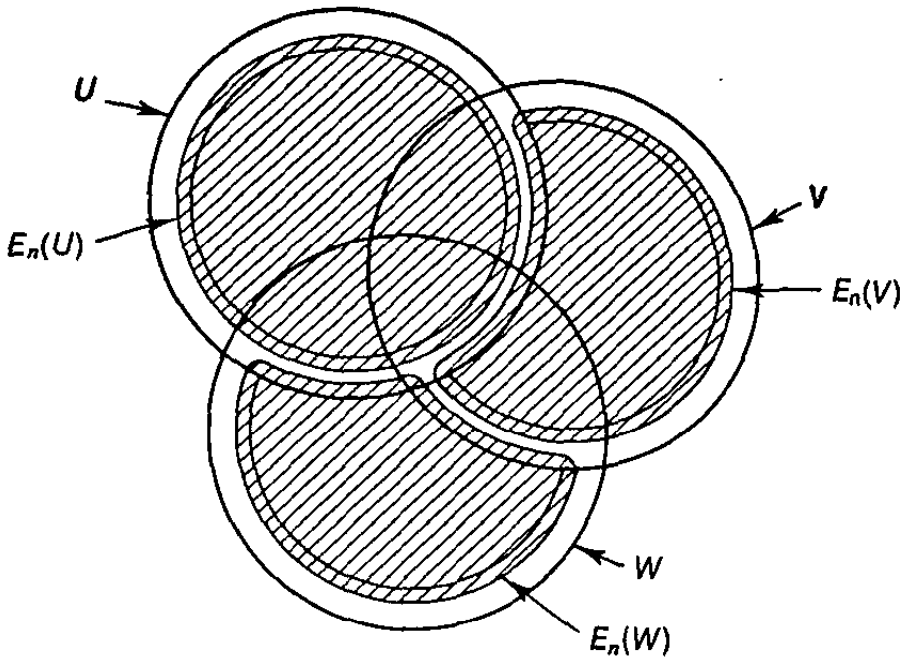


Figure 2

But we assert that the collection

$$\mathcal{E} = \bigcup_{n \in Z_+} \mathcal{E}_n$$

does cover X.

Let x be a point of X. The collection $\mathcal{A}$ with which we began covers X; let us choose U to be the first element of $\mathcal{A}$ (in the well-ordering $<$) that contains x. Since U is open, we can choose n so that $B(x, 1/n) \subset U$. Then, by definition, $x \in S_n(U)$. Now because U is the first element of $\mathcal{A}$ that contains x, the point x belongs to $S'_n(U)$. Then x also belongs to the element $E_n(U)$ of $\mathcal{E}_n$, as desired.

The lemma follows; $\mathcal{E}$ is the required collection of open sets. □

Theorem 3.2. *Let* X *be a metrizable space. Then* X *has a basis that is countably locally finite.*

Proof. Choose a metric for X. Given m, let $\mathcal{A}^m$ be the open covering of X by all open balls of radius $1/m$;

$$\mathcal{A}^m = \{B(x, 1/m) \mid x \in X\}.$$

By the preceding lemma, there is an open covering $\mathcal{D}^m$ of X refining $\mathcal{A}^m$ such that $\mathcal{D}^m$ is countably locally finite. Note that each element of $\mathcal{D}^m$ has diameter at most $2/m$. Let

$$\mathcal{D} = \bigcup_{m \in Z_+} \mathcal{D}^m.$$

Because each collection $\mathcal{D}^m$ is a countable union of locally finite collections, so is $\mathcal{D}$. We assert that $\mathcal{D}$ is a basis for X; then our theorem is proved.

We prove that given $x \in X$ and given $\epsilon > 0$, there is an element D of $\mathcal{D}$ containing x and contained in $B(x, \epsilon)$. First choose m so that $1/m < \epsilon/2$. Then because $\mathcal{D}^m$ covers X, we can choose an element D of $\mathcal{D}^m$ containing x. Since D contains x and has diameter at most $2/m < \epsilon$, it is contained in $B(x, \epsilon)$.

It follows from Lemma 2.3 of Chapter 2 that $\mathcal{D}$ is a basis for X. □

Exercises

1. Let $\mathcal{A}$ be the following collection of subsets of R:

$$\mathcal{A} = \{(n, n+2) \mid n \in Z\}.$$

Which of the following collections refines $\mathcal{A}$?

$$\mathcal{B} = \{(x, x+1) \mid x \in R\},$$
$$\mathcal{C} = \{(n, n+\tfrac{3}{2}) \mid n \in Z\},$$
$$\mathcal{D} = \{(x, x+\tfrac{3}{2}) \mid x \in R\}.$$

2. A collection $\mathcal{A}$ of subsets of X is said to be **locally discrete** if each point of X has a neighborhood that intersects at most one element of $\mathcal{A}$. A collection $\mathcal{B}$ is **countably locally discrete** (or "σ-locally discrete") if it equals a countable union of locally discrete collections. Prove the following:

Theorem (*Bing metrization theorem*). *A space X is metrizable if and only if it is regular and has a basis that is countably locally discrete.*

6-4 Paracompactness

The concept of paracompactness is one of the most useful generalizations of compactness that has been discovered in recent years. Particularly is it useful for applications in algebraic topology and differential geometry. We shall give just one application, a metrization theorem, which we prove in the next section.

Many of the spaces that are familiar to us already are paracompact. For instance, every compact Hausdorff space is paracompact; this will be an immediate consequence of the definition. It is also true that every metrizable space is paracompact; this is a theorem of A. H. Stone, which we shall prove. Thus the class of paracompact spaces includes the two most important classes of spaces we have studied, the compact Hausdorff spaces and the metrizable spaces. It includes other spaces as well. (See Exercise 2.)

To see how paracompactness generalizes compactness, we recall the definition of compactness: A space X is said to be *compact* if every open covering $\mathcal{A}$ of X contains a finite subcollection $\mathcal{A}'$ that covers X. An equivalent way of saying this is the following:

A space X is compact if every open covering $\mathcal{A}$ of X has a finite open refinement $\mathcal{B}$ that covers X.

This definition is equivalent to the usual one; given such a refinement $\mathcal{B}$, one can choose for each element of $\mathcal{B}$ an element of $\mathcal{A}$ containing it; in this way one obtains a finite subcollection of $\mathcal{A}$ that covers X.

This new formulation of compactness is an awkward one, but it suggests a way to generalize:

Definition. A space X is **paracompact** if it is Hausdorff and if every open covering $\mathcal{A}$ of X has a locally finite open refinement $\mathcal{B}$ that covers X.

EXAMPLE 1. Any compact Hausdorff space is paracompact, trivially. The real line R is a paracompact space that is not compact. The fact that R is paracompact is a consequence of the theorem that every metrizable space is paracompact. But a direct proof can be given as follows:

Suppose we are given an open covering $\mathcal{A}$ of R. For each integer n, choose a finite number of elements of $\mathcal{A}$ that cover the interval $[n, n+1]$ and intersect each one with the open interval $(n-1, n+2)$. Let the resulting collection of open sets be denoted $\mathcal{B}_n$. Then the collection

$$\mathcal{B} = \bigcup_{n \in Z} \mathcal{B}_n$$

is a locally finite open refinement of $\mathcal{A}$ that covers R, as you can check.

Some of the properties of a paracompact space are similar to those of a compact Hausdorff space. For instance, a subspace of a paracompact space is not necessarily paracompact; but a closed subspace is paracompact. Also, a paracompact space is necessarily normal. In other ways, a paracompact space is not similar to a compact Hausdorff space; in particular, the product of two paracompact spaces need not be paracompact.

One of the most useful properties of a paracompact space is the fact that given an indexed open covering $\{U_\alpha\}$ of a paracompact space X, there exists a partition of unity dominated by $\{U_\alpha\}$. A proof of this fact was outlined in the exercises of §4-5. We shall prove in this section the other properties of paracompact spaces we have mentioned.

Theorem 4.1. *Every paracompact space* X *is normal.*

Proof. The proof is somewhat similar to the proof that a compact Hausdorff space is normal.

First one proves regularity. Let a be a point of X and let B be a closed set of X disjoint from a. The Hausdorff condition enables us to choose, for each b in B, an open set U_b about b whose closure is disjoint from a. Cover X by the open sets U_b, along with the open set $X - B$; take a locally finite open refinement $\mathcal{C}$ that covers X. Form the subcollection $\mathcal{D}$ of $\mathcal{C}$ consisting of every element of $\mathcal{C}$ that intersects B. Then $\mathcal{D}$ covers B. Furthermore, if $D \in \mathcal{D}$, then $\bar{D}$ is disjoint from a. For D intersects B, so it lies in some set U_b, whose closure is disjoint from a. Let

$$V = \bigcup_{D \in \mathcal{D}} D;$$

then V is an open set in X containing B. Because $\mathcal{D}$ is locally finite,

$$\bar{V} = \bigcup_{D \in \mathcal{D}} \bar{D},$$

so that $\bar{V}$ is disjoint from a. Thus regularity is proved.

To prove normality, one merely repeats the same argument, replacing a by the closed set A throughout and replacing the Hausdorff condition by regularity. □

Theorem 4.2. (a) *Every closed subspace of a paracompact space is paracompact.*

(b) *An arbitrary subspace of a paracompact space need not be paracompact.*

(c) *A product of paracompact spaces need not be paracompact.*

Proof. (a) Let Y be a closed subspace of the paracompact space X; let $\mathcal{A}$ be a covering of Y by sets open in Y. For each $A \in \mathcal{A}$, choose an open set A' of X such that $A' \cap Y = A$. Cover X by the open sets A', along with the open set $X - Y$. Let $\mathcal{B}$ be a locally finite open refinement of this covering that covers X. The collection

$$\mathcal{C} = \{B \cap Y \mid B \in \mathcal{B}\}$$

is the required locally finite open refinement of $\mathcal{A}$.

Examples demonstrating (b) and (c) follow. □

EXAMPLE 2. The space $\bar{S}_\Omega \times \bar{S}_\Omega$ is compact Hausdorff, so it is paracompact. The subspace $S_\Omega \times \bar{S}_\Omega$ is not paracompact, for it is not even normal (see §4-2).

EXAMPLE 3. The space R_l is paracompact; we leave the proof to the exercises. But R_l^2 is not paracompact, for it is not normal (see §4-2).

Theorem 4.3 (Stone's theorem). Every metrizable space is paracompact.

Proof. Let X be a metrizable space. We already know from Lemma 3.1 that every open covering of X has an open refinement that covers X and is *countably* locally finite. It remains to prove that the latter condition implies that every open covering of X has an open refinement that covers X and is locally finite. This is a consequence of the following lemma. □

Lemma 4.4. Let X *be regular. Then the following conditions on* X *are equivalent:*

Every open covering of X *has a refinement that is:*

(1) *An open covering of* X *and countably locally finite.*

(2) *A covering of* X *and locally finite.*

(3) *A closed covering of* X *and locally finite.*

(4) *An open covering of* X *and locally finite.*

Proof. It is trivial that (4) $\Rightarrow$ (1). What we need for Stone's theorem is the converse. In order to prove the converse, we must go through the steps (1) $\Rightarrow$ (2) $\Rightarrow$ (3) $\Rightarrow$ (4) anyway, so we have for convenience listed these conditions in the statement of the lemma.

(1) $\Rightarrow$ (2). Let $\mathcal{A}$ be an open covering of X. Let $\mathcal{B}$ be an open refinement

of $\mathcal{A}$ that covers X and is countably locally finite; let

$$\mathcal{B} = \bigcup \mathcal{B}_n$$

where each $\mathcal{B}_n$ is locally finite. Denote the elements of $\mathcal{B}$ generically by $U, V, W, \ldots$.

Now we apply essentially the same sort of shrinking trick we have used before to make sets from different $\mathcal{B}_n$'s disjoint. Given i, let

$$V_i = \bigcup_{U \in \mathcal{B}_i} U.$$

Then for each $n \in Z_+$ and each element U of $\mathcal{B}_n$, define

$$S_n(U) = U - \bigcup_{i<n} V_i.$$

[Note that $S_n(U)$ is not necessarily open, nor closed.] Let

$$\mathcal{C}_n = \{S_n(U) \mid U \in \mathcal{B}_n\}.$$

Then $\mathcal{C}_n$ is a refinement of $\mathcal{B}_n$, because $S_n(U) \subset U$ for each $U \in \mathcal{B}_n$.

Let $\mathcal{C} = \bigcup \mathcal{C}_n$. We assert that $\mathcal{C}$ is the required locally finite refinement of $\mathcal{A}$, covering X.

Let x be a point of X. We wish to prove that x lies in an element of $\mathcal{C}$, and that x has a neighborhood intersecting only finitely many elements of $\mathcal{C}$. Consider the covering $\mathcal{B} = \bigcup \mathcal{B}_n$; let N be the smallest integer such that x lies in an element of $\mathcal{B}_N$. Let U be an element of $\mathcal{B}_N$ containing x. First, note that since x lies in no element of $\mathcal{B}_i$ for $i < N$, the point x lies in the element $S_N(U)$ of $\mathcal{C}$. Second, note that since each collection $\mathcal{B}_n$ is locally finite, we can choose for each $n = 1, \ldots, N$ a neighborhood W_n of x that intersects only finitely many elements of $\mathcal{B}_n$. Now if W_n intersects the element $S_n(V)$ of $\mathcal{C}_n$, it must intersect the element V of $\mathcal{B}_n$, since $S_n(V) \subset V$. Therefore, W_n intersects only finitely many elements of $\mathcal{C}_n$. Furthermore, because U is in $\mathcal{B}_N$, U intersects no element of $\mathcal{C}_n$ for $n > N$. As a result, the neighborhood

$$W_1 \cap W_2 \cap \cdots \cap W_N \cap U$$

of x intersects only finitely many elements of $\mathcal{C}$.

(2) $\Rightarrow$ (3). Let $\mathcal{A}$ be an open covering of X. Let $\mathcal{B}$ be the collection of all open sets U of X such that $\bar{U}$ is contained in an element of $\mathcal{A}$. By regularity, $\mathcal{B}$ covers X. Using (2), we can find a refinement $\mathcal{C}$ of $\mathcal{B}$ that covers X and is locally finite. Let

$$\mathcal{D} = \{\bar{C} \mid C \in \mathcal{C}\}.$$

Then $\mathcal{D}$ also covers X; it is locally finite by Lemma 1.1; and it refines $\mathcal{A}$.

(3) $\Rightarrow$ (4). Let $\mathcal{A}$ be an open covering of X. Using (3), choose $\mathcal{B}$ to be a refinement of $\mathcal{A}$ that covers X and is locally finite. (We can take $\mathcal{B}$ to be a closed refinement if we like, but that is irrelevant.) We seek to expand each element B of $\mathcal{B}$ slightly to an open set, making the expansion slight enough that the resulting collection of open sets will be still be locally finite and will still refine $\mathcal{A}$.

This step involves a new trick. The previous trick, used several times, consisted of ordering the sets in some way and forming a new set by subtracting off all the previous ones.† That trick shrinks the sets; to expand them we need something different. We shall introduce an auxiliary locally finite closed covering $\mathcal{C}$ of X and use it to expand the elements of $\mathcal{B}$.

For each point x of X, there is a neighborhood of x that intersects only finitely many elements of $\mathcal{B}$. The collection of all open sets that intersect only finitely many elements of $\mathcal{B}$ is thus an open covering of X. Using (3) again, let $\mathcal{C}$ be a closed refinement of this covering that covers X and is locally finite. Each element of $\mathcal{C}$ intersects only finitely many elements of $\mathcal{B}$.

For each element B of $\mathcal{B}$, let

$$\mathcal{C}(B) = \{C \mid C \in \mathcal{C} \quad \text{and} \quad C \subset X - B\}.$$

Then define

$$E(B) = X - \bigcup_{C \in \mathcal{C}(B)} C.$$

Because $\mathcal{C}$ is a locally finite collection of closed sets, the union of the elements of any subcollection of $\mathcal{C}$ is closed, by Lemma 1.1. Therefore, the set $E(B)$ is an open set. Furthermore, $E(B) \supset B$ by definition. (See Figure 3, in which

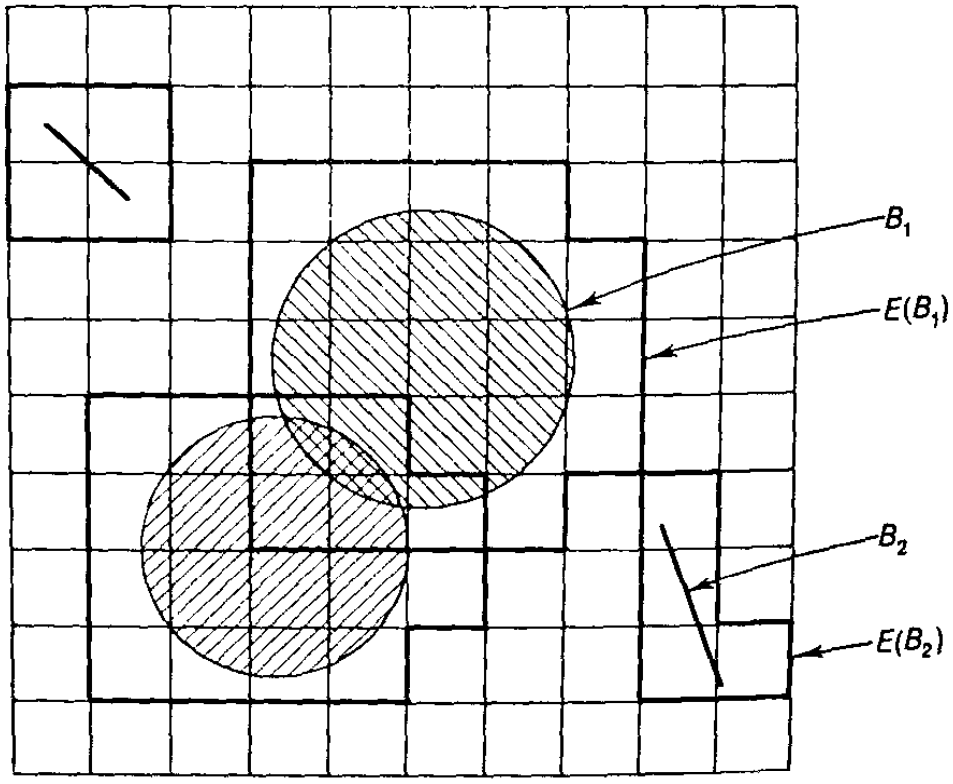


Figure 3

the elements of $\mathcal{B}$ are represented as closed circular regions and line segments, and the elements of $\mathcal{C}$ are represented as closed square regions.)

Now we may have expanded each B too much; the collection $\{E(B)\}$ may not be a refinement of $\mathcal{A}$. This is easily remedied. For each $B \in \mathcal{B}$, choose an

†See Lemma 3.1 and the proof (1) ⇒ (2) just given.

element $F(B)$ of $\mathcal{A}$ containing B. Then define

$$\mathcal{D} = \{E(B) \cap F(B) \mid B \in \mathcal{B}\}.$$

The collection $\mathcal{D}$ is a refinement of $\mathcal{A}$. Because $B \subset (E(B) \cap F(B))$ and $\mathcal{B}$ covers X, the collection $\mathcal{D}$ also covers X.

We have finally to prove that $\mathcal{D}$ is locally finite. Given a point x of X, choose a neighborhood W of x that intersects only finitely many elements of $\mathcal{C}$, say $C_1, \ldots, C_k$. Then $W \subset C_1 \cup \cdots \cup C_k$, because $\mathcal{C}$ is a covering of X. Now if an element C of $\mathcal{C}$ intersects the set $E(B) \cap F(B)$, it cannot lie in $X - B$ (by definition of $E(B)$); thus C must intersect B. Because the element C of $\mathcal{C}$ intersects only finitely many elements B of $\mathcal{B}$, it intersects at most the same number of elements of the collection $\mathcal{D} = \{E(B) \cap F(B)\}$. Then since each C_i intersects only finitely many elements of $\mathcal{D}$, so does the set W. □

Exercises

1. Give an example to show that if X is paracompact, it does not follow that for every open covering $\mathcal{A}$ of X, there is a locally finite *subcollection* of $\mathcal{A}$ that covers X.

2. (a) Show that every regular Lindelöf space is paracompact.
   (b) Conclude that R_l is paracompact.

3. (a) Show that the product of a paracompact space and a compact Hausdorff space is paracompact. [*Hint:* Use the tube lemma of §3-5.]
   (b) Conclude that S_Ω is not paracompact.
   (c) Show that $S_\Omega \times [0, 1)$ in the dictionary order is not paracompact. [*Hint:* S_Ω is homeomorphic to the closed subspace $S_\Omega \times 0$.]

4. Let X be a Hausdorff space. Suppose there exists a countable open covering $\{U_n\}$ of X such that for each n, $\bar{U}_n$ is compact and $\bar{U}_n \subset U_{n+1}$. Show that X is paracompact. [*Hint:* See Example 1.]

5. Let X be a regular space. Show that if there is a countable covering $\{U_n\}$ of X by open sets whose closures are paracompact, then X is paracompact.

6. Is every locally compact Hausdorff space paracompact?

7. One has a "shrinking lemma" for point-finite open coverings of a normal space; see the exercises of §4-5. Here is an "expansion lemma" for locally finite indexed families in a paracompact space:

   Lemma. Let $\{A_\alpha\}_{\alpha \in J}$ be a locally finite indexed family of subsets of the paracompact space X. Then there is a locally finite indexed family $\{U_\alpha\}_{\alpha \in J}$ of open sets such that $A_\alpha \subset U_\alpha$ for each $\alpha \in J$.

8. Determine which of the following spaces are paracompact in the dictionary order topology:
   (a) $[0, 1] \times [0, 1]$
   (b) $[0, 1) \times [0, 1]$
   *(c) $[0, 1] \times [0, 1)$

***9.** Let G be a locally compact, connected, Hausdorff topological group. Show that G is paracompact. [*Hint:* Choose a neighborhood U_1 of e having compact closure. In general, define $U_{n+1} = \bar{U}_n \cdot U_1$. Then $G = \bigcup \bar{U}_n$.]

***10.** Show that if G is a paracompact topological group and H is a compact subgroup, then G/H is paracompact. [*Hint:* Show that $p: G \longrightarrow G/H$ is a closed map. Then show that if $\mathcal{B}$ is a locally finite closed covering of G, the collection $\{p(B) \mid B \in \mathcal{B}\}$ is a locally finite closed covering of G/H.]

6-5 *The Smirnov Metrization Theorem*

The Nagata–Smirnov metrization theorem gives one set of necessary and sufficient conditions for metrizability of a space. In this section we give another such metrization theorem. It is a corollary of the Nagata–Smirnov theorem and was first proved by Smirnov.

Definition. A space X is **locally metrizable** if every point x of X has a neighborhood U that is metrizable in the subspace topology.

Theorem 5.1 (Smirnov metrization theorem). A space X *is metrizable if and only if it is paracompact and locally metrizable.*

Proof. Suppose that X is metrizable. Then X is locally metrizable; it is also paracompact, by Stone's theorem.

Conversely, suppose that X is paracompact and locally metrizable. We shall show that X has a basis $\mathcal{B}$ that is countably locally finite. Since X is regular (being paracompact), it will then follow from the Nagata–Smirnov theorem that X is metrizable.

The proof is an adaptation of the proof of Theorem 3.2. Cover X by open sets that are metrizable; then choose a locally finite open refinement $\mathcal{C}$ of this covering that covers X. Each element C of $\mathcal{C}$ is metrizable; let the function $d_C : C \times C \longrightarrow R$ be a metric that gives the topology of C. Given $x \in C$, let $B_C(x, \epsilon)$ denote the set of all points y of C such that $d_C(x, y) < \epsilon$. Being open in C, the set $B_C(x, \epsilon)$ will also be open in X.

Given $m \in Z_+$, let $\mathcal{A}^m$ be the covering of X by all open balls of radius $1/m$ in sets C of $\mathcal{C}$;

$$\mathcal{A}^m = \{B_C(x, 1/m) \mid C \in \mathcal{C} \quad \text{and} \quad x \in C\}.$$

Let $\mathcal{D}^m$ be a locally finite open refinement of $\mathcal{A}^m$ that covers X. (Here we use paracompactness.) Let

$$\mathcal{D} = \bigcup \mathcal{D}^m;$$

then $\mathcal{D}$ is countably locally finite. We assert that $\mathcal{D}$ is a basis for X; our theorem follows.

Let x be a point of X and let U be a neighborhood of x. We seek to find an element D of $\mathcal{D}$ such that $x \in D \subset U$. Now x belongs to only finitely many elements of $\mathcal{C}$, say to $C_1, \ldots, C_k$. Then $U \cap C_i$ is a neighborhood of x in the set C_i, so there is an $\epsilon_i > 0$ such that

$$B_{C_i}(x, \epsilon_i) \subset (U \cap C_i).$$

Choose m so that $1/m < \frac{1}{2} \min \{\epsilon_1, \ldots, \epsilon_k\}$. Because the collection $\mathcal{D}^m$ covers X, there must be an element D of $\mathcal{D}^m$ containing x. Because $\mathcal{D}^m$ refines $\mathcal{A}^m$, there must be an element $B_C(y, 1/m)$ of $\mathcal{A}^m$, for some $C \in \mathcal{C}$ and some $y \in C$, that contains D. Then x belongs to C, so that C must be one of the sets $C_1, \ldots, C_k$. Say that $C = C_i$. Then, using the triangle inequality, we have

$$D \subset B_{C_i}(y, 1/m) \subset B_{C_i}(x, \epsilon_i) \subset U,$$

as desired. □

EXAMPLE 1. We show that the space S_Ω is not paracompact. One proof of this fact appeared in the preceding set of exercises; here is another.

Note first that S_Ω is locally metrizable: Let $x \in S_\Omega$; choose y so that $x < y < \Omega$. The open set S_y has a countable basis; indeed, the collection of *all* intervals with end points in S_y is countable. It follows from the Urysohn metrization theorem that S_y is metrizable.

If S_Ω were paracompact, it would follow from the Smirnov theorem that S_Ω is metrizable. But we know that S_Ω is not metrizable, for it is limit point compact but not compact (see Example 1 of §3-7).

Exercises

1. Compare Theorem 5.1 with Exercises 5 and 6 of §4-4.
2. Show that every connected locally compact metric space has a countable basis. [Hint: Let $\mathcal{C}$ be a locally finite covering of X by open sets having compact closures. Let U_1 be a nonempty element of $\mathcal{C}$, and in general let U_{n+1} be the union of all elements of $\mathcal{C}$ that intersect $\bar{U}_n$. Show $\bar{U}_n$ is compact and $X = \bigcup U_n$.]
3. A more general notion of manifold than that defined in §4-5 is the following: A space X is called a **paracompact-manifold** if it is a paracompact space and there is an integer m such that each point of X has a neighborhood that is homeomorphic to an open subset of R^m.
   (a) Show that a paracompact-manifold is metrizable.
   (b) Show that if X is a paracompact-manifold having only countably many components, then X is a manifold; and conversely.

7. Complete Metric Spaces and Function Spaces

The concept of completeness for a metric space is one you may have seen already. It is basic for all aspects of analysis. Although completeness is a metric property rather than a topological one, there are a number of theorems involving complete metric spaces that are topological in character. In this chapter, we shall study the most important examples of complete metric spaces and shall prove some of these theorems.

The most familiar example of a complete metric space is euclidean space in either of its usual metrics. Another example, just as important, is the set $\mathcal{C}(X, Y)$ of all continuous functions mapping a space X into a space Y. This set has a metric called the *uniform metric*, analogous to the uniform metric defined for R^J in §2-9. If Y is a complete metric space, then $\mathcal{C}(X, Y)$ is complete in the uniform metric. This we demonstrate in §7-1. As an application, we construct in §7-2 the well-known *Peano space-filling curve.*

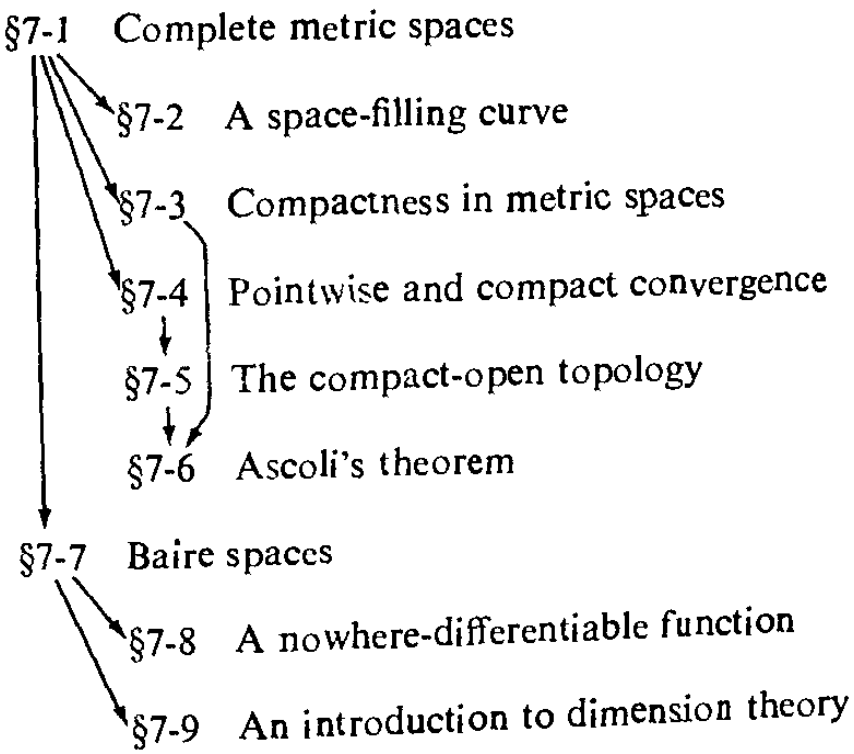
One theorem of topological character concerning complete metric spaces is a theorem relating compactness of a space to completeness. We prove it in §7-3. An immediate corollary is a theorem concerning compact sets in the function space $\mathcal{C}(X, R^n)$; it is the classical version of a famous theorem called *Ascoli's theorem.*

There are other useful topologies on the function space $\mathcal{C}(X, Y)$ besides the one derived from the uniform metric. We study some of them in §7-4 and §7-5. We also prove a general version of Ascoli's theorem, in §7-6.

Another theorem of topological character concerning complete metric spaces is a theorem we prove in §7-7, to the effect that every complete metric space belongs to the class of topological spaces called the *Baire spaces*. The defining condition for a Baire space is a bit complicated to state, but it is often useful in the applications. One application is the proof we give in §7-8 of the existence of a continuous nowhere-differentiable real-valued function.

Another application arises in that branch of topology called *dimension theory*. In §7-9 we define a topological notion of dimension (due to Lebesgue), and prove the classic theorem that every compact metrizable space of topological dimension m can be imbedded in euclidean space R^N of dimension $N = 2m + 1$. This theorem generalizes the imbedding theorem for manifolds proved in §4-5.

Some of the sections of this chapter are independent of one another. The dependence among them is expressed in the following diagram:



Throughout the chapter, we assume §3-8, Local compactness. When we study dimension theory, we shall make use of §4-5, Partitions of unity, as well as a bit of linear algebra.

7-1 Complete Metric Spaces

In this section we define the notion of completeness and show that if Y is a complete metric space, then the function space $\mathcal{C}(X, Y)$ is complete in the uniform metric. We also show that every metric space can be imbedded isometrically in a complete metric space.

Definition. Let (X, d) be a metric space. A sequence (x_n) of points of X is said to be a **Cauchy sequence** in (X, d) if it has the property that given $\epsilon > 0$, there is an integer N such that

$$d(x_n, x_m) < \epsilon \quad \text{whenever } n, m \geq N.$$

The metric space (X, d) is said to be **complete** if every Cauchy sequence in X converges.

Any convergent sequence in X is necessarily a Cauchy sequence, of course; completeness requires that the converse hold.

Note that a closed subset A of a complete metric space (X, d) is necessarily complete in the restricted metric. For a Cauchy sequence in A is also a Cauchy sequence in X, whence it converges in X. Because A is a closed subset of X, the limit must lie in A.

Note also that if X is complete under the metric d, then X is complete under the standard bounded metric

$$\bar{d}(x, y) = \min \{d(x, y), 1\}$$

corresponding to d, and conversely. For a sequence (x_n) is a Cauchy sequence under $\bar{d}$ if and only if it is a Cauchy sequence under d. And a sequence converges under $\bar{d}$ if and only if it converges under d.

A useful criterion for a metric space to be complete is the following:

Lemma 1.1. A metric space X *is complete if every Cauchy sequence in* X *has a convergent subsequence.*

Proof. Let (x_n) be a Cauchy sequence in (X, d). We show that if (x_n) has a subsequence (x_{n_i}) that converges to a point x, then the sequence (x_n) itself converges to x.

Given $\epsilon > 0$, first choose N large enough that

$$d(x_n, x_m) < \epsilon/2$$

for all $n, m \geq N$ [using the fact that (x_n) is a Cauchy sequence]. Then choose an integer i large enough that $n_i \geq N$ and

$$d(x_{n_i}, x) < \epsilon/2$$

[using the fact that $n_1 < n_2 < \ldots$ is an increasing sequence of integers and x_{n_i} converges to x]. Putting these facts together, we have the desired result that for $n \geq N$,

$$d(x_n, x) \leq d(x_n, x_{n_i}) + d(x_{n_i}, x) < \epsilon. \quad \square$$

Theorem 1.2. *Euclidean space* R^k *is complete in either of its usual metrics, the euclidean metric* d *or the square metric* ρ.

Proof. To show the metric space (R^k, ρ) is complete, let (x_n) be a Cauchy sequence in (R^k, ρ). Then the *set* $\{x_n\}$ is a bounded subset of (R^k, ρ). For if

we choose N so that

$$\rho(x_n, x_m) \leq 1$$

for all $n, m \geq N$, then the number

$$M = \max \{\rho(x_1, 0), \ldots, \rho(x_{N-1}, 0), \rho(x_N, 0) + 1\}$$

is an upper bound for $\rho(x_n, 0)$. Thus the points of the sequence (x_n) all lie in the cube $[-M, M]^k$. Since this cube is compact, the sequence (x_n) has a convergent subsequence, by Theorem 7.4 of Chapter 3. Then (R^k, ρ) is complete, by Lemma 1.1.

To show that (R^k, d) is complete, note that a sequence is a Cauchy sequence relative to d if and only if it is a Cauchy sequence relative to ρ, and a sequence converges relative to d if and only if it converges relative to ρ. □

EXAMPLE 1. An example of a noncomplete metric space is the space Q of rational numbers in the usual metric $d(x, y) = |x - y|$. For instance, the sequence

$$1.4, 1.41, 1.414, 1.4142, 1.41421, \ldots$$

of finite decimals converging (in R) to $\sqrt{2}$ is a Cauchy sequence in Q that does not converge (in Q).

EXAMPLE 2. Another noncomplete space is the open interval $(-1, 1)$ in R, in the metric $d(x, y) = |x - y|$. In this space the sequence (x_n) defined by

$$x_n = 1 - 1/n$$

is a Cauchy sequence that does not converge. This example shows that completeness is not a topological property, that is, it is not preserved by homeomorphisms. For $(-1, 1)$ is homeomorphic to the real line R, and R is complete in its usual metric.

EXAMPLE 3. There is a metric for the product topology on R^ω relative to which R^ω is complete. Consider the metric

$$D(\mathbf{x}, \mathbf{y}) = \text{lub } \{\bar{d}(x_i, y_i)/i\}$$

on R^ω, where $\bar{d}(a, b) = \min \{|a - b|, 1\}$. We showed in Theorem 9.5 of Chapter 2 that D is a metric for the product space R^ω. To show R^ω complete under D, let $(\mathbf{x}_n)$ be a Cauchy sequence under D. Let i be fixed for the moment. Let $\pi_i : R^\omega \longrightarrow R$ be projection on the ith coordinate. Since

$$\bar{d}(\pi_i(\mathbf{x}), \pi_i(\mathbf{y})) \leq iD(\mathbf{x}, \mathbf{y}),$$

the sequence $(\pi_i(\mathbf{x}_n))$ is a Cauchy sequence in $(R, \bar{d})$. Therefore, it converges, say to a_i.

Let $\mathbf{a}$ be the point $(a_i)_{i \in Z_+}$ of R^ω. We assert that $\mathbf{x}_n \longrightarrow \mathbf{a}$. For a typical basis element containing $\mathbf{a}$ is of the form $\prod U_i$, where $U_i = R$ for $i \geq m$. Choose N_i large enough that $\pi_i(\mathbf{x}_n) \in U_i$ for $n \geq N_i$. Let $N = \max \{N_1, \ldots, N_m\}$; then $\mathbf{x}_n \in \prod U_i$ for $n \geq N$.

Although both the product spaces R^n and R^ω have metrics relative to which they are complete, one cannot hope to prove the same result for the product space R^J in general, because R^J is not even metrizable if J is uncountable (see §2-10). There is, however, a different topology on the set R^J, the one given by the uniform metric. Relative to this metric, R^J *is* complete, as we shall see.

We define the uniform metric for Y^J in general as follows:

Definition. Let (Y, d) be a metric space. Let $\bar{d}(a, b) = \min\{d(a, b), 1\}$ be the standard bounded metric on Y corresponding to d. Given an index set J, define a metric on the set Y^J of all functions $f: J \to Y$ by the rule

$$\bar{\rho}(f, g) = \text{lub}\,\{\bar{d}(f(\alpha), g(\alpha)) |\ \alpha \in J\}.$$

The function $\bar{\rho}$ is called the **uniform metric** on Y^J corresponding to the metric d on Y; it is easy to check that it is a metric.

Note that here we have used functional notation for the elements of Y^J rather than "tuple" notation. We shall follow this practice thoughout the chapter.

Theorem 1.3. *If the space* Y *is complete in the metric* d, *then the space* Y^I *is complete in the uniform metric* $\bar{\rho}$ *corresponding to* d.

Proof. Recall that if (Y, d) is complete, so is $(Y, \bar{d})$, where $\bar{d}$ is the bounded metric corresponding to d. Now suppose that $f_1, f_2, \ldots$ is a sequence of points of Y^J that is a Cauchy sequence relative to $\bar{\rho}$. Given α in J, the fact that

$$\bar{d}(f_n(\alpha),\ f_m(\alpha)) \le \bar{\rho}(f_n, f_m)$$

for all n, m means that the sequence $f_1(\alpha), f_2(\alpha), \ldots$ is a Cauchy sequence in $(Y, \bar{d})$. Hence this sequence converges, say to a point y_α. Let $f: J \to Y$ be the function defined by $f(\alpha) = y_\alpha$. We assert that the sequence (f_n) converges to f in the metric $\bar{\rho}$.

Given $\epsilon > 0$, first choose N large enough that $\bar{\rho}(f_n, f_m) < \epsilon/2$ whenever $n, m \ge N$. Then, in particular,

$$\bar{d}(f_n(\alpha),\ f_m(\alpha)) < \epsilon/2$$

for $n, m \ge N$ and $\alpha \in J$. Letting n and α be fixed and m become arbitrarily large, we see that

$$\bar{d}(f_n(\alpha), f(\alpha)) \le \epsilon/2.$$

This inequality holds for any α in J, provided merely that $n \ge N$. Therefore,

$$\bar{\rho}(f_n, f) \le \epsilon/2 < \epsilon$$

for $n \ge N$, as desired. □

Now let us specialize somewhat, and consider the set Y^X where X is a *topological space* rather than merely a set. Of course, this has no effect on

what has gone before; the topology of X is irrelevant when considering the set of *all* functions $f: X \longrightarrow Y$. But suppose that we consider the subset $\mathcal{C}(X, Y)$ of Y^X consisting of all *continuous* functions $f: X \longrightarrow Y$. It turns out that if Y is complete, this subset is also complete in the uniform metric.

Theorem 1.4. *Let* X *be a topological space and let* (Y, d) *be a metric space. The set* $\mathcal{C}$(X, Y) *of continuous functions is closed in* Y^X *under the uniform metric. Therefore, if* Y *is complete,* $\mathcal{C}$(X, Y) *is complete in the uniform metric.*

Proof. This theorem is just the uniform limit theorem (Theorem 10.6 of Chapter 2) in a new guise. First, we show that if a sequence of elements f_n of Y^X converges to the element f of Y^X relative to the metric $\bar{\rho}$ on Y^X, then it converges to f uniformly in the sense defined in §2-10, relative to the metric $\bar{d}$ on Y. Given $\epsilon > 0$, choose an integer N such that

$$\bar{\rho}(f, f_n) < \epsilon$$

for all $n \geq N$. Then for all $x \in X$ and all $n \geq N$,

$$\bar{d}(f_n(x), f(x)) \leq \bar{\rho}(f_n, f) < \epsilon.$$

Thus (f_n) converges uniformly to f.

Now we show that $\mathcal{C}(X, Y)$ is closed in Y^X relative to the metric $\bar{\rho}$. Let f be an element of Y^X that is a limit point of $\mathcal{C}(X, Y)$. Then there is a sequence (f_n) of elements of $\mathcal{C}(X, Y)$ converging to f in the metric $\bar{\rho}$. By the uniform limit theorem, f is continuous, so that $f \in \mathcal{C}(X, Y)$.

Since Y^X is complete under $\bar{\rho}$, so is the closed subset $\mathcal{C}(X, Y)$. □

Definition. Let (Y, d) be a metric space; let X be either a set or a topological space. Suppose that $\mathcal{F}$ is a subset of Y^X having the property that for each pair f, g of elements of $\mathcal{F}$, the set

$$\{d(f(x), g(x)) \mid x \in X\}$$

is bounded. We define a metric ρ on $\mathcal{F}$ by the formula

$$\rho(f, g) = \text{lub}\,\{d(f(x), g(x)) \mid x \in X\};$$

it is called the square metric or, more commonly, the sup (supremum) metric.

How does this metric compare with the uniform metric $\bar{\rho}$ defined above? It is an easy exercise to show that for $f, g \in \mathcal{F}$,

$$\bar{\rho}(f, g) = \min\{\rho(f, g), 1\}.$$

This means that on a set $\mathcal{F}$ where both $\bar{\rho}$ and ρ are defined, $\bar{\rho}$ is just the standard bounded metric derived from ρ. (This is the reason we originally introduced the notation $\bar{\rho}$ for the uniform metric.) It follows that ρ and $\bar{\rho}$ give the same topology on $\mathcal{F}$. And it follows that $\mathcal{F}$ is complete under ρ if and only if it is complete under $\bar{\rho}$. Therefore, for all practical purposes, they are equivalent metrics.

Many authors do not bother to define the metric $\bar{\rho}$ at all, but simply

limit themselves to spaces $\mathcal{F}$ where ρ is defined. We prefer to deal with the more general case and specialize to the metric ρ whenever it is useful (and possible) to do so.

EXAMPLE 4. Let (Y, d) be a metric space and let J be a finite index set. Then the square metric ρ is defined on all of Y^J. Of course, in this case all the various topologies for Y^J—the box topology, the uniform topology, and the product topology—are the same.

EXAMPLE 5. Consider the space $\mathcal{C}(X, R)$ of all continuous real-valued functions on the compact space X. Each such function is bounded, so the metric

$$\rho(f, g) = \text{lub}\,\{|f(x) - g(x)|\}$$

is defined on this space. Indeed, the maximum value theorem tells us that

$$\rho(f, g) = \max\,\{|f(x) - g(x)|\}.$$

The space $\mathcal{C}(X, R)$ is complete under the metric ρ.

We now prove a classical theorem, to the effect that every metric space can be imbedded isometrically in a complete metric space. (A different proof, somewhat more direct, is outlined in Exercise 8.) Although we shall not need this theorem, it is useful in other parts of mathematics.

First, a lemma:

Lemma 1.5. *Let* X *be a topological space. The set* $\mathcal{B}(X, R)$ *of all bounded functions* $f : X \to R$ *is complete under the sup metric* ρ.

Proof. We show that $\mathcal{B}(X, R)$ is closed in R^X in the uniform metric $\bar{\rho}$. It follows that $\mathcal{B}(X, R)$ is complete under $\bar{\rho}$. Since the metric ρ is defined on this set, it is also complete under ρ.

Let f be the limit of a sequence (f_n) of elements of $\mathcal{B}(X, R)$. Choose N so that $\bar{\rho}(f_n, f) < 1$ for $n \geq N$. Then since f_N belongs to $\mathcal{B}(X, R)$, there is an M such that $|f_N(x)| \leq M$ for all $x \in X$. Then

$$|f(x)| \leq M + 1$$

for all $x \in X$, so that f is bounded. □

Theorem 1.6. *Let* (X, d) *be a metric space. There is an isometric imbedding of* X *into a complete metric space.*

Proof. Let $\mathcal{B}(X, R)$ be the set of all bounded functions mapping X into R. Let x_0 be a fixed point of X. Given $a \in X$, define $\phi_a : X \to R$ by the equation

$$\phi_a(x) = d(x, a) - d(x, x_0).$$

We assert that ϕ_a is bounded. For it follows from the inequalities

$$d(x, a) \leq d(x, x_0) + d(a, x_0),$$
$$d(x, x_0) \leq d(x, a) + d(a, x_0)$$

that we have $|\phi_a(x)| \leq d(a, x_0)$.

Define $\Phi : X \to \mathcal{B}(X, R)$ by setting

$$\Phi(a) = \phi_a.$$

We show that Φ is an isometric imbedding of (X, d) into the complete metric space $(\mathcal{B}(X, R), \rho)$. That is, we show that for every pair of points $a, b \in X$, we have

$$\rho(\phi_a, \phi_b) = d(a, b).$$

By definition,

$$\rho(\phi_a, \phi_b) = \text{lub}\{|\phi_a(x) - \phi_b(x)| ; x \in X\}$$
$$= \text{lub}\{|d(x, a) - d(x, b)| ; x \in X\}.$$

By the triangle inequality,

$$|d(x, a) - d(x, b)| \leq d(a, b),$$

from which we conclude that

$$\rho(\phi_a, \phi_b) \leq d(a, b).$$

On the other hand, this inequality cannot be strict, for when $x = a$,

$$|d(x, a) - d(x, b)| = d(a, b). \quad \square$$

Definition. Let X be a metric space. If $h : X \to Y$ is an isometric imbedding of X into a complete metric space Y, then the subspace $\overline{h(X)}$ of Y is a complete metric space. It is called the completion of X.

The completion of X is uniquely determined up to an isometry. See Exercise 9.

Exercises

1. Let X be a metric space.
   (a) Suppose that for some $\epsilon > 0$, every ϵ-ball in X has compact closure. Show that X is complete.
   (b) Suppose that for each $x \in X$ there is an $\epsilon > 0$ such that the ball $B(x, \epsilon)$ has compact closure. Show by means of an example that X need not be complete.
2. Let (X, d_X) and (Y, d_Y) be metric spaces; let Y be complete. Let $A \subset X$. Show that if $f: A \to Y$ is uniformly continuous, then f can be uniquely extended to a continuous function $g : \bar{A} \to Y$, and g is uniformly continuous.
3. Consider the subset R^∞ of R^ω consisting of sequences that are eventually zero. Is R^∞ complete in the uniform metric?

4. Show that the metric space (X, d) is complete if and only if for any nested sequence $A_1 \supset A_2 \supset \cdots$ of nonempty closed sets of X such that diameter $A_n \to 0$,

$$\bigcap_{n \in Z_+} A_n \neq \emptyset.$$

5. Let (X, d) be a complete metric space. Recall that a continuous map $f: X \to X$ is said to be a *contraction* if there is a number $\alpha < 1$ such that

$$d(f(x), f(y)) \leq \alpha d(x, y)$$

for all $x, y \in X$. Show that if f is a contraction, then there is a unique point x of X such that $f(x) = x$.

6. A space X is said to be **topologically complete** if there exists a metric for the topology of X relative to which X is complete.
   (a) Show that a closed subspace of a topologically complete space is topologically complete.
   (b) Show that a countable product of topologically complete spaces is topologically complete (in the product topology).
   (c) Show that an open subset of a topologically complete space is topologically complete. [*Hint:* If $U \subset X$ and X is complete under the metric d, define $\phi : U \to R$ by the equation

$$\phi(x) = 1/d(x, X - U),$$

   where $d(x, A) = \text{glb}\,\{d(x, a) \mid a \in A\}$. Then define $f: U \to X \times R$ by the equation

$$f(x) = x \times \phi(x).]$$

   (d) Show that if A is a G_δ set in a topologically complete space, then A is topologically complete. [*Hint:* Let A be the intersection of the open sets U_n, for $n \in Z_+$. Consider the diagonal imbedding $f(a) = (a, a, \ldots)$ of A into $\prod U_n$.]
   (e) Show that the irrationals are topologically complete.

7. Show that ℓ^2 is complete in the ℓ^2-metric (see Exercise 9 of §2-9).

8. Let (X, d) be a metric space. Show that there is an isometric imbedding h of X into a complete metric space (Y, D), as follows: Let $\tilde{X}$ denote the set of all Cauchy sequences

$$\mathbf{x} = (x_1, x_2, \ldots)$$

of points of X. Define $\mathbf{x} \sim \mathbf{y}$ if

$$d(x_n, y_n) \to 0.$$

Let $[\mathbf{x}]$ denote the equivalence class of $\mathbf{x}$; and let Y denote the set of these equivalence classes. Define a metric D on Y by the equation

$$D([\mathbf{x}], [\mathbf{y}]) = \lim_{n \to \infty} d(x_n, y_n).$$

   (a) Show that $\sim$ is an equivalence relation, and show that D is a well-defined metric.
   (b) Define $h : X \to Y$ by letting $h(x)$ be the equivalence class of the constant sequence $(x, x, \ldots)$;

$$h(x) = [(x, x, \ldots)].$$

Show that h is an isometric imbedding.

(c) Show that $h(X)$ is dense in Y; indeed, given $\mathbf{x} = (x_1, x_2, \ldots) \in \tilde{X}$, show the sequence $h(x_n)$ of points of Y converges to the point $[\mathbf{x}]$ of Y.

(d) Show that if A is a dense subset of a metric space (Z, ρ), and if every Cauchy sequence in A converges in Z, then Z is complete.

(e) Show that (Y, D) is complete.

9. *Theorem (Uniqueness of the completion). Let $h : X \to Y$ and $h' : X \to Y'$ be isometric imbeddings of the metric space (X, d) in the complete metric spaces (Y, D) and (Y', D'), respectively. Then there is an isometry of $(\overline{h(X)}, D)$ with $(\overline{h'(X)}, D')$ that equals $h'h^{-1}$ on the subspace $h(X)$.*

7-2 A Space-Filling Curve

As an application of the completeness of the metric space $\mathcal{C}(X, Y)$ in the uniform metric when Y is complete, we shall construct the famous "Peano space-filling curve."

Theorem 2.1. *Let $\mathrm{I} = [0, 1]$. There exists a continuous map $\mathrm{f} : \mathrm{I} \to \mathrm{I}^2$ whose image fills up the entire square I^2.*

The existence of this path violates one's naive geometric intuition in much the same way as does the existence of the continuous nowhere-differentiable function (which we shall come to later).

Proof. Step 1. We shall construct the map f as the limit of a sequence of continuous functions f_n. First we describe a particular operation on paths which will be used to generate the sequence f_n.

Begin with an arbitrary closed interval $[a, b]$ in the real line and an arbitrary square in the plane with sides parallel to the coordinate axes, and consider the triangular path g pictured in Figure 1. It is a continuous map of $[a, b]$ into the square. The operation we wish to describe replaces the path g by the path g' pictured in Figure 2. It is made up of four triangular paths, each half the size of g. Note that g and g' have the same initial point and the same final point. You can write the equations for g and g' if you like.

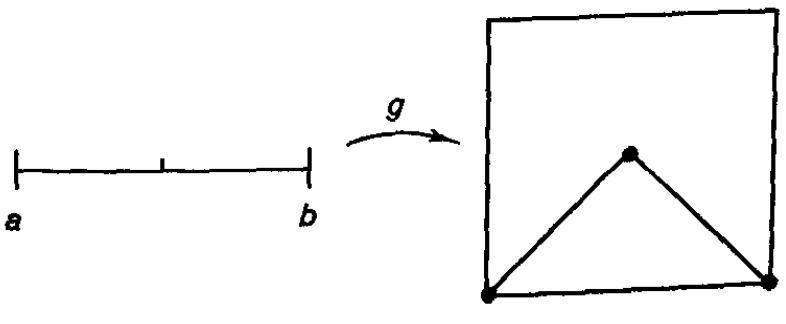


Figure 1

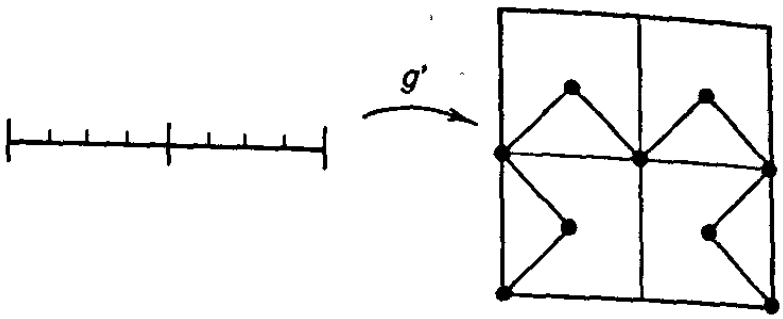


Figure 2

This same operation can also be applied to any triangular path connecting two adjacent corners of the square. For instance, when applied to the path h pictured in Figure 3, it gives the path h'.

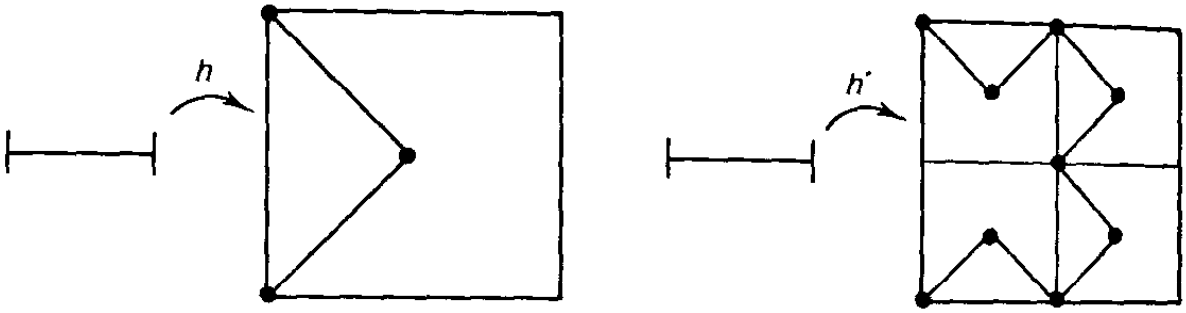


Figure 3

Step 2. Now we define a sequence of functions $f_n : I \to I^2$. The first function, which we label f_0 for convenience, is the triangular path pictured in Figure 1, letting $a = 0$ and $b = 1$. The next function f_1 is the function obtained by applying the operation described in Step 1 to the function f_0; it is pictured in Figure 2. The next function f_2 is the function obtained by applying this same operation to each of the four triangular paths that make up f_1. It is pictured in Figure 4. The next function f_3 is obtained by applying the operation to each of the 16 triangular paths that make up f_2; it is pictured

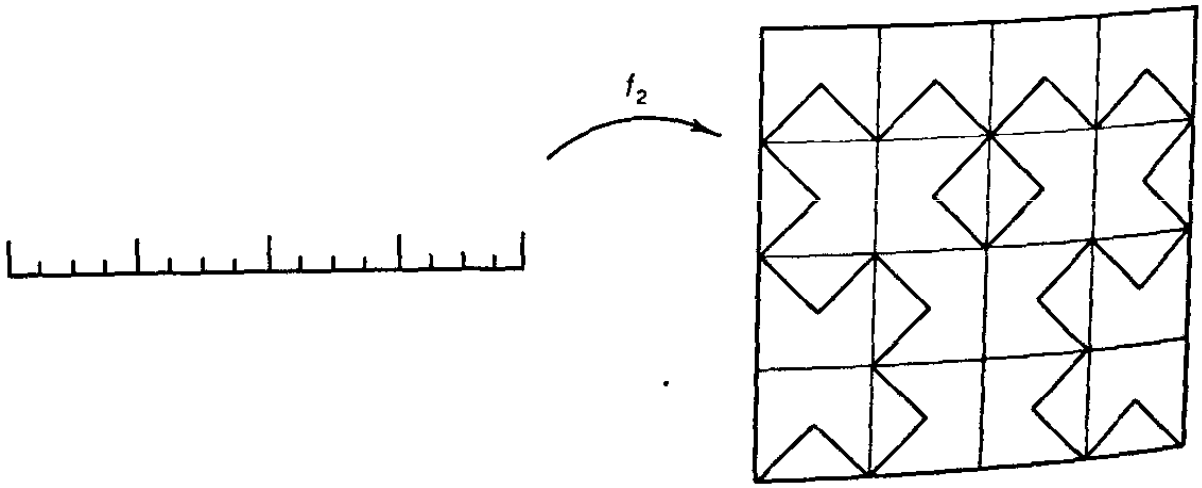


Figure 4

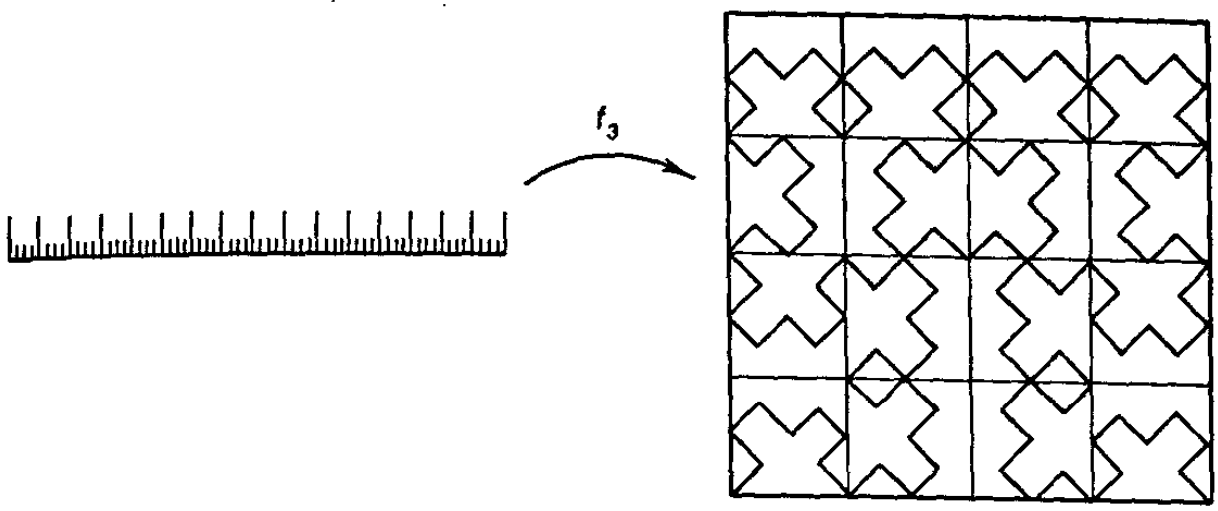


Figure 5

in Figure 5. And so on. At the general step, f_n is a path made up of 4^n triangular paths of the type considered in Step 1, each lying in a square of edge length $1/2^n$. The function f_{n+1} is obtained by applying the operation of Step 1 to these triangular paths, replacing each one by four smaller triangular paths.

Step 3. For purposes of this proof, let $d(\mathbf{x}, \mathbf{y})$ denote the square metric on R^2,

$$d(\mathbf{x}, \mathbf{y}) = \max \{|x_1 - y_1|, |x_2 - y_2|\}.$$

Then we can let ρ denote the corresponding sup metric on $\mathcal{C}(I, I^2)$;

$$\rho(f, g) = \operatorname{lub} \{d(f(t), g(t)) \mid t \in I\}$$

Since I^2 is closed in R^2, it is complete in the square metric; then $\mathcal{C}(I, I^2)$ is complete in the metric ρ.

We assert that the sequence of functions (f_n) defined in Step 2 is a Cauchy sequence under ρ. To prove this fact, let us examine what happens when we pass from f_n to f_{n+1}. Each small triangular path in f_n lies in a square of edge length $1/2^n$. The operation by which we obtain f_{n+1} replaces this triangular path by four triangular paths that lie in the same square. Therefore, in the square metric on I^2, the distance between $f_n(t)$ and $f_{n+1}(t)$ is at most $1/2^n$. As a result, $\rho(f_n, f_{n+1}) \leq 1/2^n$. It follows that (f_n) is a Cauchy sequence, since

$$\rho(f_n, f_{n+m}) \leq 1/2^n + 1/2^{n+1} + \cdots + 1/2^{n+m-1} < 2/2^n$$

for all n and m.

Step 4. Because $\mathcal{C}(I, I^2)$ is complete, the sequence f_n converges to a continuous function $f : I \longrightarrow I^2$. We prove that f is surjective.

Let $\mathbf{x}$ be a point of I^2; we show that $\mathbf{x}$ belongs to $f(I)$. First we note that, given n, the path f_n comes within a distance of $1/2^n$ of the point $\mathbf{x}$. For the

path f_n touches each of the little squares of edge length $1/2^n$ into which we have divided I^2.

Using this fact, we shall prove that, given $\epsilon > 0$, the ϵ-neighborhood of $\mathbf{x}$ intersects $f(I)$. Choose N large enough that

$$\rho(f_N, f) < \epsilon/2 \quad \text{and} \quad 1/2^N < \epsilon/2.$$

By the result of the preceding paragraph, there is a point $t_0 \in I$ such that $d(\mathbf{x}, f_N(t_0)) \leq 1/2^N$. Then since $d(f_N(t), f(t)) < \epsilon/2$ for all t, it follows that

$$d(\mathbf{x}, f(t_0)) < \epsilon,$$

so the ϵ-neighborhood of $\mathbf{x}$ intersects $f(I)$.

It follows that $\mathbf{x}$ belongs to the closure of $f(I)$. But I is compact, so $f(I)$ is compact and is therefore closed. Hence $\mathbf{x}$ belongs to $f(I)$, as desired. □

Exercises

1. Given n, show there is a continuous surjective map $g : I \longrightarrow I^n$. [*Hint:* Consider $f \times f : I \times I \longrightarrow I^2 \times I^2$.]
2. Is there is continuous surjective map $f : I \longrightarrow R^2$?
3. Show there is a continuous surjective map $f : R \longrightarrow R^n$.
4. Is there a continuous surjective map $f : R \longrightarrow X$ if X is the subspace R^∞ of R^ω consisting of sequences that are eventually zero? What if $X = R^\omega$?

*5. (a) Let X be a Hausdorff space. Show that if there is a continuous surjective mapping $f : I \longrightarrow X$, then X is compact, connected, locally connected, and metrizable.

(b) The converse also holds; it is a famous theorem of point set topology called the *Hahn–Mazurkiewicz theorem* (see [H-Y], p. 129). Assuming this theorem, show there is a continuous surjective map $f : I \longrightarrow I^\omega$.

A Hausdorff space that is the continuous image of the closed unit interval is often called a **Peano space.**

7-3 Compactness in Metric Spaces

We have already shown that compactness, limit point compactness, and sequential compactness are equivalent for metric spaces. There is still another formulation of compactness for metric spaces, one that involves the notion of completeness. We study it in this section. As an application, we shall prove a theorem characterizing those subsets of $\mathcal{C}(X, R^n)$ which are compact in the uniform metric; it is the classical version of *Ascoli's theorem.*

How is compactness of a metric space X related to completeness of X? It follows from Lemma 1.1 that every compact metric space is necessarily

complete. The converse does not hold—a complete metric space need not be compact. It is reasonable to ask what extra condition one needs to impose on a complete space to be assured of its compactness. Such a condition is the one called *total boundedness.*

Definition. A metric space (X, d) is said to be **totally bounded** if for every $\epsilon > 0$, there is a finite covering of X by ϵ-balls.

Total boundedness of a metric space clearly implies boundedness, but the converse does not hold.

EXAMPLE 1. Under the metric $d(x, y) = |x - y|$, the real line R is neither bounded nor totally bounded. Under the metric $\bar{d}(x, y) = \min\{|x - y|, 1\}$, the real line is bounded but not totally bounded.

EXAMPLE 2. Under the metric $|x - y|$, the subspace $(-1, 1)$ of R is totally bounded, and so is the subspace $Q \cap [-1, 1]$. Neither of these spaces is complete, however. The subspace $[-1, 1]$ is both complete and totally bounded.

Theorem 3.1. *A metric space* (X, d) *is compact if and only if it is complete and totally bounded.*

Proof. If X is a compact metric space, then X is automatically complete, as noted above. The fact that X is totally bounded is a consequence of the fact that the covering of X by all open ϵ-balls must contain a finite subcovering.

Conversely, let X be complete and totally bounded. We shall prove that X is sequentially compact. This will suffice.

Let (x_n) be a sequence of points of X. We shall construct a subsequence of (x_n) that is a Cauchy sequence, so that it necessarily converges. First cover X by finitely many balls of radius 1. At least one of these balls, say B_1, contains x_n for infinitely many values of n. Let J_1 be the subset of Z_+ consisting of those indices n for which $x_n \in B_1$.

Next, cover X by finitely many balls of radius $\frac{1}{2}$. Because J_1 is infinite, at least one of these balls, say B_2, must contain x_n for infinitely many values of n in J_1. Choose J_2 to be the set of those indices n for which $n \in J_1$ and $x_n \in B_2$. In general, given an infinite set J_k of positive integers, choose J_{k+1} to be an infinite subset of J_k such that there is a ball B_{k+1} of radius $1/(k+1)$ which contains x_n for all $n \in J_{k+1}$.

Choose $n_1 \in J_1$. Given n_k, choose $n_{k+1} \in J_{k+1}$ such that $n_{k+1} > n_k$; this we can do because J_{k+1} is an infinite set. Now for $i, j \geq k$, the indices n_i and n_j both belong to J_k (because $J_1 \supset J_2 \supset \cdots$ is a nested sequence of sets). Therefore, for all $i, j \geq k$, the points x_{n_i} and x_{n_j} are contained in a ball B_k of radius $1/k$. It follows that the sequence (x_{n_i}) is a Cauchy sequence, as desired. □

Now we apply this result to the function space $\mathcal{C}(X, R^n)$, in the uniform topology. We are going to be concerned only about the case where X is compact, so we shall use the metric

$$\rho(f, g) = \max \{d(f(x), g(x))\}$$

throughout. [Here d may denote either the euclidean metric or the square metric on R^n.]

We have proved earlier that a subset of (R^n, d) is compact if and only if it is closed and bounded. One might hope that the corresponding theorem is true for $(\mathcal{C}(X, R^n), \rho)$. But it is not. One needs to assume that the subset of $\mathcal{C}(X, R^n)$ satisfies an additional condition, called *equicontinuity*. It is defined as follows:

Definition. Let (Y, d) be a metric space. Let $\mathcal{F}$ be a subset of the function space $\mathcal{C}(X, Y)$. If $x_0 \in X$, the set $\mathcal{F}$ of functions is said to be **equicontinuous at x_0** if given $\epsilon > 0$, there is a neighborhood U of x_0 such that for all $x \in U$ and all $f \in \mathcal{F}$,

$$d(f(x), f(x_0)) < \epsilon.$$

If the set $\mathcal{F}$ is equicontinuous at x_0 for each $x_0 \in X$, it is said simply to be **equicontinuous.**

Continuity of the function f at x_0 means that given f and given $\epsilon > 0$, there exists a neighborhood U of x_0 such that $d(f(x), f(x_0)) < \epsilon$ for $x \in U$. Equicontinuity of $\mathcal{F}$ means that a single neighborhood U can be chosen that will work for all the functions f in the collection $\mathcal{F}$.

Equicontinuity has the following interesting interpretation when *both* X and Y are compact:

Lemma 3.2. *Let* X *be a compact space; let* (Y, d) *be a compact metric space. Let* $\mathcal{F}$ *be a subset of* $\mathcal{C}$(X, Y). *Then* $\mathcal{F}$ *is equicontinuous if and only if* $\mathcal{F}$ *is totally bounded in the sup metric* ρ.

Proof. Suppose that $\mathcal{F}$ is totally bounded under ρ. Given x_0, we show $\mathcal{F}$ is equicontinuous at x_0. Let $\epsilon > 0$ be given. Choose positive numbers ϵ_1 and ϵ_2 so that $2\epsilon_1 + \epsilon_2 \leq \epsilon$. Cover $\mathcal{F}$ by finitely many open ϵ_1-balls

$$B_\rho(f_1, \epsilon_1), \ldots, B_\rho(f_n, \epsilon_1).$$

Each function f_i is continuous; therefore, we can choose a neighborhood U of x_0 such that for $x \in U$ and $i = 1, \ldots, n$,

$$d(f_i(x), f_i(x_0)) < \epsilon_2.$$

We assert that if $x \in U$ and $f \in \mathcal{F}$, then $d(f(x), f(x_0)) < \epsilon$; this proves equicontinuity.

Let f be an arbitrary element of $\mathcal{F}$. Then f belongs to at least one of the above ϵ_1-balls, say to $B_\rho(f_i, \epsilon_1)$. Then

$$d(f(x), f_i(x)) < \epsilon_1,$$
$$d(f_i(x), f_i(x_0)) < \epsilon_2,$$
$$d(f_i(x_0), f(x_0)) < \epsilon_1.$$

The first and third inequalities hold because f is in $B_\rho(f_i, \epsilon_1)$, and the second holds because $x \in U$. It follows from the triangle inequality that for all $x \in U$, we have $d(f(x), f(x_0)) < \epsilon$, as desired.

Conversely, suppose that $\mathcal{F}$ is equicontinuous. Let $\epsilon > 0$ be given. We wish to cover $\mathcal{F}$ by finitely many open ϵ-balls. Choose ϵ_1 and ϵ_2 so that $2\epsilon_1 + \epsilon_2 \leq \epsilon$. Using equicontinuity of $\mathcal{F}$ and compactness of X, cover X by finitely many open sets $U_1, \ldots, U_k$, containing points $x_1, \ldots, x_k$, respectively, such that

$$d(f(x), f(x_i)) < \epsilon_1$$

for $x \in U_i$ and all $f \in \mathcal{F}$. Cover Y by finitely many open sets $V_1, \ldots, V_m$ of diameter less than ϵ_2.

Let J be the collection of all functions $\alpha : \{1, \ldots, k\} \to \{1, \ldots, m\}$. Given $\alpha \in J$, if there exists a function f of $\mathcal{F}$ such that $f(x_i) \in V_{\alpha(i)}$ for each $i = 1, \ldots, k$, choose one such function and label it f_α. The collection $\{f_\alpha\}$ is indexed by a subset J' of the set J and is thus finite. We assert that the open balls $B_\rho(f_\alpha, \epsilon)$, for $\alpha \in J'$, cover $\mathcal{F}$.

Let f be an element of $\mathcal{F}$. For each $i = 1, \ldots, k$, choose an integer $\alpha(i)$ such that $f(x_i) \in V_{\alpha(i)}$. Then the function α is in J'. We assert that f belongs to the ball $B_\rho(f_\alpha, \epsilon)$.

Let x be a point of X. Choose i so that $x \in U_i$. Then

$$d(f(x), f(x_i)) < \epsilon_1,$$
$$d(f(x_i), f_\alpha(x_i)) < \epsilon_2,$$
$$d(f_\alpha(x_i), f_\alpha(x)) < \epsilon_1.$$

The first and third inequalities hold because $x \in U_i$, and the second holds because $f(x_i)$ and $f_\alpha(x_i)$ are in $V_{\alpha(i)}$. We conclude that $d(f(x), f_\alpha(x)) < \epsilon$. Since this inequality holds for every $x \in X$,

$$\rho(f, f_\alpha) = \max \{d(f(x), f_\alpha(x))\} < \epsilon.$$

Thus f belongs to $B_\rho(f_\alpha, \epsilon)$, as asserted. □

Now we prove the classical version of Ascoli's theorem. A more general version, whose proof does not depend on this one, appears in §7-6.

Theorem 3.3 (Ascoli's theorem, classical version). *Let* X *be a compact space; consider* $\mathcal{C}(X, \mathbb{R}^n)$ *in the sup metric* ρ. *A subset* $\mathcal{F}$ *of* $\mathcal{C}(X, \mathbb{R}^n)$ *is compact if and only if it is closed, bounded, and equicontinuous.*

Proof. Step 1. We first show that if $\mathcal{F}$ is bounded under ρ, then there is a compact subset Y of R^n having the property that $f(x) \in Y$ for all $f \in \mathcal{F}$

and all $x \in X$. It then follows that $\mathcal{F}$ is contained in the subspace $\mathcal{C}(X, Y)$ of $\mathcal{C}(X, R^n)$.

Choose an element $f_0 \in \mathcal{F}$. Because $\mathcal{F}$ is bounded under ρ, there is a number M such that $\rho(f_0, f) < M$ for all $f \in \mathcal{F}$. Because X is compact, so is $f_0(X)$; therefore, we can choose a number N so that $f_0(X)$ lies in the ball $B_d(\mathbf{0}, N)$ in R^n. Then $f(X)$ lies in the ball $B_d(\mathbf{0}, M + N)$, for every $f \in \mathcal{F}$. Choose Y to be the closure of this ball; then $f(x) \in Y$ for every $x \in X$ and every $f \in \mathcal{F}$.

Step 2. Suppose that $\mathcal{F}$ is compact. Then $\mathcal{F}$ is closed and bounded under ρ. By Step 1, $\mathcal{F}$ is contained in some subspace $\mathcal{C}(X, Y)$ of $\mathcal{C}(X, R^n)$, where Y is compact. Since $\mathcal{F}$ is compact, $\mathcal{F}$ is totally bounded under ρ, by Theorem 3.1. Because both X and Y are compact, Lemma 3.2 applies to show that $\mathcal{F}$ is equicontinuous.

Step 3. Suppose that $\mathcal{F}$ is closed and bounded and equicontinuous. Since $\mathcal{F}$ is a closed subset of the complete metric space $(\mathcal{C}(X, R^n), \rho)$, it is necessarily complete. Because $\mathcal{F}$ is bounded, by Step 1 it is contained in some subspace $\mathcal{C}(X, Y)$ of $\mathcal{C}(X, R^n)$, where Y is compact. Then since $\mathcal{F}$ is equicontinuous, and X and Y are compact, Lemma 3.2 implies that $\mathcal{F}$ is totally bounded. It follows from Theorem 3.1 that $\mathcal{F}$ is compact. □

Exercises

1. Let (X, d) be a metric space.
   (a) Show that if X is totally bounded, then X is bounded.
   (b) Show that X is totally bounded under d if and only if it is totally bounded under $\bar{d}(x, y) = \min\{d(x, y), 1\}$.
2. Let A be a subset of the complete metric space X. Show that A is totally bounded if and only if $\bar{A}$ is compact.
3. Consider a countable product $X = \prod X_n$ of metrizable spaces; use on X the metric D that gives the product topology. (See Exercise 3 of §2-10.) Show that if each space X_n is totally bounded, so is X. Conclude without using the Tychonoff theorem that a countable product of compact metrizable spaces is compact.
4. (a) Show that any finite set of continuous functions is equicontinuous.
   (b) Suppose that $\mathcal{F}$ is a collection of differentiable functions $f: [0, 1] \to R$ whose derivatives are uniformly bounded. [This means there is an M such that $|f'(x)| \leq M$ for all f in $\mathcal{F}$ and all x.] Show that $\mathcal{F}$ is equicontinuous.
   (c) Let $f_n: [0, 1] \to R$ be the function $f_n(x) = x^n$; let $\mathcal{F}$ be the collection $\{f_n\}$. Show that $\mathcal{F}$ is closed and bounded but has no limit point in $\mathcal{C}(I, R)$. At what point or points does $\mathcal{F}$ fail to be equicontinuous?
5. Let X be compact; give $\mathcal{C}(X, R^n)$ the sup metric ρ. We say that a subset $\mathcal{F}$ of $\mathcal{C}(X, R^n)$ is **uniformly bounded** if it is bounded under ρ. We say it is **pointwise**

bounded if for each $x \in X$, the set

$$\mathcal{F}_x = \{f(x) \mid f \in \mathcal{F}\}$$

is a bounded subset of R^n.

(a) Show that if $\mathcal{F}$ is equicontinuous and pointwise bounded, it is uniformly bounded.

(b) Show that if $\mathcal{F}$ is equicontinuous, so is its closure.

(c) *Theorem. Let X be compact; consider $\mathcal{C}(X, R^n)$ in the metric ρ. A subset $\mathcal{F}$ of $\mathcal{C}(X, R^n)$ has compact closure if and only if it is equicontinuous and pointwise bounded.*

(d) *Theorem (Arzela's theorem). Let X be compact; let $f_n \in \mathcal{C}(X, R^k)$. If the collection $\{f_n\}$ is pointwise bounded ond equicontinuous, then the sequence (f_n) has a uniformly convergent subsequence.*

6. Let X be locally compact Hausdorff. A subset $\mathcal{F}$ of $\mathcal{C}(X, R)$ is said to **vanish uniformly at infinity** if given $\epsilon > 0$, there is a compact subset C of X such that $|f(x)| < \epsilon$ for $x \in X - C$ and $f \in \mathcal{F}$. If $\mathcal{F}$ consists of a single function f, we say simply that f **vanishes at infinity**. Let $\mathcal{C}_0(X, R)$ denote the set of continuous functions $f: X \longrightarrow R$ that vanish at infinity. The sup metric ρ is defined on $\mathcal{C}_0(X, R)$.

Theorem. A subset $\mathcal{F}$ of $(\mathcal{C}_0(X, R), \rho)$ is compact if and only if it is closed, bounded, equicontinuous, and vanishes uniformly at infinity.

[*Hint:* Let Y denote the one-point compactification of X. Show that $\mathcal{C}_0(X, R)$ is isometric with a subspace of $\mathcal{C}(Y, R)$.]

***7.** Let (X, d) be a metric space. If $A \subset X$ and $\epsilon > 0$, define

$$U(A, \epsilon) = \bigcup_{a \in A} B_d(a, \epsilon).$$

Let $\mathcal{H}$ be the collection of all (nonempty) closed, bounded subsets of X. If $A, B \in \mathcal{H}$, define

$$D(A, B) = \text{glb}\,\{\epsilon \mid A \subset U(B, \epsilon) \quad \text{and} \quad B \subset U(A, \epsilon)\}.$$

(a) Show that D is a metric on $\mathcal{H}$; it is called the **Hausdorff metric.**

(b) Show that if (X, d) is complete, so is $(\mathcal{H}, D)$. [*Hint:* Let A_n be a Cauchy sequence in $\mathcal{H}$; by passing to a subsequence, assume $D(A_n, A_{n+1}) < 1/2^n$. Define A to be the set of all points x that are the limits of sequences $x_1, x_2, \ldots$ such that $x_i \in A_i$ for each i and $d(x_i, x_{i+1}) < 1/2^i$. Show $A_n \longrightarrow A$.]

(c) Show that if (X, d) is totally bounded, so is $(\mathcal{H}, D)$. [*Hint:* Given ϵ, choose $\delta < \epsilon$ and let S be a finite subset of X such that the collection $\{B_d(x, \delta) \mid x \in S\}$ covers X. Let $\mathcal{A}$ be the collection of all nonempty subsets of S; show that $\{B_D(A, \epsilon) \mid A \in \mathcal{A}\}$ covers $\mathcal{H}$.]

(d) *Theorem. If X is compact in the metric d, then the space $\mathcal{H}$ of all nonempty closed bounded subsets of X is compact in the Hausdorff metric D.*

8. Which half of the proof of Lemma 3.2 uses compactness of Y?

7-4 Pointwise and Compact Convergence

There are other useful topologies on the spaces Y^X and $\mathcal{C}(X, Y)$ in addition to the uniform topology. We shall consider two of them here; they are called the *topology of pointwise convergence* and the *topology of compact convergence.*

Definition. Given a point x of the set X and an open set U of the space Y, let

$$S(x, U) = \{f \mid f \in Y^X \text{ and } f(x) \in U\}.$$

The sets $S(x, U)$ are a subbasis for topology on Y^X, which is called the **topology of pointwise convergence** (or the **point-open topology**).

The general basis element for this topology is a finite intersection of subbasis elements $S(x, U)$. Thus a typical basis element about the function f consists of all functions g that are "close" to f at finitely many points. Such a neighborhood is illustrated in Figure 6; it consists of all functions g whose graphs intersect the three vertical intervals pictured.

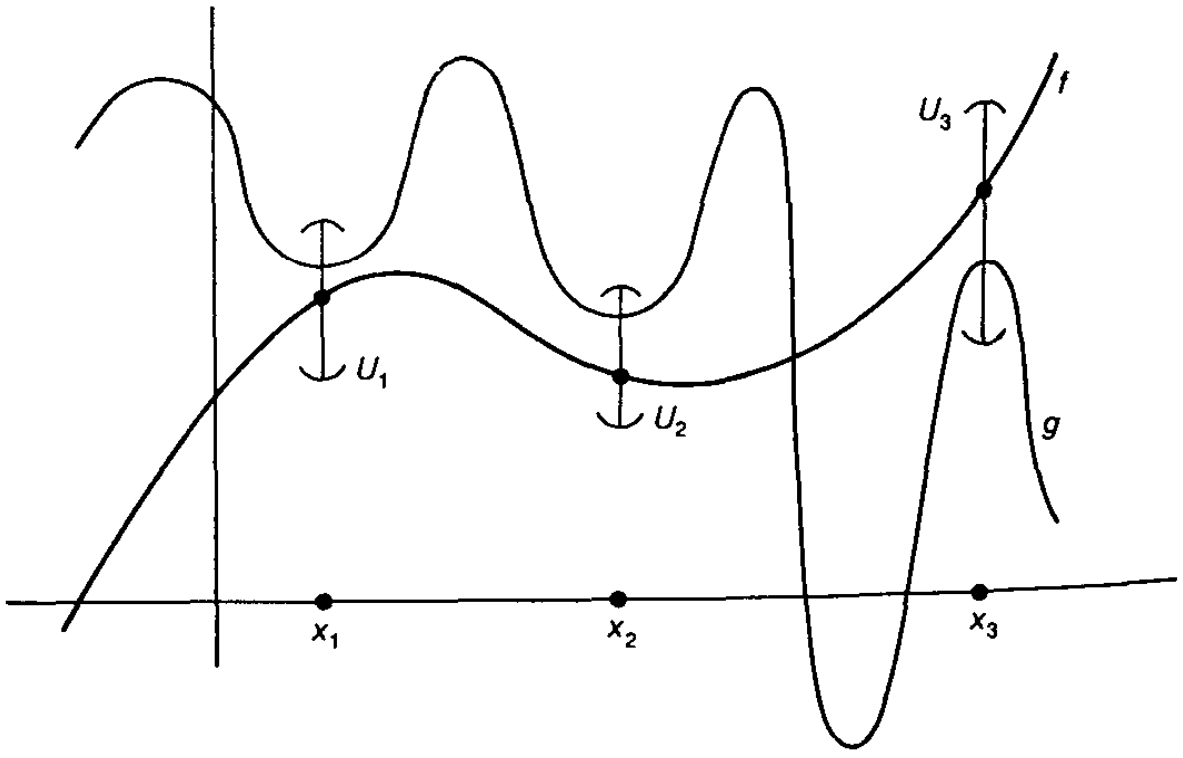


Figure 6

The topology of pointwise convergence on Y^X is nothing new. It is just the product topology we have already studied. If we replace X by J and denote the general element of J by α to make it look more familiar, then the set $S(\alpha, U)$ of all functions $\mathbf{x} : J \to Y$ such that $\mathbf{x}(\alpha) \in U$ is just the subset

$\pi_\alpha^{-1}(U)$ of Y^J, which is the standard subbasis element for the product topology.

The reason for calling it the topology of pointwise convergence comes from the following theorem:

Theorem 4.1. *A sequence* f_n *of functions converges to the function* f *in the topology of pointwise convergence if and only if for each* x *in* X, *the sequence* $f_n(x)$ *of points of* Y *converges to the point* $f(x)$.

Proof. This is a standard fact about the product topology; we prove it here using functional notation. Suppose that f_n converges in the topology of pointwise convergence. Given $x \in X$, and given an open set U about $f(x)$, the set $S(x, U)$ is a neighborhood of f. Therefore, there is an integer N such that $f_n \in S(x, U)$ for all $n \geq N$. Then $f_n(x) \in U$ for all $n \geq N$.

Conversely, suppose $f_n(x)$ converges to $f(x)$ for each x. To show that f_n converges to f in the topology of pointwise convergence, it suffices to show that if $S(x, U)$ is an arbitrary subbasis element about f, then $S(x, U)$ contains all f_n for n sufficiently large. (Why?) But since $f_n(x)$ converges to $f(x)$ and $f(x) \in U$, there must be an integer N such that $f_n(x) \in U$ for $n \geq N$. Then $f_n \in S(x, U)$ for $n \geq N$. □

EXAMPLE 1. Consider the space R^I, where $I = [0, 1]$. The sequence (f_n) of continuous functions given by $f_n(x) = x^n$ converges in the topology of pointwise convergence to the function f defined by

$$f(x) = \begin{cases} 0 & \text{for } 0 \leq x < 1, \\ 1 & \text{for } x = 1. \end{cases}$$

This example shows that the subspace $\mathcal{C}(I, R)$ of continuous functions is not closed in R^I in the topology of pointwise convergence.

We know that a sequence (f_n) of continuous functions which converges in the uniform topology has a continuous limit, and the preceding example shows that a sequence which converges only in the topology of pointwise convergence need not. One can ask whether there is a topology intermediate between these two which nevertheless will ensure that the limit of a convergent sequence of continuous functions is continuous. The answer is "yes"; assuming the (fairly mild) restriction that the space X be compactly generated, it will suffice if f_n converges to f in the topology of compact convergence, which we now define.

Definition. Let (Y, d) be a metric space; let X be a topological space. Given an element f of Y^X, a compact subset C in X, and a number $\epsilon > 0$, let $B_C(f, \epsilon)$ denote the set of all those elements g of Y^X for which

$$\text{lub } \{d(f(x), g(x)) \mid x \in C\} < \epsilon.$$

The sets $B_C(f, \epsilon)$ form a basis for a topology on Y^X. It is called the **topology of compact convergence** (or sometimes the "topology of uniform convergence on compact sets").

This topology differs from the preceding topology in that the general basis element containing f consists of functions that are "close" to f not just at finitely many points, but at all points of some compact set.

Note that if $g \in B_C(f, \epsilon)$, then there is a $\delta > 0$ such that we have $B_C(g, \delta) \subset B_C(f, \epsilon)$; simply let

$$\delta = \epsilon - \text{lub}\,\{d(f(x), g(x)) \mid x \in C\}.$$

It is then easy to check that these sets form a basis.

The justification for the choice of terminology comes from the following theorem, whose proof is immediate.

Theorem 4.2. *A sequence* $f_n : X \to Y$ *of functions converges to the function* f *in the topology of compact convergence if and only if for each compact subset* C *of* X, *the sequence* $f_n | C$ *converges uniformly to* $f | C$.

Definition. A space X is said to be **compactly generated** if it satisfies the following condition: A set A is open in X if and only if $A \cap C$ is open in C for each compact set C in X.

This condition is equivalent to requiring that a set B be closed in X if and only if $B \cap C$ is closed in C for each compact set C. It is a fairly mild restriction on the space; many familiar spaces are compactly generated. For instance:

Lemma 4.3. *If* X *is locally compact, or if* X *satisfies the first countability axiom, then* X *is compactly generated.*

Proof. Suppose that X is locally compact. Let $A \cap C$ be open in C for every compact subset C of X. We show A is open. Given $x \in A$, choose a neighborhood U of x which lies in a compact set C. Since $A \cap C$ is open in C by hypothesis, $A \cap U$ is open in U, and hence open in X. Then $A \cap U$ is a neighborhood of x contained in A, so that A is open in X.

Suppose that X satisfies the first countability axiom. If $B \cap C$ is closed in C for each compact subset C of X, we show that B is closed in X. Let x be a point of $\bar{B}$; we show that $x \in B$. Since X has a countable basis at x, there is a sequence (x_n) of points of B converging to x. The set

$$C = \{x\} \cup \{x_n \mid n \in Z_+\}$$

is compact, so that $B \cap C$ is by assumption closed in C. Since $B \cap C$ contains x_n for every n, it contains x as well. Therefore, $x \in B$, as desired. □

Theorem 4.4. *Let* X *be a compactly generated space; let* (Y, d) *be a metric space. Then* $\mathcal{C}(X, Y)$ *is closed in* Y^X *in the topology of compact convergence.*

Proof. Let $f \in Y^X$ be a limit point of $\mathcal{C}(X, Y)$; we wish to show f is continuous.

First we show that $f \mid C$ is continuous for each compact set C in X. For each n, consider the neighborhood $B_C(f, 1/n)$ of f; it intersects $\mathcal{C}(X, Y)$, so we can choose a function $f_n \in \mathcal{C}(X, Y)$ lying in this neighborhood. The sequence of functions $f_n \mid C : C \longrightarrow Y$ converges uniformly to the function $f \mid C$, so that by the uniform limit theorem, $f \mid C$ is continuous.

It follows that f is continuous. Let V be an open subset of Y; we show that $f^{-1}(V)$ is open in X. Given any subset C of X,

$$f^{-1}(V) \cap C = (f \mid C)^{-1}(V).$$

If C is compact, this set is open in C because $f \mid C$ is continuous. Since X is compactly generated, it follows that $f^{-1}(V)$ is open in X. □

Corollary 4.5. *Let* X *be a compactly generated space; let* (Y, d) *be a metric space. If a sequence of continuous functions* $f_n : X \longrightarrow Y$ *converges to* f *in the topology of compact convergence, then* f *is continuous.*

Now we have three topologies for the function space Y^X, when Y is metric. The relation between them is stated in the following theorem, whose proof is straightforward.

Theorem 4.6. *Let* X *be a space; let* (Y, d) *be a metric space. For the function space* Y^X, *one has the following inclusions of topologies:*

$$(\textit{uniform}) \supset (\textit{compact convergence}) \supset (\textit{pointwise convergence}).$$

If X *is compact, the first two coincide, and if* X *is discrete, the second two coincide.*

Exercises

1. Show that the sets $B_C(f, \epsilon)$ form a basis for a topology on Y^X.
2. Prove Theorem 4.2.
3. Prove Theorem 4.6.
4. Consider Y^X in the topology of compact convergence.
   (a) Show Y^X is regular. Is it normal?
   (b) Show that if X equals a countable union of open sets with compact closures, then Y^X satisfies the first countability axiom.
5. Show that in general the set of bounded functions $f: X \longrightarrow R$ is not closed in R^X in the topology of compact convergence.

6. Consider the sequence of continuous functions $f_n : R_+ \longrightarrow R$ defined by

$$f_n(x) = 1/nx.$$

In which of the three topologies of Theorem 4.6 does this sequence converge? What about the sequence given in Exercise 9 of §2-10?

7. Consider the sequence of functions $f_n : (-1, 1) \longrightarrow R$, defined by

$$f_n(x) = \sum_{k=1}^{n} kx^k.$$

(a) Show that (f_n) converges in the topology of compact convergence; conclude that the limit function is continuous. (This is a standard fact about power series.)

(b) Show that (f_n) does not converge in the uniform topology.

8. A function $f: X \longrightarrow R$ is said to have compact support if f vanishes outside some compact set. Let $\mathcal{C}_c(R, R)$ be the set of all continuous functions $f: R \longrightarrow R$ having compact support. Define a metric on this set by the rule

$$D(f, g) = \left[\int_{-\infty}^{+\infty} (f(x) - g(x))^2 \, dx\right]^{1/2}.$$

(a) Show that D is a metric. The topology given by D is called the topology of mean convergence. [*Hint:* Define a "scalar product" by the rule

$$f \cdot g = \int_{-\infty}^{+\infty} f(x)g(x) \, dx;$$

then imitate the proof given in Exercise 8 of §2-9 for the euclidean metric.]

(b) In which of the four topologies we have for $\mathcal{C}_c(X, Y)$ does the sequence f_n pictured in Figure 7 converge? What about the sequence g_n pictured in Figure 8?

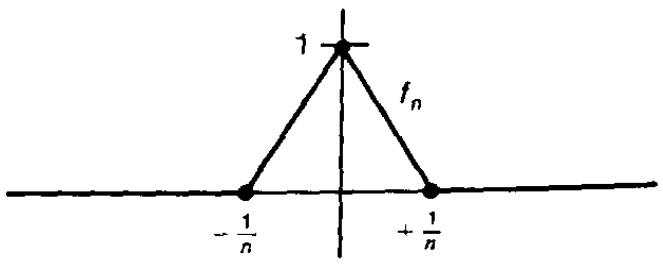


Figure 7

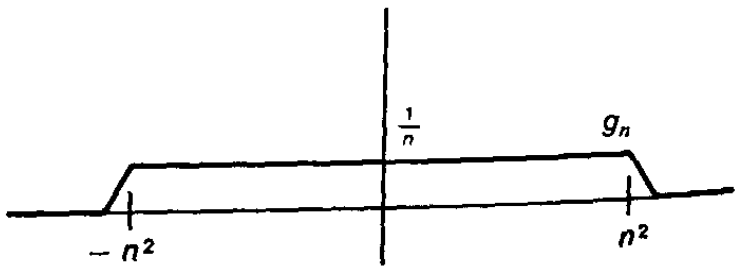


Figure 8

(c) Show that the mean convergence topology is not comparable with the other three.

9. Let (Y, d) be a metric space; let X be a space. Define a topology on $\mathcal{C}(X, Y)$ as follows: Given $f \in \mathcal{C}(X, Y)$, and given a positive continuous function $\delta : X \to R_+$ on X, let

$$B(f, \delta) = \{g \mid d(f(x), g(x)) < \delta(x) \quad \text{for all} \quad x \in X\}.$$

(a) Show that the sets $B(f, \delta)$ form a basis for a topology on $\mathcal{C}(X, Y)$. We call it the fine topology.
(b) Show that the fine topology contains the uniform topology.
(c) Show that if X is compact, the fine and uniform topologies agree.
(d) Show that if X is discrete, then $\mathcal{C}(X, Y) = Y^X$ and the fine and box topologies agree.

10. Here is a theorem about compactly generated spaces. We say that $f: X \to Y$ is a proper map if for each compact set D in Y, the set $f^{-1}(D)$ is compact in X.

Theorem. Let X be a space; let Y be a compactly generated Hausdorff space. If $f: X \to Y$ is continuous, injective, and proper, then f is an imbedding and $f(X)$ is closed in Y.

7-5 The Compact-Open Topology

In defining the uniform topology and the topology of compact convergence on the function space Y^X, we needed to assume a metric for the space Y; both definitions involved this metric. (The topology of pointwise convergence on the function space Y^X does not involve a metric for Y.)

In general, topologists are concerned primarily with matters that depend only on the topology of a space, not on its metric. So one naturally asks whether either of these first two topologies depends only on the topology of Y, rather than on the particular metric d.

There is no satisfactory answer to this question for the space Y^X of *all* functions mapping X into Y. But for the space $\mathcal{C}(X, Y)$ of continuous functions, one *can* prove something; for this space, the topology of compact convergence is independent of the particular metric chosen for Y.

To prove this, we define a certain topology on $\mathcal{C}(X, Y)$ called the *compact-open topology*, whose definition does not involve a metric for Y. Then we show that if Y happens to have a metric d, this topology agrees with the topology of compact convergence. Therefore, the latter topology on $\mathcal{C}(X, Y)$ depends only on the space Y, not on the metric d.

The compact-open topology is very important in its own right; it turns out to have many useful properties. We shall use it in the next section in proving a generalized version of Ascoli's theorem. It also has close connections with the notion of "homotopy," which is one of the fundamental concepts in algebraic topology.

Definition. Let X and Y be topological spaces. If C is a compact subset of X and U is an open subset of Y, define

$$S(C, U) = \{f \mid f \in \mathcal{C}(X, Y) \text{ and } f(C) \subset U\}.$$

The sets $S(C, U)$ form a subbasis for a topology on $\mathcal{C}(X, Y)$, called the **compact-open topology.**

Note that both the compact-open topology and the point-open topology are defined without assuming a metric on Y. It is clear from the definition that the compact-open topology is in general finer than the point-open topology.

Theorem 5.1. *Let* X *be a space and let* (Y, d) *be a metric space. For the space* $\mathcal{C}$(X, Y), *the compact-open topology and the topology of compact convergence coincide.*

Proof. Step 1. If A is a subset of Y and $\epsilon > 0$, we define

$$U(A, \epsilon) = \bigcup_{a \in A} B_d(a, \epsilon).$$

We call this set the "ϵ-neighborhood" of A. If A is compact and V is an open set containing A, we show there is an $\epsilon > 0$ such that $U(A, \epsilon) \subset V$.

This was given earlier as an exercise (Exercise 4 of §3-6); we give a proof here. For each $a \in A$, choose $\delta(a) > 0$ so that $B_d(a, \delta(a)) \subset V$. Cover A by finitely many open sets of the form

$$B_d(a_1, \tfrac{1}{2}\delta(a_1)), \ldots, B_d(a_n, \tfrac{1}{2}\delta(a_n)).$$

Let $\epsilon = \min \{\frac{1}{2}\delta(a_i)\}$. Then if $a \in A$, the point a is in at least one of the sets $B_d(a_i, \frac{1}{2}\delta(a_i))$; the triangle inequality implies that

$$B_d(a, \epsilon) \subset B_d(a_i, \delta(a_i)).$$

Since this holds for each a in A, it follows that $U(A, \epsilon) \subset V$, as desired.

Step 2. We prove that the topology of compact convergence is finer than the compact-open topology. Let $S(C, U)$ be a subbasis element for the compact-open topology on $\mathcal{C}(X, Y)$ and let f be an element of $S(C, U)$. Because f is continuous, $f(C)$ is a compact subset of the open set U. By Step 1, we can choose ϵ so the ϵ-neighborhood of $f(C)$ lies in U. Then

$$B_C(f, \epsilon) \subset S(C, U).$$

Step 3. We prove the compact-open topology is finer than the topology of compact convergence. Given an open set about f in the topology of compact convergence, it contains a basis element of the form $B_C(f, \epsilon)$. We shall find a basis element for the compact-open topology that contains f and lies in $B_C(f, \epsilon)$.

Each point x of X has a neighborhood V_x such that $f(\bar{V}_x)$ lies in an open set U_x of Y having diameter less than ϵ. [For example, choose V_x so that $f(V_x)$ lies in the $\epsilon/4$-neighborhood of $f(x)$. Then $f(\bar{V}_x)$ lies in the $\epsilon/3$-

neighborhood of $f(x)$, which has diameter at most $2\epsilon/3$.] Cover C by finitely many such sets V_x, say for $x = x_1, \ldots, x_n$. Let $C_x = \bar{V}_x \cap C$. Then C_x is compact, and the basis element

$$S(C_{x_1}, U_{x_1}) \cap \cdots \cap S(C_{x_n}, U_{x_n})$$

contains f and lies in $B_C(f, \epsilon)$, as desired. □

Corollary 5.2. Let Y *be a metric space. The compact convergence topology on the subspace* $\mathcal{C}(X, Y)$ *of* Y^X *does not depend on the metric of* Y.

The fact that the definition of the compact-open topology does not involve any metric is just one of its useful features. Another is the fact that it satisfies the requirement of "joint continuity." Roughly speaking, this means that the expression $f(x)$ is continuous not only in the single "variable" x, but is continuous jointly in both the "variables" x and f. More precisely, one has the following theorem:

Theorem 5.3. *Let* X *be locally compact Hausdorff; let* $\mathcal{C}(X, Y)$ *have the compact-open topology. Then the map*

$$e: X \times \mathcal{C}(X, Y) \longrightarrow Y$$

defined by the equation

$$e(x, f) = f(x)$$

is continuous.

The map e is called the **evaluation map.**

Proof. Given a point (x, f) of $X \times \mathcal{C}(X, Y)$ and an open set V in Y about the image point $e(x, f) = f(x)$, we wish to find an open set about (x, f) that e maps into V. First, using the continuity of f and the fact that X is locally compact Hausdorff, we can choose an open set U about x having compact closure $\bar{U}$, such that f carries $\bar{U}$ into V. Then consider the open set

$$U \times S(\bar{U}, V)$$

in $X \times \mathcal{C}(X, Y)$. It is an open set containing (x, f). And if (x', f') belongs to this set, then $e(x', f') = f'(x')$ belongs to V, as desired. □

A consequence of this theorem is the following theorem. We shall not need to use it, but it is important in algebraic topology.

Corollary 5.4. Let X *be a locally compact Hausdorff space; let* $\mathcal{C}(X, Y)$ *have the compact-open topology; let* Z *be an arbitrary topological space. Then a map* $F : X \times Z \longrightarrow Y$ *is continuous if and only if the induced map*

$$\hat{F}: Z \longrightarrow \mathcal{C}(X, Y)$$

is continuous, where $\hat{F}$ *is defined by the rule*

$$(\hat{F}(z))(x) = F(x, z).$$

Proof. Suppose that $\hat{F}$ is continuous. It follows that F is continuous, since F equals the composite

$$X \times Z \xrightarrow{i_X \times \hat{F}} X \times \mathcal{C}(X, Y) \xrightarrow{e} Y.$$

Suppose that F is continuous; we prove that

$$\hat{F}: Z \to \mathcal{C}(X, Y)$$

is continuous. (This part of the proof does not use the local compactness of X.) To prove continuity, we take a point z_0 of Z and a subbasis element $S(C, U)$ of $\mathcal{C}(X, Y)$ containing $\hat{F}(z_0)$, and find a neighborhood W of z_0 that is mapped by $\hat{F}$ into $S(C, U)$. This will suffice.

The statement that $\hat{F}(z_0)$ lies in $S(C, U)$ means simply that $(\hat{F}(z_0))(x) = F(x, z_0)$ is in U for all $x \in C$. That is, $F(C \times z_0) \subset U$. Continuity of F implies that $F^{-1}(U)$ is an open set in $X \times Z$ containing $C \times z_0$. Then

$$F^{-1}(U) \cap (C \times Z)$$

is an open set in $C \times Z$ about the slice $C \times z_0$; the tube lemma of §3-5 shows that there is a neighborhood W of z_0 in Z such that the entire tube $C \times W$ lies in $F^{-1}(U)$. See Figure 9. Then for $z \in W$ and $x \in C$, we have $F(x, z) \in U$. Hence $\hat{F}(W) \subset S(C, U)$, as desired. □

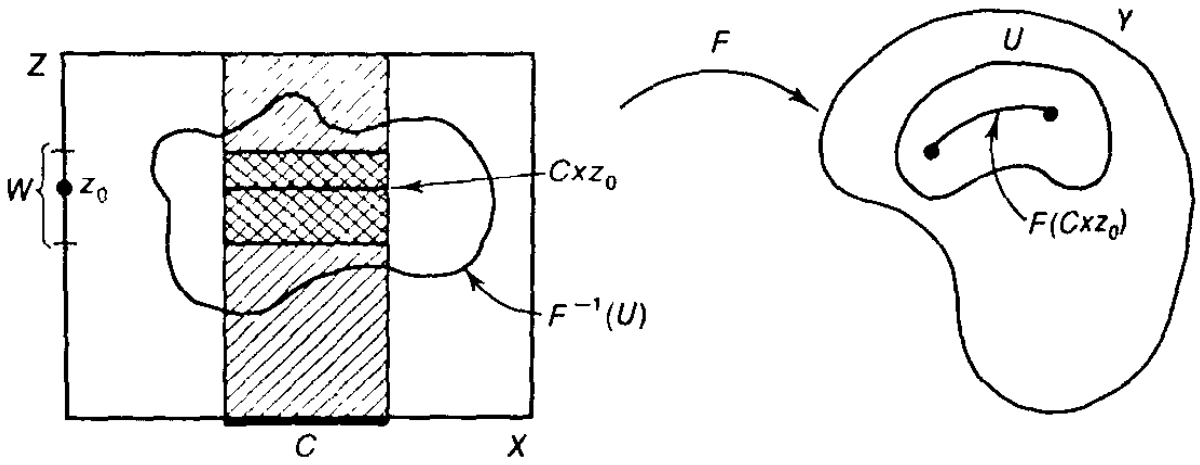


Figure 9

We discuss briefly the connections between the compact-open topology and the concept of *homotopy*.

If f and g are continuous maps of X into Y, we say that f and g are **homotopic** if there is a continuous map

$$F: X \times [0, 1] \to Y$$

such that $F(x, 0) = f(x)$ and $F(x, 1) = g(x)$ for each $x \in X$. The map F is called a **homotopy** between f and g.

Roughly speaking, a homotopy is a "continuous one-parameter family"

of maps from X to Y. More precisely, we note that a homotopy F gives rise to a map

$$\hat{F}: [0, 1] \longrightarrow \mathcal{C}(X, Y)$$

which assigns, to each parameter value t in $[0, 1]$, the corresponding continuous map from X to Y. Assuming that X is locally compact Hausdorff, we see that F is continuous if and only if $\hat{F}$ is continuous. This means that a homotopy F between two continuous maps $f, g : X \longrightarrow Y$ corresponds precisely to a *path* $\hat{F}$ in the function space $\mathcal{C}(X, Y)$ joining the point f of $\mathcal{C}(X, Y)$ to the point g.

We shall return to a more detailed study of homotopy in the next chapter.

Exercises

1. Let C be a subspace of X. Show that the restriction map $r : \mathcal{C}(X, Y) \longrightarrow \mathcal{C}(C, Y)$ is continuous if both spaces have the point-open or compact-open topologies.

2. Show that in the compact-open topology, $\mathcal{C}(X, Y)$ is Hausdorff if Y is Hausdorff, and regular if Y is regular. [*Hint:* If $\bar{U} \subset V$, then $\overline{S(C, U)} \subset S(C, V)$.]

3. Let Y be a metric space. Show that if X is compact, the uniform topology on $\mathcal{C}(X, Y)$ is independent of the metric chosen for Y.

4. Show that if Y is locally compact Hausdorff, then composition of maps

$$\mathcal{C}(X, Y) \times \mathcal{C}(Y, Z) \longrightarrow \mathcal{C}(X, Z)$$

is continuous, provided the compact-open topology is used throughout. [*Hint:* If $g \circ f \in S(C, U)$, find V such that $f(C) \subset V$ and $g(\bar{V}) \subset U$.]

5. Let $\mathcal{C}'(X, Y)$ denote the set $\mathcal{C}(X, Y)$ in some topology $\mathfrak{I}$. Show that if the evaluation map

$$e : X \times \mathcal{C}'(X, Y) \longrightarrow Y$$

is continuous, then $\mathfrak{I}$ contains the compact-open topology. [*Hint:* Consider $\hat{e} : \mathcal{C}'(X, Y) \longrightarrow \mathcal{C}(X, Y)$. Note that local compactness of X is not needed.]

6. Here is an (unexpected) application of Corollary 5.4 to quotient maps: Show that if $p : A \longrightarrow B$ is a quotient map and X is locally compact Hausdorff, then $i_X \times p : X \times A \longrightarrow X \times B$ is a quotient map. Compare Exercise 11 of §3-8. [*Hint:* Let $(X \times B)'$ denote the set $X \times B$ in the quotient topology induced by $i_X \times p$; let $\pi : X \times A \longrightarrow (X \times B)'$ be the quotient map. If $g : X \times B \longrightarrow (X \times B)'$ is the identity map, show that $\hat{g} \circ p = \hat{\pi}$; conclude that $\hat{g}$ is continuous.]

7-6 Ascoli's Theorem

Now we prove a generalized version of Ascoli's theorem. It includes the earlier one as a special case. We shall use the Tychonoff theorem in the course of the proof.

Theorem 6.1 (Ascoli's theorem). *Let* X *be locally compact Hausdorff; let* (Y, d) *be a metric space. Consider* $\mathcal{C}(X, Y)$ *in the compact-open topology. A subset* $\mathcal{F}$ *of* $\mathcal{C}(X, Y)$ *has compact closure if and only if it is equicontinuous and the subset*

$$\mathcal{F}_x = \{f(x) \mid f \in \mathcal{F}\}$$

of Y *has compact closure for each* x.

Proof. Step 1. We first show that if a subset $\mathcal{F}$ of $\mathcal{C}(X, Y)$ is equicontinuous, then the topologies it inherits from the topologies of pointwise convergence and compact convergence on Y^X are the same.

In general, the compact convergence topology is finer than the pointwise convergence topology; we wish to prove the reverse inclusion holds for the subspace $\mathcal{F}$. Given $f \in \mathcal{F}$, consider the basis element $B_C(f, \epsilon)$ for the compact convergence topology. It consists of all functions g such that

$$\text{lub } \{d(g(x), f(x)) \mid x \in C\} < \epsilon.$$

We find a basis element B for the pointwise convergence topology such that

$$f \in B \cap \mathcal{F} \subset B_C(f, \epsilon) \cap \mathcal{F}.$$

Choose ϵ_1 and ϵ_2 so that $2\epsilon_1 + \epsilon_2 \leq \epsilon$. Using equicontinuity of $\mathcal{F}$ and compactness of C, cover C by finitely many open sets $U_1, \ldots, U_n$ of X such that on each set U_i, each of the functions in $\mathcal{F}$ varies by at most ϵ_1. Choose $x_i \in U_i$ for each i. Consider the subset B of Y^X given by the equation

$$B = \{g \mid d(g(x_i), f(x_i)) < \epsilon_2 \text{ for } i = 1, \ldots, n\};$$

we assert that it is the desired basis element for the pointwise convergence topology.

Let g be an element of $B \cap \mathcal{F}$. Given $x \in C$, choose i so that x belongs to U_i. Then

$$d(g(x), g(x_i)) \leq \epsilon_1,$$
$$d(g(x_i), f(x_i)) < \epsilon_2,$$
$$d(f(x_i), f(x)) \leq \epsilon_1,$$

so that $d(g(x), f(x)) < \epsilon$. This inequality holds for each x in C. Therefore,

$$\max \{d(g(x), f(x)) \mid x \in C\} < \epsilon,$$

so that $g \in B_C(f, \epsilon)$, as desired.

Step 2. Second, we show that if a subset $\mathcal{F}$ of Y^X is equicontinuous, then its closure $\mathcal{G}$ in Y^X relative to the topology of pointwise convergence is also equicontinuous.

This is easy. Given $x_0 \in X$ and $\epsilon > 0$, choose a neighborhood U of x_0 such that

$$(*) \qquad d(f(x), f(x_0)) < \epsilon/3 \quad \text{for all } f \in \mathcal{F} \text{ and all } x \in U.$$

We assert that if g belongs to $\mathcal{G}$, then $d(g(x), g(x_0)) < \epsilon$ for all $x \in U$. It follows that $\mathcal{G}$ is equicontinuous.

To prove this assertion, let x be a point of U. Let V_x be the subset of Y^X, open in the pointwise convergence topology, consisting of all elements h of Y^X such that

$$(**) \qquad d(h(x), g(x)) < \epsilon/3 \quad \text{and} \quad d(h(x_0), g(x_0)) < \epsilon/3.$$

Because g belongs to the closure of $\mathcal{F}$ in this topology, the neighborhood V_x of g must contain an element f of $\mathcal{F}$. Applying the triangle inequality to $(*)$ and $(**)$, we see that $d(g(x), g(x_0)) < \epsilon$, as asserted.

Step 3. Now we prove the theorem. Suppose first that $\mathcal{F}$ is equicontinuous and that the set $\mathcal{F}_x$ has compact closure in Y for each $x \in X$.

Let $\mathcal{G}$ denote the closure of $\mathcal{F}$ in Y^X relative to the pointwise convergence topology. Let $\mathcal{G}'$ denote the closure of $\mathcal{F}$ in Y^X relative to the topology of compact convergence. Then $\mathcal{F} \subset \mathcal{G}' \subset \mathcal{G}$.

Because $\mathcal{F}$ is equicontinuous, it follows from Step 2 that $\mathcal{G}$ is equicontinuous. Therefore, by Step 1, the topologies $\mathcal{G}$ inherits from the topologies of pointwise convergence and compact convergence on Y^X are the same. We conclude that $\mathcal{G}' = \mathcal{G}$.

So far, we have used only the equicontinuity of $\mathcal{F}$. Now let C_x denote the closure in Y of $\mathcal{F}_x$. Consider the product $\prod_{x \in X} C_x$; we have

$$\mathcal{F} \subset \prod_{x \in X} C_x \subset Y^X,$$

by elementary set theory. The topology of pointwise convergence on Y^X is the same as the product topology. In this topology, the set $\prod C_x$ is compact, by the Tychonoff theorem. Since $\mathcal{G}$ is a closed subset of $\prod C_x$ in this topology, it is compact in this topology. Therefore, $\mathcal{G}$ is also compact in the topology of compact convergence.

Now $\mathcal{G}$ is equicontinuous, so each function in $\mathcal{G}$ is continuous. Hence $\mathcal{G}$ is contained in the set $\mathcal{C}(X, Y)$, a set on which the topology of compact convergence equals the compact-open topology. Therefore, the closure of $\mathcal{F}$ in $\mathcal{C}(X, Y)$ in the compact-open topology is the compact set $\mathcal{G}$, as desired.

Step 4. Give $\mathcal{C}(X, Y)$ the compact-open topology; suppose that $\mathcal{F}$ has compact closure $\bar{\mathcal{F}}$ in $\mathcal{C}(X, Y)$. We show that $\mathcal{F}$ is equicontinuous and $\mathcal{F}_x$ has compact closure. The proof makes use of the fact that the evaluation map

$$e: X \times \mathcal{C}(X, Y) \to Y$$

is continuous; here we make our first use of the local compactness of X.

First, given $x \in X$, we show $\mathcal{F}_x$ has compact closure. This is easy. The set $x \times \bar{\mathcal{F}}$ is compact in $X \times \mathcal{C}(X, Y)$; therefore, $e(x \times \bar{\mathcal{F}})$ is a compact set in Y. But $e(x \times \bar{\mathcal{F}})$ contains $e(x \times \mathcal{F}) = \mathcal{F}_x$.

Second, we show $\mathcal{F}$ equicontinuous at x_0. In fact, we show $\bar{\mathcal{F}}$ is equicon-

tinuous at x_0. Choose a compact set C that contains a neighborhood of x_0. It will suffice to show the collection of functions

$$\mathcal{R} = \{f|C\,; f \in \bar{\mathcal{F}}\}$$

is equicontinuous, since C contains a neighborhood of x_0.

Consider the restriction map

$$r\colon \mathcal{C}(X, Y) \longrightarrow \mathcal{C}(C, Y)$$

defined by $r(f) = f|C$. Then $\mathcal{R} = r(\bar{\mathcal{F}})$. If we give both spaces the compact-open topology, r is continuous, as you can easily check. Therefore, $\mathcal{R}$ is a compact subset of $\mathcal{C}(C, Y)$ in the compact-open topology.

On the space $\mathcal{C}(C, Y)$, the compact-open topology equals the uniform topology (because C is compact). Therefore, $\mathcal{R}$ is compact in the uniform topology; by Theorem 3.1, $\mathcal{R}$ is totally bounded in the metric ρ. We assert that there is a compact subset Y_0 of Y such that $\mathcal{R}$ lies in the subspace $\mathcal{C}(C, Y_0)$ of $\mathcal{C}(C, Y)$. The theorem then follows at once from Lemma 3.2, for since C and Y_0 are compact, total boundedness of $\mathcal{R}$ implies equicontinuity of $\mathcal{R}$.

To find Y_0 is easy. The set $C \times \bar{\mathcal{F}}$ is compact in $X \times \mathcal{C}(X, Y)$; let Y_0 denote its image under e. Because e is continuous, Y_0 is a compact subset of Y; it contains $f(x)$ for every $x \in C$ and every $f \in \bar{\mathcal{F}}$. It follows that $\mathcal{R} \subset \mathcal{C}(C, Y_0)$. □

An even more general version of Ascoli's theorem may be found in [K] or [Wd]. There it is not assumed that Y is a metric space, but only that it has what is called a *uniform structure*, which is a generalization of the notion of metric.

Exercises

1. (a) Which step in the proof of Ascoli's theorem uses the local compactness of X? Which step uses the Tychonoff theorem?
   (b) Show that the classical version of Ascoli's theorem is a special case of the general version, and so is Exercise 5(c) of §7-3.

2. Prove the following generalized version of Arzela's theorem:

   Theorem. Let X be locally compact Hausdorff with a countable basis; let $f_n\colon X \longrightarrow R^k$. If the collection $\{f_n\}$ is pointwise bounded and equicontinuous, then the sequence (f_n) has a subsequence that converges in the topology of compact convergence.

3. Which of the following subsets of $\mathcal{C}(R, R)$ are pointwise bounded? Which have compact closure relative to the topology of compact convergence?
   (a) The collection $\{f_n\}$, where $f_n(x) = x + \sin nx$.
   (b) The collection $\{g_n\}$, where $g_n(x) = x^n$.
   (c) The collection $\{h_n\}$, where $h_n(x) = |x|^{1/n}$.

(d) The collection of all polynomials of degree at most 10, whose coefficients have absolute value less than 1.

4. Let Y be a metric space; let $f_n: X \longrightarrow Y$ be a sequence of continuous functions; let $f: X \longrightarrow Y$ be a function (not necessarily continuous).
   (a) Show that if $\{f_n\}$ is equicontinuous and $f_n \longrightarrow f$ in the topology of pointwise convergence, then $f_n \longrightarrow f$ in the topology of compact convergence.
   (b) Prove the converse of (a) if X is locally compact Hausdorff.
   (c) Show that in either case, f is continuous.

7-7 *Baire Spaces*

The defining condition for a Baire space is probably as "unnatural looking" as any condition we have yet introduced in this book. But bear with us awhile.

In this section, we shall define Baire spaces and shall show that two important classes of spaces—the complete metric spaces and the compact Hausdorff spaces—are contained in the class of Baire spaces. Then we shall give some applications which, even if they do not make the Baire condition seem any more natural, will at least show what a useful tool it can be. In fact, it turns out to be a very useful and fairly sophisticated tool in both analysis and topology.

Definition. Recall that if A is a subset of a space X, the *interior* of A is defined as the union of all open sets of X that are contained in A. To say that A has empty interior is to say then that A contains no open set of X other than the empty set.

EXAMPLE 1. The set Q of rationals has empty interior as a subset of R, but the interval $[0, 1]$ has nonempty interior. The interval $[0, 1] \times 0$ has empty interior as a subset of the plane R^2, and so does the subset $Q \times R$.

Definition. A space X is said to be a **Baire space** if the following condition holds: Given any countable collection $\{A_n\}$ of closed sets of X each of which has empty interior in X, their union $\bigcup A_n$ also has empty interior in X.

EXAMPLE 2. The set Q of rationals is not a Baire space. For each one-point set in Q is closed and has empty interior in Q; and Q is the countable union of its one-point subsets.

The set Z_+, on the other hand, does form a Baire space. Every subset of Z_+ is open, so that there exist no subsets of Z_+ having empty interior, except for the empty set. Therefore, Z_+ satisfies the Baire condition vacuously.

More generally, every closed subset of R, being a complete metric space, is a Baire space. Somewhat surprising is the fact that the irrationals in R also form a Baire space; see Exercise 7.

The terminology originally used by R. Baire for this concept involved the word "category." A space X was said to be of the *first category* if it equaled the union of a countable collection of closed sets in X having empty interiors; otherwise, it was said to be of the *second category*. Using this terminology, we can say the following:

A space X is a Baire space if and only if every nonempty open set in X is of the second category.

We shall not use the terms "first category" and "second category" in this book.

The preceding definition is the "closed set definition" of a Baire space. There is also a formulation involving open sets that is frequently useful. It is given in the following lemma.

Lemma 7.1. X *is a Baire space if and only if given any countable collection* $\{U_n\}$ *of open sets in* X, *each of which is dense in* X, *their intersection* $\bigcap U_n$ *is also dense in* X.

Proof. Recall that a set C is dense in X if $\bar{C} = X$. The theorem now follows at once from the two remarks:

(1) A is closed in X if and only if $X - A$ is open in X.

(2) B has empty interior in X if and only if $X - B$ is dense in X. □

There are a number of theorems giving conditions under which a space is a Baire space. The most important is the following:

Theorem 7.2. If X *is a compact Hausdorff space, or a complete metric space, then* X *is a Baire space.*

Proof. Given a countable collection $\{A_n\}$ of closed sets in X having empty interiors, we want to show that their union $\bigcup A_n$ also has empty interior in X. So, given the nonempty open set U_0 of X, we must find a point x of U_0 that does not lie in any of the sets A_n.

Consider the first set A_1. By hypothesis, A_1 does not contain U_0. Therefore, we may choose a point y of U_0 that is not in A_1. Regularity of X, along with the fact that A_1 is closed, enables us to choose a neighborhood U_1 of y such that

$$\bar{U}_1 \cap A_1 = \varnothing,$$
$$\bar{U}_1 \subset U_0.$$

If X is metric, we also choose U_1 small enough that its diameter is less than 1.

In general, given the nonempty open set U_{n-1}, we choose a point of U_{n-1} that is not in the closed set A_n, and then we choose U_n to be a neighborhood of this point such that

$$\bar{U}_n \cap A_n = \varnothing,$$
$$\bar{U}_n \subset U_{n-1},$$
$$\operatorname{diam} U_n < 1/n \quad \text{in the metric case.}$$

We assert that the intersection $\bigcap \bar{U}_n$ is nonempty. From this fact our theorem will follow. For if x is a point of $\bigcap \bar{U}_n$, then x is in U_0, because $\bar{U}_1 \subset U_0$. And for each n, the point x is not in A_n, because $\bar{U}_n$ is disjoint from A_n.

The proof that $\bigcap \bar{U}_n$ is nonempty splits into two parts, depending on whether X is compact Hausdorff or complete metric. If X is compact Hausdorff, we consider the nested sequence $\bar{U}_1 \supset \bar{U}_2 \supset \cdots$ of nonempty subsets of X. The collection $\{\bar{U}_n\}$ satisfies the finite intersection condition; since X is compact, the intersection $\bigcap \bar{U}_n$ must be nonempty.

If X is complete metric, we apply the following lemma. □

Lemma 7.3. Let $C_1 \supset C_2 \supset \cdots$ *be a nested sequence of nonempty closed sets in the complete metric space* X. *If diam* $C_n \to 0$, *then* $\bigcap C_n \neq \varnothing$.

Proof. We gave this as an exercise in §7-1. Here is a proof: Choose $x_n \in C_n$ for each n. Because $x_n, x_m \in C_N$ for $n, m \geq N$, and because diam C_N can be made less than any given ϵ by choosing N large enough, the sequence (x_n) is a Cauchy sequence. Suppose that it converges to x. Then for given k, the subsequence $x_k, x_{k+1}, \ldots$ also converges to x. Thus x necessarily belongs to $\bar{C}_k = C_k$. Then $x \in \bigcap C_k$, as desired. □

The theorem just proved about Baire spaces is the one most frequently applied, but it is not the strongest one that can be proved. One can prove for instance that any G_δ set in a compact Hausdorff space or a complete metric space is a Baire space. These and other generalizations are given in the exercises.

EXAMPLE 3. *The set Q of rationals is not a G_δ set in the reals.*

Suppose that Q is the intersection of the countable collection $\{W_n\}$ of open sets of R. For each $q \in Q$, let V_q be the open set $R - \{q\}$. Then the collection

$$\mathcal{A} = \{W_n \mid n \in Z_+\} \cup \{V_q \mid q \in Q\}$$

is a countable collection of open sets of R. Each W_n is dense in R, since $W_n \supset Q$. And each set V_q is dense in R, being the complement of a single point. Since R is a Baire space, the intersection A of the elements of the countable collection $\mathcal{A}$ must be dense in R. But any point belonging to all the sets W_n is rational, and any point belonging to all the sets V_q is irrational. Hence A is empty.

EXAMPLE 4. The second application is more amusing than profound. It arises from a question that could be posed to a class in elementary calculus: *Is there a function $f: R \to R$ that is continuous precisely at the rational points of R?*

The answer is no, and the proof goes as follows: Let $f: R \to R$ be an arbitrary function. Give the positive integer n, let U_n be the union of the elements of the following collection:

$$\{U \mid U \text{ is open in } R \text{ and diameter } f(U) < 1/n\}.$$

Then U_n is open in R. Furthermore, the set

$$C = \bigcap_{n \in Z_+} U_n$$

is precisely the set of points at which the function f is continuous, as you can readily check. Thus the set of points at which f is continuous forms a G_δ set in R. The set of rationals is not a G_δ set in R; therefore, there can be no function $f: R \to R$ that is continuous precisely at the rational points.

Surprisingly enough, there does exist a function h that is continuous precisely at the *irrational* points. Choose a bijective function $f: Z_+ \to Q$; let $q_n = f(n)$. Then define $h: R \to R$ by the rule

$$h(q_n) = 1/n \quad \text{for } n \in Z_+,$$
$$h(x) = 0 \quad \text{for } x \text{ irrational.}$$

We leave it to you to check that h is the desired function.

Exercises

1. Let X equal the countable union $\bigcup B_n$. Show that if X is a Baire space and nonempty, at least one of the sets $\bar{B}_n$ has a nonempty interior.
2. The Baire theorem implies that R cannot be written as a countable union of closed subsets having empty interiors. Show this fails if the sets are not required to be closed.
3. Show that every open subset of a Baire space is a Baire space.
4. Show that every locally compact Hausdorff space is a Baire space.
5. Show that if every point x of X has a neighborhood that is a Baire space, then X is a Baire space. [*Hint:* Use the open set formulation of the Baire condition.]
6. Show that if Y is a G_δ set in X, and if X is compact Hausdorff or complete metric, then Y is a Baire space in the subspace topology. [*Hint:* Suppose that $Y = \bigcap W_n$, where W_n is open in X, and that B_n is closed in Y and has empty interior in Y. Given U_0 open in X with $U_0 \cap Y \neq \varnothing$, find a sequence of open sets U_n of X with $U_n \cap Y$ nonempty, such that

$$\bar{U}_n \subset U_{n-1},$$
$$\bar{U}_n \cap \bar{B}_n = \varnothing,$$
$$\operatorname{diam} U_n < 1/n \quad \text{in the metric case,}$$
$$\bar{U}_n \subset W_n.]$$

7. Show that the irrationals are a Baire space.
8. Show that every G_δ set in a locally compact Hausdorff space is a Baire space.

9. Check the statements made in Example 4 about C and h.

10. *Theorem (Uniform boundedness principle). Let X be a complete metric space, and let $\mathcal{F}$ be a subset of $\mathcal{C}(X, R)$ such that for each $x \in X$, the set*

$$\mathcal{F}_x = \{f(x) \mid f \in \mathcal{F}\}$$

is bounded. Then there is a nonempty open set U of X on which the functions in $\mathcal{F}$ are uniformly bounded; that is, there is such a set U and a number M such that $|f(x)| \leq M$ for all $x \in U$ and all $f \in \mathcal{F}$.

[*Hint:* Let $A_N = \{x; |f(x)| \leq N \text{ for all } f \in \mathcal{F}\}$.]

11. Determine whether or not R_l is a Baire space.

12. Show that R^J is a Baire space in the box, product, and uniform topologies.

***13.** Show that if A is any G_δ set in R, there is a function $f: R \to R$ that is continuous at all points of A and discontinuous at all other points.

***14.** Let X be a topological space; let Y be a complete metric space. Show that $\mathcal{C}(X, Y)$ is a Baire space in the fine topology (see Exercise 9 of §7-4). [*Hint:* Given basis elements $B(f_i, \delta_i)$ such that $\delta_1 \leq 1$ and $\delta_{i+1} \leq \delta_i/3$ and $f_{i+1} \in B(f_i, \delta_i/3)$, show that

$$\bigcap B(f_i, \delta_i) \neq \varnothing.]$$

7-8 A Nowhere-Differentiable Function

We prove the following result from analysis:

Theorem 8.1. *Let* $h : [0, 1] \to R$ *be a continuous function. Given* $\epsilon > 0$, *there is a function* $g : [0, 1] \to R$ *with* $|h(x) - g(x)| < \epsilon$ *for all* x, *such that* g *is continuous and nowhere-differentiable.*

Proof. Let $I = [0, 1]$. Consider the space $\mathcal{C} = \mathcal{C}(I, R)$ of continuous maps from I to R, in the metric

$$\rho(f, g) = \max \{|f(x) - g(x)|\}.$$

This space is a complete metric space, and therefore is a Baire space. We shall define, for each n, a certain subset U_n of $\mathcal{C}$ that is open in $\mathcal{C}$ and dense in $\mathcal{C}$, and has the property that the functions belonging to the intersection

$$\bigcap_{n \in \mathbb{Z}_+} U_n$$

are nowhere-differentiable. Since $\mathcal{C}$ is a Baire space, this intersection is dense in $\mathcal{C}$, by Lemma 7.1. Therefore, given h and ϵ, this intersection must contain a function g such that $\rho(h, g) < \epsilon$. The theorem follows.

The tricky part is to define the set U_n properly. We first take a function f and consider its difference quotients. Given $x \in I$ and given $0 < h \leq \frac{1}{2}$, consider the expressions

$$\left|\frac{f(x+h) - f(x)}{h}\right| \quad \text{and} \quad \left|\frac{f(x-h) - f(x)}{-h}\right|.$$

Since $h \leq \frac{1}{2}$, at least one of the numbers $x + h$ and $x - h$ belongs to I, so that at least one of these expressions is defined. Let $\Delta f(x, h)$ denote the larger of the two if both are defined; otherwise let it denote the one that is defined. If the derivative $f'(x)$ of f at x exists, it equals the limit of these difference quotients, so that

$$|f'(x)| = \lim_{h \to 0} \Delta f(x, h).$$

We seek to find a continuous function for which this limit does not exist.

This gives us the idea for defining the set U_n. Roughly speaking, U_n will consist of those functions such that $\Delta f(x, h)$ is large (bounded below by a number greater than n) for values of h which are small (less than $1/n$). More precisely, given positive numbers α and h, with $h \leq \frac{1}{2}$, we define a set $U(\alpha, h) \subset \mathcal{C}(I, R)$ by setting

$$U(\alpha, h) = \{f \mid \Delta f(x, h) \geq \alpha \text{ for every } x \in I\}.$$

Then define U_n to be the union of all those sets $U(\alpha, h)$ for which $\alpha > n$ and $h < 1/n$. Note that $U_n \supset U_{n+1} \supset \cdots$.

EXAMPLE 1. Let α be given. The function $f(x) = 4\alpha x(1 - x)$, whose graph is a parabola, belongs to $U(\alpha, h)$ for $h = \frac{1}{4}$. Geometrically speaking, what this says is that for each x, at least one of the indicated secant lines of the parabola in Figure 10 has slope of absolute value at least α. The function g pictured in Figure 10 belongs to $U(\alpha, h)$ for any $h \leq \frac{1}{4}$; and the function k belongs to $U(\alpha, h)$ for any $h \leq \frac{1}{8}$.

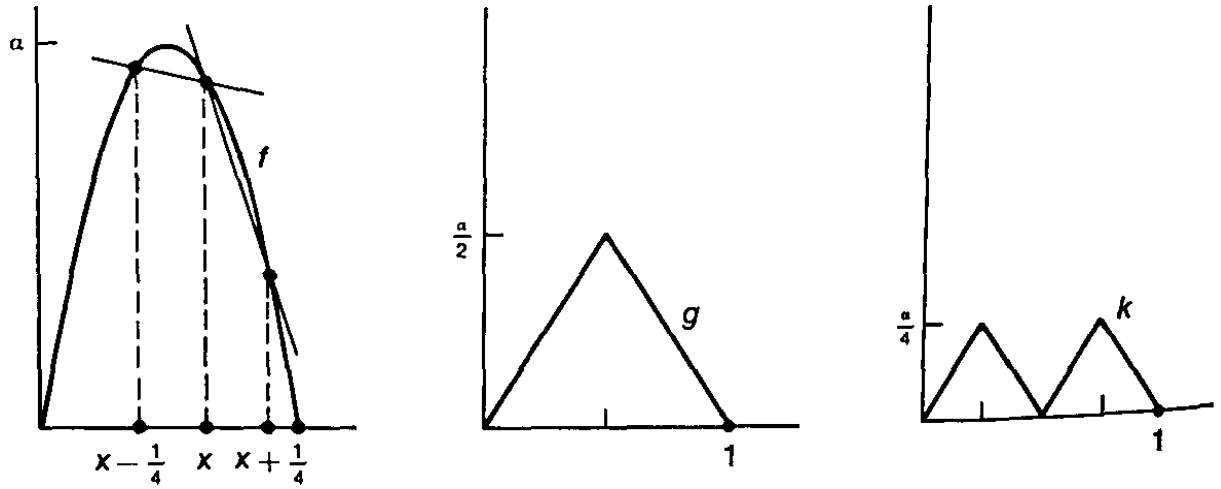


Figure 10

Now we prove the following facts about the set U_n:

(1) *U_n is open in $\mathcal{C}$.* Suppose that $f \in U_n$. Then $f \in U(\alpha, h)$ for some $\alpha > n$ and some $h < 1/n$. Let

$$\delta = \frac{h(\alpha - n)}{4}.$$

We assert that if g is any function such that $\rho(f, g) < \delta$, then g belongs to $U(\alpha', h)$, where $\alpha' = (\alpha + n)/2$. Since $\alpha' > n$, this means that g belongs to U_n. Thus the δ-neighborhood of f belongs to U_n, so U_n is open in $\mathcal{C}$.

To prove the claim, suppose that $\Delta f(x, h) = |f(x + h) - f(x)|/h$. We compute

$$\left|\frac{f(x+h)-f(x)}{h} - \frac{g(x+h)-g(x)}{h}\right|$$
$$= \left|\frac{[f(x+h)-g(x+h)]-[f(x)-g(x)]}{h}\right| \leq \frac{2\delta}{h} = \frac{\alpha - n}{2}.$$

If the first difference quotient is at least α in absolute value, then the second one is in absolute value at least $\alpha - (\alpha - n)/2 = \alpha'$. A similar computation applies if $\Delta f(x, h)$ equals the other difference quotient.

(2) *U_n is dense in $\mathcal{C}$.* We must show that given f in $\mathcal{C}$, given $\epsilon > 0$, and given n, we can find an element g of U_n within ϵ of f.

Choose $\alpha > n$. We shall construct g as a "piecewise-linear" function, that is, a function whose graph is a broken line segment; each line segment in the graph of g will have slope at least α in absolute value. It follows at once that such a function g belongs to U_n. For let

$$0 = x_0 < x_1 < x_2 < \cdots < x_k = 1$$

be a partition of the interval [0, 1] such that the restriction of g to each subinterval $I_i = [x_{i-1}, x_i]$ is a linear function. Then choose h so that

$$h < 1/n,$$
$$h \leq \tfrac{1}{2} \min \{|x_i - x_{i-1}|; i = 1, \ldots, k\}.$$

If x is in [0, 1], then x belongs to some subinterval I_i. If x belongs to the first half of the subinterval I_i, then $x + h$ belongs to I_i and $(g(x + h) - g(x))/h$ equals the slope of the linear function $g \mid I_i$. Similarly, if x belongs to the second half of I_i, then $x - h$ belongs to I_i and $(g(x - h) - g(x))/(-h)$ equals the slope of $g \mid I_i$. In either case, $\Delta g(x, h) \geq \alpha$, so $g \in U(\alpha, h) \subset U_n$, as desired.

Now given f, ϵ, and α, we must show how to construct the desired piecewise-linear function g. First, we use uniform continuity of f to choose a partition of the interval

$$0 = t_0 < t_1 < \cdots < t_m = 1$$

having the property that f varies by at most $\epsilon/4$ on each subinterval $[t_{i-1}, t_i]$ of this partition. For each $i = 1, \ldots, m$, choose a point $a_i \in (t_{i-1}, t_i)$. We then define a piecewise-linear function g_1 by the equations

$$g_1(x) = \begin{cases} f(t_{i-1}) & \text{for } x \in [t_{i-1}, a_i], \\ f(t_{i-1}) + \dfrac{f(t_i) - f(t_{i-1})}{t_i - a_i}(x - a_i) & \text{for } x \in [a_i, t_i]. \end{cases}$$

The graphs of f and g_1 are pictured in Figure 11.

We have some freedom of choice in choosing the point a_i. If $f(t_i) \neq$

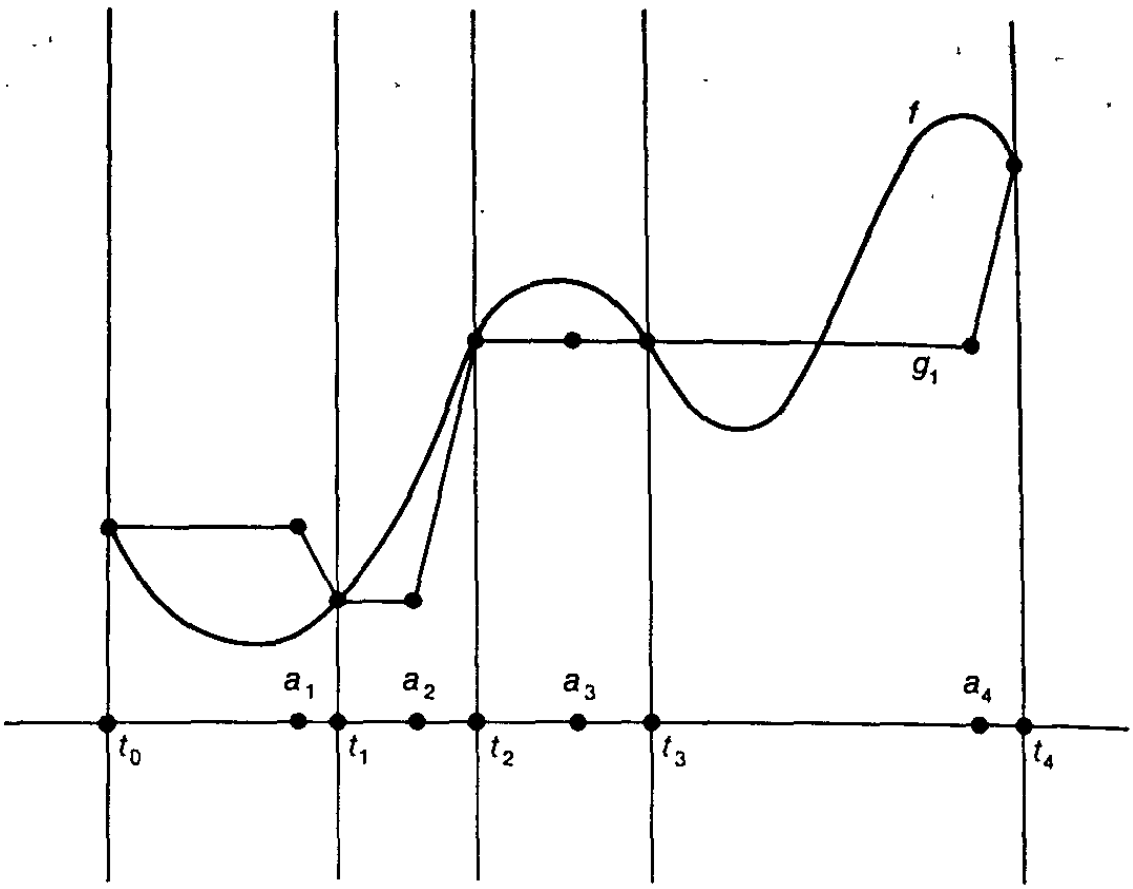


Figure 11

$f(t_{i-1})$, we require a_i to be close enough to t_i that

$$t_i - a_i < \frac{|f(t_i) - f(t_{i-1})|}{\alpha}.$$

Then the graph of g_1 will consist entirely of line segments of slope zero and line segments of slope at least α in absolute value.

Furthermore, we assert that $\rho(g_1, f) < \epsilon/2$: On the interval I_i, both $g_1(x)$ and $f(x)$ vary by at most $\epsilon/4$ from $f(t_{i-1})$; therefore, they are within $\epsilon/2$ of each other. Then $\rho(g_1, f) = \max\{|g_1(x) - f(x)|\} < \epsilon/2$.

The function g_1 is not yet the function we want. We now define a function g by replacing each horizontal line segment in the graph of g_1 by a "sawtooth" graph that lies within $\epsilon/2$ of the graph of g_1 and has the property that each edge of the sawtooth has slope at least α in absolute value. We leave this part of the construction to you; it is not hard. The result is the desired piecewise-linear function g. See Figure 12.

(3) $\bigcap U_n$ *consists of nowhere-differentiable functions.* Let $f \in \bigcap U_n$. We shall prove that given x in $[0, 1]$, the limit

$$\lim_{h \to 0} \Delta f(x, h)$$

does not exist: Given n, the fact that f belongs to U_n means that we can find a number h_n with $0 < h_n < 1/n$ such that

$$\Delta f(x, h_n) > n.$$

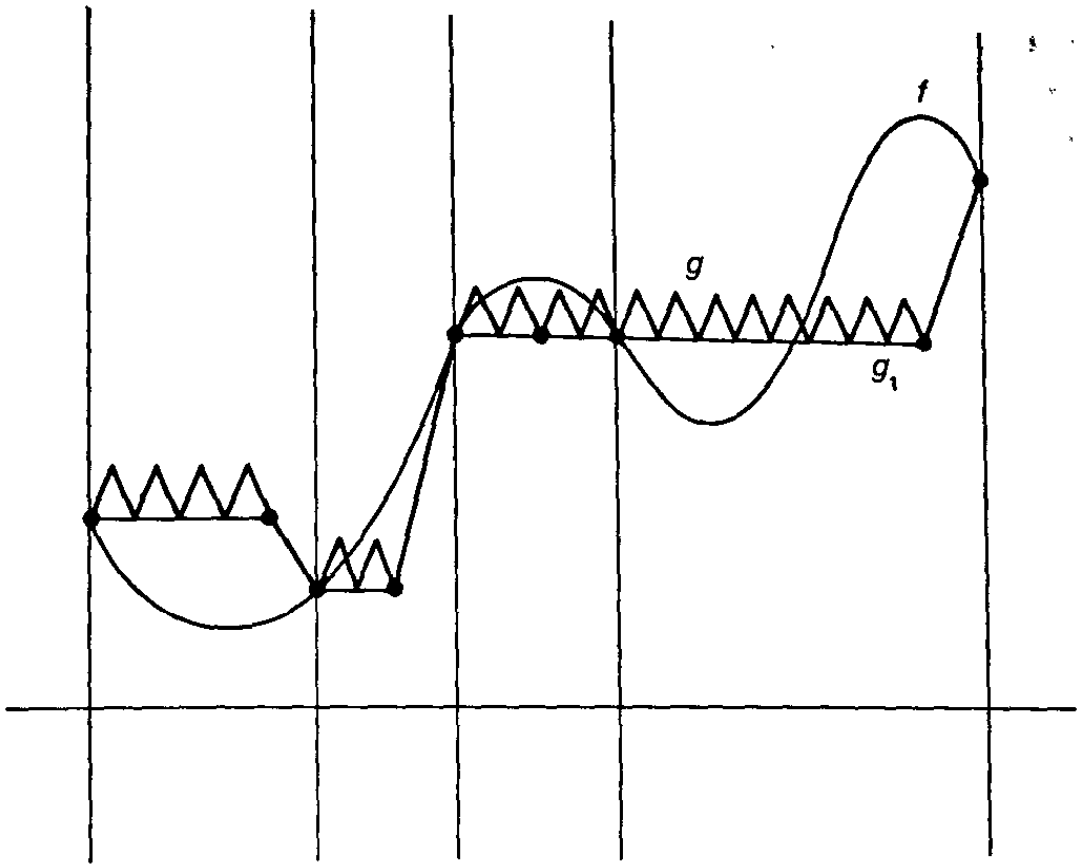


Figure 12

Then the sequence (h_n) converges to zero, but the sequence $(\Delta f(x, h_n))$ does not converge. As a result, f is not differentiable at x. □

You may find this proof frustrating, in that it seems so abstract and nonconstructive. Implicit in the proof, however, is a procedure for constructing a specific sequence f_n of piecewise-linear functions that converges uniformly to the nowhere-differentiable function f. And defining the function f in this way is just as constructive as the usual definition of the sine function, for instance, as the limit of an infinite series.

Exercises

1. Check the stated properties of the functions f, g, and k of Example 1.
2. Given n and ϵ, define a continuous function $f: I \longrightarrow R$ such that $f \in U_n$ and $|f(x)| \leq \epsilon$ for all x.
3. Complete the construction of the function g in the proof of Theorem 8.1.

7-9 An Introduction to Dimension Theory

We showed in §4-5 that if X is a compact manifold, then X can be imbedded in R^N for some positive integer N. In this section we generalize this

theorem to arbitrary compact metrizable spaces, and we also determine how large a value of N is required.

We shall define, for an arbitrary topological space X, a notion of topological dimension. It is the "covering dimension" originally defined by Lebesgue. We shall prove that each compact subset of R^m has topological dimension at most m. We shall also prove that the topological dimension of any compact m-manifold is at most m. (It is, in fact, precisely m, but this we shall not prove.)

The major theorem of this section is the theorem that any compact metrizable space of topological dimension m can be imbedded in R^N for $N = 2m + 1$. This value of N is in general the best possible. The proof is an application of the Baire theorem. It follows that every compact m-manifold can be imbedded in R^{2m+1}. It follows also that a compact metrizable space can be imbedded in R^N for some N if and only if it has finite topological dimension.

Much of what we shall do holds without requiring the space in question to be compact. But we shall restrict ourselves to that case whenever it is convenient to do so. Generalizations to the noncompact case are given in the exercises.

Definition. A collection $\mathcal{A}$ of subsets of the space X is said to have **order** $m + 1$ if some point of X lies in $m + 1$ elements of $\mathcal{A}$, and no point of X lies in more than $m + 1$ elements of $\mathcal{A}$.

Now we define what we mean by the *topological dimension* of a space X. Recall that given a collection $\mathcal{A}$ of subsets of X, a collection $\mathcal{B}$ is said to *refine* $\mathcal{A}$, or to be a *refinement* of $\mathcal{A}$, if each element B of $\mathcal{B}$ is contained in at least one element of $\mathcal{A}$.

Definition. A space X is said to be **finite-dimensional** if there is some integer m such that for every open covering $\mathcal{A}$ of X, there is an open covering $\mathcal{B}$ of X that refines $\mathcal{A}$ and has order at most $m + 1$. The **topological dimension** of X is defined to be the smallest value of m for which this statement holds.

Some basic facts about topological dimension are given in the following theorems:

Theorem 9.1. If Y *is a closed subset of* X, *and if* X *has finite dimension, then so does* Y; *and* dim Y $\leq$ dim X.

Proof. Let dim $X = m$. Let $\mathcal{A}$ be a covering of Y by sets open in Y. For each $A \in \mathcal{A}$, choose an open set A' of X such that $A' \cap Y = A$. Cover X by the open sets A', along with the open set $X - Y$. Let $\mathcal{B}$ be a refinement of this covering that is an open covering of X and has order at most $m + 1$. Then the collection

$$\{B \cap Y \mid B \in \mathcal{B}\}$$

is a covering of Y by sets open in Y, it has order at most $m + 1$, and it refines $\mathcal{A}$. □

Theorem 9.2. Let $X = Y \cup Z$, *where* Y *and* Z *are closed sets in* X *having finite topological dimension. Then*

$$\dim X = \max \{\dim Y, \dim Z\}.$$

Proof. Let $m = \max \{\dim Y, \dim Z\}$. By the preceding theorem, we know that if X is finite dimensional, $\dim X \geq m$. It remains to prove that $\dim X$ exists and is at most m.

Step 1. First we prove the following: Suppose Y is a closed subspace of X, and $\dim Y \leq m$. If $\mathcal{A}$ is an open covering of X, then there is an open covering $\mathcal{C}$ of X that refines $\mathcal{A}$, such that no point of Y lies in more than $m+1$ elements of $\mathcal{C}$. (If $\mathcal{C}$ satisfies this condition, we say that "the restriction of $\mathcal{C}$ to Y has order at most $m+1$".)

To prove this fact, consider the collection

$$\{A \cap Y \mid A \in \mathcal{A}\}.$$

It is an open covering of Y, so it has a refinement $\mathcal{B}$ that is an open covering of Y and has order at most $m + 1$. Given $B \in \mathcal{B}$, choose an open set U_B of X such that $U_B \cap Y = B$. Choose also an element A_B of $\mathcal{A}$ such that $B \subset A_B$. Let $\mathcal{C}$ be the collection consisting of all the sets $U_B \cap A_B$, for $B \in \mathcal{B}$, along with all the sets $A - Y$, for $A \in \mathcal{A}$. Then $\mathcal{C}$ is the desired open covering of X.

Step 2. Now let $\mathcal{A}$ be an open covering of X. Let $\mathcal{B}$ be an open covering of X refining $\mathcal{A}$, whose restriction to Y has order at most $m + 1$. Then let $\mathcal{C}$ be an open covering of X refining $\mathcal{B}$, whose restriction to Z has order at most $m + 1$.

We form a new covering $\mathcal{D}$ of X as follows: Define $f : \mathcal{C} \to \mathcal{B}$ by choosing for each $C \in \mathcal{C}$ an element $f(C)$ of $\mathcal{B}$ such that $C \subset f(C)$. Given $B \in \mathcal{B}$, define $D(B)$ to be the union of all those elements C of $\mathcal{C}$ for which $f(C) = B$. (Of course, $D(B)$ is empty if B is not in the image of f.) Let $\mathcal{D}$ be the collection of all the sets $D(B)$, for $B \in \mathcal{B}$.

Now $\mathcal{D}$ refines $\mathcal{B}$, because $D(B) \subset B$ for each B; therefore, $\mathcal{D}$ refines $\mathcal{A}$. Also, $\mathcal{D}$ covers X because $\mathcal{C}$ covers X and $C \subset D(f(C))$ for each $C \in \mathcal{C}$. We show that $\mathcal{D}$ has order at most $m + 1$; then our theorem will be proved.

Suppose $x \in D(B_1) \cap \cdots \cap D(B_k)$, where the sets $D(B_i)$ are distinct. We wish to prove that $k \leq m + 1$. Note that the sets $B_1, \ldots, B_k$ must be distinct because the sets $D(B_i)$ are. Because $x \in D(B_i)$, we can choose for each i, a set $C_i \in \mathcal{C}$ such that $x \in C_i$ and $f(C_i) = B_i$. The sets C_i are distinct because the sets B_i are. Furthermore,

$$x \in [C_1 \cap \cdots \cap C_k] \subset [D(B_1) \cap \cdots \cap D(B_k)] \subset [B_1 \cap \cdots \cap B_k].$$

If x happens to lie in Y, then $k \leq m + 1$ because the restriction of $\mathcal{B}$ to Y has order at most $m + 1$; and if x is in Z, then $k \leq m + 1$ because the restriction of $\mathcal{C}$ to Z has order at most $m + 1$. □

Corollary 9.3. Let $X = Y_1 \cup \cdots \cup Y_k$, *where each* Y_i *is closed in* X *and is finite-dimensional. Then*

$$\dim X = \max \{\dim Y_1, \ldots, \dim Y_k\}.$$

Obviously, the dimension of X (if it exists) is a topological invariant. In the case where X is compact and metric, the definition can be formulated in a way that involves the metric:

Lemma 9.4. Let X *be a compact metric space. Then* $\dim X \leq m$ *if and only if for every* $\epsilon > 0$ *there is a finite open covering* $\mathcal{B}$ *of* X *by sets of diameter less than* ϵ *such that* $\mathcal{B}$ *has order at most* $m + 1$.

Proof. Suppose that $\dim X \leq m$. Take a covering $\mathcal{A}$ of X by open $\epsilon/3$-balls, and choose $\mathcal{B}'$ to be a refinement of $\mathcal{A}$ that is an open covering of X and has order at most $m + 1$. Let $\mathcal{B}$ be a finite subcollection of $\mathcal{B}'$ that covers X. Then $\mathcal{B}$ has order at most $m + 1$, and the elements of $\mathcal{B}$ have diameter less than ϵ.

To prove the reverse implication, let $\mathcal{A}$ be an arbitrary open covering of X. Let $\epsilon > 0$ be a Lebesgue number for $\mathcal{A}$, and choose a finite open covering $\mathcal{B}$ of X by sets of diameter less than ϵ, such that $\mathcal{B}$ has order at most $m + 1$. Then $\mathcal{B}$ is the desired refinement of $\mathcal{A}$. □

EXAMPLE 1. *The interval* $[a, b]$ *has topological dimension* 1. Given $\epsilon > 0$, one can choose a subdivision

$$a = t_0 < t_1 < \cdots < t_n = b$$

such that $t_i - t_{i-1} < \epsilon/2$ for each i. Then the collection

$$\mathcal{B} = \{[a, t_1), (t_0, t_2), (t_1, t_3), (t_2, t_4), \ldots, (t_{n-2}, t_n), (t_{n-1}, b]\}$$

is an open covering by sets of diameter less than ϵ, and $\mathcal{B}$ has order 2. Hence $\dim [a, b] \leq 1$.

On the other hand, given $\epsilon \leq b - a$, let $\mathcal{B}$ be any open covering of $[a, b]$ by sets of diameter less than ϵ. We assert that $\mathcal{B}$ has order at least 2. Suppose that $\mathcal{B}$ has order 1; then no two elements of $\mathcal{B}$ intersect. Since the elements of $\mathcal{B}$ have diameter less than ϵ, the collection $\mathcal{B}$ must contain more than one element; let U be one element of $\mathcal{B}$, and let V be the union of all the others. Then U and V form a separation of $[a, b]$, contrary to the fact that this interval is connected.

EXAMPLE 2. A (finite) **linear graph** G is a space that is homeomorphic to the union of finitely many line segments, each pair of which intersect in at most a common end point. The end points of the line segments are called the vertices of the linear graph. Two examples of linear graphs are sketched in Figure 13. The first is a diagram of the familiar "gas–water–electricity problem"; the second is called the "complete graph on five vertices." Neither of them can be

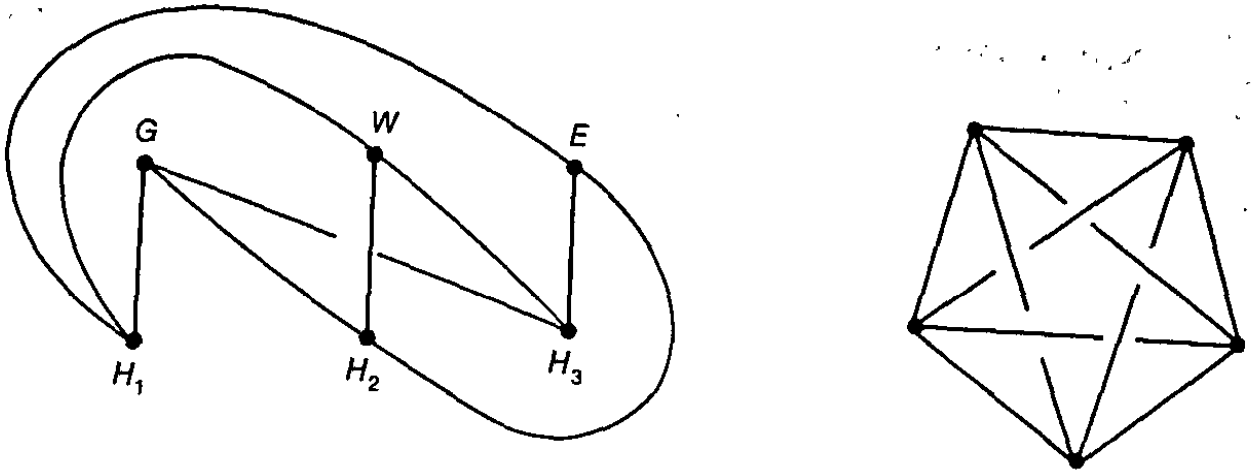


Figure 13

imbedded in the plane. [This is a highly nontrivial fact to prove; it depends on the Jordan curve theorem. See the exercises of §8-13.]

Any linear graph has topological dimension 1. There are of course many spaces that are not linear graphs which also have dimension 1 (see Exercise 3).

EXAMPLE 3. *Any compact subset C of R^2 has topological dimension at most* 2.

To prove this fact, we construct a certain open covering $\mathcal{A}$ of R^2 that has order 3. We begin by defining $\mathcal{A}_2$ to be the collection of all open unit squares in R_2 of the following form:

$$\mathcal{A}_2 = \{(n, n+1) \times (m, m+1) \mid n, m \text{ integers}\}.$$

Note that the elements of $\mathcal{A}_2$ are disjoint. Then we define a collection $\mathcal{A}_1$ by taking each (open) edge e of one of these squares,

$$e = \{n\} \times (m, m+1) \qquad \text{or} \qquad e = (n, n+1) \times \{m\},$$

and expanding it slightly to an open set U_e of R^2, being careful to ensure that if $e \neq e'$, the sets U_e and $U_{e'}$ are disjoint. We also choose each U_e so that its diameter is at most 2. Finally, we define $\mathcal{A}_0$ to be the collection consisting of all open balls of radius $\frac{1}{2}$ about the points $n \times m$. See Figure 14.

The collection of open sets $\mathcal{A} = \mathcal{A}_2 \cup \mathcal{A}_1 \cup \mathcal{A}_0$ covers R^2. Each of its elements has diameter at most 2. And it has order 3, since no point of R^2 can lie in more than one set from each $\mathcal{A}_i$.

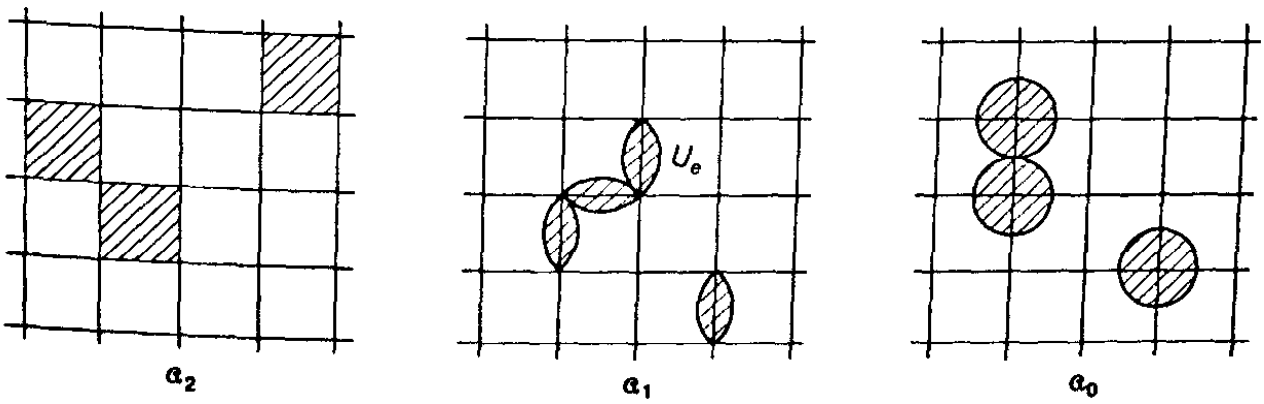


Figure 14

Now we prove that any compact set C in R^2 has topological dimension at most 2. Given $\epsilon > 0$, consider the homeomorphism $f: R^2 \to R^2$ defined by $f(\mathbf{x}) = (\frac{1}{3}\epsilon)\mathbf{x}$. Take the images under f of the open sets of the collection $\mathcal{A}$. They will form an open covering of R^2 by sets of diameter less than ϵ. Choose finitely many of them that cover C; this will be the desired open covering of C.

Of course, not every compact subset of R^2 has dimension exactly 2. A one point set has dimension 0, for instance, and a closed interval on the x-axis has dimension 1. It is, in fact, surprisingly nontrivial to prove that *some* subsets of R^2 do have topological dimension 2; the techniques of algebraic topology are required. See Example 6 below.

EXAMPLE 4. *Every compact 2-manifold X has topological dimension at most* 2. For X can be written as a finite union of subspaces that are homeomorphic to the closed unit ball in R^2; then Corollary 9.3 applies. The 2-sphere, the torus, and the Klein bottle are typical examples of 2-manifolds (see Figure 27 of §8-10).

EXAMPLE 5. A **finite simplicial 2-complex** (called a **2-complex** for short) is a space X that is homeomorphic to a finite union of closed triangles and line segments such that any two of them intersect in at most a common vertex or (in the case of two triangles) a common edge. By Corollary 9.3, every 2-complex has topological dimension at most 2. Pictured in Figure 15 is a 2-complex that is not a manifold.

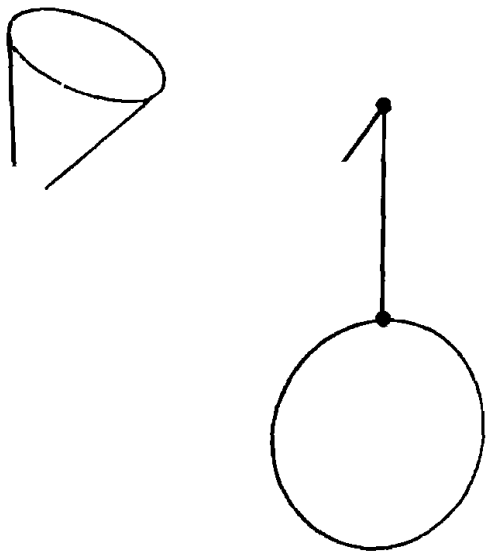


Figure 15

Properly speaking, a 2-complex must contain at least one triangle; otherwise it would be a linear graph.

EXAMPLE 6. Assuming a theorem we shall prove in the next chapter (in §8-8), we now show that any closed triangular region T in R^2 has topological dimension 2. It follows that *every compact 2-manifold, and every 2-complex, has topological dimension precisely* 2.

The theorem in question is the following:

Let Bd T *denote the union of the edges of the closed triangular region T. Then there is no continuous map $f: T \to$ Bd T that carries each edge of T into itself.*

We know that T has dimension at most 2. To prove T has dimension at least 2, we find an $\epsilon > 0$ such that any open covering of T by sets of diameter less than ϵ has order at least 3.

Choose $\epsilon > 0$ small enough that any subset of R^2 which intersects all three edges of T has diameter at least ϵ. Let

$$\mathcal{A} = \{U_1, \ldots, U_n\}$$

be an open covering of T by sets of diameter less than ϵ. Suppose $\mathcal{A}$ has order less than 3. For each $i = 1, \ldots, n$, choose a vertex v_i of T as follows: If U_i intersects two edges of T, let v_i denote the vertex common to these edges. If U_i intersects only one edge of T, let v_i be one of the vertices of this edge. If U_i intersects no edge of T, let v_i be any vertex of T.

Now let $\{\phi_i\}$ be a partition of unity dominated by $\{U_1, \ldots, U_n\}$. (See §4-5.) Define $g : T \to R^2$ by the equation

$$g(x) = \sum_{i=1}^{n} \phi_i(x) v_i.$$

Then g is continuous. Now each point of T lies in at most two elements of $\mathcal{A}$. Therefore, for each $x \in T$, at most two of the functions $\phi_i(x)$ are nonzero. Then

$$g(x) = v_i$$

if x lies in only one open set U_i, and

$$g(x) = tv_i + (1 - t)v_j$$

for some t with $0 \leq t \leq 1$ if x lies in two open sets U_i and U_j. [Here we use the fact that $\sum \phi_i(x) = 1$.] Thus g maps all of T into the union of the edges of T.

We claim further that g maps each edge of T into itself. Suppose that x belongs to the edge vw of T. Then any open set U_i containing x necessarily intersects this edge, so that the vertex v_i must be either v or w. Likewise, if $x \in U_j$, we have $v_j = v$ or w. The above equations then show that $g(x)$ belongs to vw.

The existence of the map g violates the above-quoted theorem.

Now we prove the imbedding theorem mentioned earlier, to the effect that every compact metrizable space of topological dimension m can be imbedded in R^{2m+1}. This theorem is another "deep" theorem; it is not at all obvious, for instance, why $2m + 1$ should be the crucial dimension. We shall discuss some special cases of the theorem before turning to the proof; this discussion should illuminate why the dimension $2m + 1$ occurs.

First let us consider the simplest possible case, that of a finite linear graph G. In this case, the imbedding theorem states that every finite linear graph can be imbedded in R^3. It is not hard to give an *ad hoc* argument that will work here. We shall give instead an argument that will generalize to higher dimensions, an argument that involves the notion of *general position*.

A set A of points of R^3 is said to be in **general position** in R^3 if no two of the points are equal, no three are collinear, and no four are coplanar. It is easy to find sets of points in general position in R^3. For instance, the set consisting of all points of the curve

$$\{(t, t^2, t^3) \mid t \in R\}$$

is in general position in R^3, as you can prove.

Now, given a finite linear graph G with vertices $v_1, \ldots, v_n$, let us choose a set $\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$ of points in general position in R^3. We define a continuous map $f : G \longrightarrow R^3$ by letting f map v_i to $\mathbf{z}_i$, and letting f map the point $tv_i + (1 - t)v_j$ of the line segment in G joining v_i and v_j (if any) to the point $t\mathbf{z}_i + (1 - t)\mathbf{z}_j$ of the line segment in R^3 joining $\mathbf{z}_i$ and $\mathbf{z}_j$. It is clear that f is continuous. Also, f is injective: Let $e = v_iv_j$ and $e' = v_kv_l$ be two line segments of G. If e and e' have no vertex in common, then $f(e)$ and $f(e')$ are disjoint, for otherwise the points $\mathbf{z}_i$, $\mathbf{z}_j$, $\mathbf{z}_k$, and $\mathbf{z}_l$ would be coplanar. And if e and e' have a vertex in common, say $i = k$, then $f(e)$ and $f(e')$ can intersect only in the point $\mathbf{z}_i$; for otherwise $\mathbf{z}_i$, $\mathbf{z}_j$, and $\mathbf{z}_l$ would be collinear.

Now let us consider the next simplest case, that of a 2-complex. In order to generalize the preceding argument, we must first study a bit of the analytic geometry of R^N. We shall need these results later anyway when we prove the general imbedding theorem.

Analytic geometry in R^N is really nothing more than the usual linear algebra of vector spaces translated into somewhat different language. First we make the following definition:

Definition. A set $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$ of points of R^N is said to be **geometrically independent** if

$$[\textstyle\sum_{i=0}^{k} a_i\mathbf{x}_i = \mathbf{0} \quad \text{with } \sum_{i=0}^{k} a_i = 0] \Rightarrow [\text{each } a_i = 0].$$

It is clear that any set consisting of only one point is geometrically independent. What does geometric independence mean if $k > 0$? It is simple algebra to show that the set of points $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$ is geometrically independent if and only if the set of vectors

$$\{\mathbf{x}_1 - \mathbf{x}_0, \mathbf{x}_2 - \mathbf{x}_0, \ldots, \mathbf{x}_k - \mathbf{x}_0\}$$

is linearly independent in the sense of ordinary linear algebra. This gives us something to visualize: Any two distinct points form a geometrically independent set. Three points form a geometrically independent set if they are not collinear. Four points in R^3 form a geometrically independent set if they are not coplanar. And so on.

It follows from these remarks that the points

$$\mathbf{0} = (0, 0, \ldots, 0),$$
$$\boldsymbol{\epsilon}_1 = (1, 0, \ldots, 0),$$
$$\ldots$$
$$\boldsymbol{\epsilon}_N = (0, 0, \ldots, 1)$$

are geometrically independent in R^N. It also follows that R^N contains no more than $N + 1$ geometrically independent points.

Definition. Let $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$ be a set of points of R^N that is geometrically independent. The **plane P determined by these points** is defined to be the set of all points $\mathbf{x}$ of R^N such that

$$\mathbf{x} = \sum_{i=0}^{k} t_i \mathbf{x}_i, \quad \text{where } \sum_{i=0}^{k} t_i = 1.$$

It is simple algebra to check that P can also be expressed as the set of all points $\mathbf{x}$ such that

$$(*) \qquad \mathbf{x} = \mathbf{x}_0 + \sum_{i=1}^{k} a_i(\mathbf{x}_i - \mathbf{x}_0)$$

for some scalars $a_1, \ldots, a_k$. Thus P can be described not only as "the plane determined by the points $\mathbf{x}_0, \ldots, \mathbf{x}_k$," but also as "the plane passing through the point $\mathbf{x}_0$ parallel to the vectors $\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_k - \mathbf{x}_0$."

Consider now the homeomorphism $T : R^N \to R^N$ defined by the equation $T(\mathbf{x}) = \mathbf{x} - \mathbf{x}_0$. It is called a **translation** of R^N. Expression (*) shows that this map carries the plane P onto the vector subspace V^k of R^N having as basis the vectors $\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_k - \mathbf{x}_0$. For this reason, we often call P a ***k*-plane** in R^N.

Two facts follow at once: First, if $k < N$, the k-plane P necessarily has empty interior in R^N (because V^k does). And second, if $\mathbf{y}$ is any point of R^N not lying in P, then the set

$$\{\mathbf{x}_0, \ldots, \mathbf{x}_k, \mathbf{y}\}$$

is geometrically independent. For if $\mathbf{y} \notin P$, then $T(\mathbf{y}) = \mathbf{y} - \mathbf{x}_0$ is not in V^k. By a standard theorem of linear algebra, the vectors $\{\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_k - \mathbf{x}_0, \mathbf{y} - \mathbf{x}_0\}$ are linearly independent, from which our result follows.

Definition. A set A of points of R^N is said to be in **general position in R^N** if every subset of A containing $N + 1$ or fewer points is geometrically independent.

In the case of R^3, this is the same as the definition given earlier, as you can check.

Lemma 9.5. *Given a finite set $\{\mathbf{x}_1, \ldots, \mathbf{x}_n\}$ of points of R^N and given $\delta > 0$, there exists a set $\{\mathbf{y}_1, \ldots, \mathbf{y}_n\}$ of points of R^N in general position in R^N, such that $|\mathbf{x}_i - \mathbf{y}_i| < \delta$ for all* i.

Proof. We proceed by induction. Set $\mathbf{y}_1 = \mathbf{x}_1$. Suppose that we are given $\mathbf{y}_1, \ldots, \mathbf{y}_p$ in general position in R^N. Consider the set of all planes in R^N determined by subsets of $\{\mathbf{y}_1, \ldots, \mathbf{y}_p\}$ that contain N or fewer elements. Every such subset is geometrically independent and determines a k-plane of R^N for some $k \leq N - 1$. Each of these planes has empty interior in R^N. Since there are only finitely many of them, their union also has empty interior in R^N. (Recall that R^N is a Baire space.) Choose $\mathbf{y}_{p+1}$ to be a point of R^N within δ of $\mathbf{x}_{p+1}$ that does not lie in any of these planes. It follows at once that the set

$$C = \{\mathbf{y}_1, \ldots, \mathbf{y}_p, \mathbf{y}_{p+1}\}$$

is in general position in R^N. For let D be any subset of C containing $N + 1$ or fewer elements. If D does not contain $\mathbf{y}_{p+1}$, then D is geometrically independent by the induction hypothesis. If D does contain $\mathbf{y}_{p+1}$, then $D - \{\mathbf{y}_{p+1}\}$ contains N or fewer points and $\mathbf{y}_{p+1}$ is not in the plane determined by these points, by construction. Then as noted above, D is geometrically independent. □

Now we can show how to imbed any 2-complex X in R^5. Given the 2-complex X with vertices $v_1, \ldots, v_n$, let us choose a set $\mathbf{z}_1, \ldots, \mathbf{z}_n$ of points that are in general position in R^5. We define a continuous map $f : X \to R^5$ as follows: The general point of the triangle with vertices v_i, v_j, and v_k can be written in the form

$$sv_i + tv_j + (1 - s - t)v_k,$$

where s, t, and $1 - s - t$ are all nonnegative. (You can convince yourself of this fact.) We let f map this point to the point $s\mathbf{z}_i + t\mathbf{z}_j + (1 - s - t)\mathbf{z}_k$ of the triangle in R^5 having $\mathbf{z}_i$, $\mathbf{z}_j$, and $\mathbf{z}_k$ as vertices. If v_iv_j is a line segment that is not the edge of a triangle, we let f map it onto the line segment $\mathbf{z}_i\mathbf{z}_j$, as before.

Because the points $\mathbf{z}_i$ are in general position, the map f will be injective. Several cases need to be checked. For instance, if v_1, v_2, v_3 and v_4, v_5, v_6 are the vertices, respectively, of two disjoint triangles σ and σ' of X, then $f(\sigma)$ and $f(\sigma')$ cannot intersect, for that would imply that the points $\mathbf{z}_i$, for $i = 1, \ldots, 6$, are not independent. [If $\mathbf{x}_0$ were a point of the intersection, we would have

$$\textstyle\sum_{i=1}^{3} s_i\mathbf{z}_i = \mathbf{x}_0 = \sum_{i=4}^{6} t_i\mathbf{z}_i$$

where $\sum s_i = \sum t_i = 1$, or

$$\textstyle\sum_{i=1}^{3} s_i\mathbf{z}_i + \sum_{i=4}^{6} (-t_i)\mathbf{z}_i = \mathbf{0},$$

where the sum of the coefficients is zero.] Similar arguments apply in the other cases.

With these special cases as motivation, let us now prove the general imbedding theorem.

Theorem 9.6 (The imbedding theorem). *Every compact metrizable space* X *of topological dimension* m *can be imbedded in* R^{2m+1}.

Proof. Let $N = 2m + 1$. Let us denote the square metric for R^N by

$$|\mathbf{x} - \mathbf{y}| = \max \{|x_i - y_i|; i = 1, \ldots, N\}.$$

Then we can use ρ to denote the corresponding sup metric on the space $\mathcal{C}(X, R^N)$;

$$\rho(f, g) = \max \{|f(x) - g(x)|; x \in X\}.$$

The space $\mathcal{C}(X, R^N)$ is complete in the metric ρ, since R^N is complete in the square metric.

Choose a metric d for the space X; because X is compact, d is bounded. Given a continuous map $f: X \to R^N$, let us define

$$\Delta(f) = \text{lub}\,\{\text{diam } f^{-1}(\{z\}) \mid z \in R^N\}.$$

The number $\Delta(f)$ measures how far f "deviates" from being injective; if $\Delta(f) = 0$, each set $f^{-1}(\{z\})$ consists of at most one point, so f is injective.

Now, given $\epsilon > 0$, define U_ϵ to be the set of all those continuous maps $f: X \to R^N$ for which $\Delta(f) < \epsilon$; it consists of all those maps that "deviate" from being injective by less than ϵ. We shall show that U_ϵ is both open and dense in $\mathcal{C}(X, R^N)$. It follows that the intersection

$$\bigcap_{n \in Z_+} U_{1/n}$$

is dense in $\mathcal{C}(X, R^N)$ and is in particular nonempty.

If f is an element of this intersection, then $\Delta(f) < 1/n$ for every n. Hence $\Delta(f) = 0$ and f is injective. Because X is compact, f is an imbedding. Thus the imbedding theorem is proved.

(1) *U_ϵ is open in $\mathcal{C}(X, R^N)$.* Given an element f of U_ϵ, we wish to find some ball $B_\rho(f, \delta)$ about f that is contained in U_ϵ. First choose a number b such that $\Delta(f) < b < \epsilon$. Note that if $f(x) = f(y) = z$, then x and y belong to the set $f^{-1}(\{z\})$, so that $d(x, y)$ must be less than b. It follows that if we let A be the following subset of $X \times X$,

$$A = \{x \times y \mid d(x, y) \geq b\},$$

then the function $|f(x) - f(y)|$ is positive on A. Now A is closed in $X \times X$ and therefore compact; hence the function $|f(x) - f(y)|$ has a positive minimum on A. Let

$$\delta = \tfrac{1}{2} \min\{|f(x) - f(y)|;\ x \times y \in A\}.$$

We assert that this value of δ will suffice.

Suppose that g is a map such that $\rho(f, g) < \delta$. If $x \times y \in A$, then $|f(x) - f(y)| \geq 2\delta$ by definition; since $g(x)$ and $g(y)$ are within δ of $f(x)$ and $f(y)$, respectively, we must have $|g(x) - g(y)| > 0$. Hence the function $|g(x) - g(y)|$ is positive on A. As a result, if x and y are two points such that $g(x) = g(y)$, then necessarily $d(x, y) < b$. We conclude that $\Delta g \leq b < \epsilon$, as desired.

(2) *U_ϵ is dense in $\mathcal{C}(X, R^N)$.* This is the hard part of the proof. We need to use the analytic geometry of R^N discussed above. Let $f \in \mathcal{C}(X, R^N)$. Given $\epsilon > 0$ and given $\delta > 0$, we wish to find a function $g \in \mathcal{C}(X, R^N)$ such that $g \in U_\epsilon$ and $\rho(f, g) < \delta$.

Let us cover X by finitely many open sets $\{U_1, \ldots, U_n\}$ such that

(1) diam $U_i < \epsilon/2$ in X.

(2) diam $f(U_i) < \delta/2$ in R^N.

(3) $\{U_1, \ldots, U_n\}$ has order $\leq m + 1$.

(A Lebesgue number argument is involved here.) Let $\{\phi_i\}$ be a partition of unity dominated by $\{U_i\}$ (see §4-5). For each i, choose a point $x_i \in U_i$. Then choose, for each i, a point $\mathbf{z}_i \in R^N$ such that $\mathbf{z}_i$ is within $\delta/2$ of the point $f(x_i)$, and such that the set

$$\{\mathbf{z}_1, \ldots, \mathbf{z}_n\}$$

is in general position in R^N. Finally, define $g : X \longrightarrow R^N$ by the equation

$$g(x) = \sum_{i=1}^{n} \phi_i(x)\mathbf{z}_i.$$

We assert that g is the desired function.

First, we show that $\rho(f, g) < \delta$. Note that

$$g(x) - f(x) = \sum_{i=1}^{n} \phi_i(x)\mathbf{z}_i - \sum_{i=1}^{n} \phi_i(x)\, f(x);$$

here we use the fact that $\sum \phi_i(x) = 1$. Then

$$g(x) - f(x) = \sum \phi_i(x)(\mathbf{z}_i - f(x_i)) + \sum \phi_i(x)(f(x_i) - f(x)).$$

Now $|\mathbf{z}_i - f(x_i)| < \delta/2$ for each i, by choice of the points $\mathbf{z}_i$. And if i is an index such that $\phi_i(x) \neq 0$, then $x \in U_i$; because diam $f(U_i) < \delta/2$, it follows that $|f(x_i) - f(x)| < \delta/2$. Since $\sum \phi_i(x) = 1$, we conclude that $|g(x) - f(x)| < \delta$. Therefore, $\rho(g, f) < \delta$, as desired.

Second, we show that $g \in U_\epsilon$. We shall prove that if $x, y \in X$ and $g(x) = g(y)$, then x and y belong to one of the open sets U_i, so that necessarily $d(x, y) < \epsilon/2$ (since diam $U_i < \epsilon/2$). As a result, $\Delta(g) \leq \epsilon/2 < \epsilon$, as desired.

So suppose $g(x) = g(y)$. Then

$$\sum_{i=1}^{n} [\phi_i(x) - \phi_i(y)]\mathbf{z}_i = \mathbf{0}.$$

Because the covering $\{U_i\}$ has order at most $m + 1$, at most $m + 1$ of the numbers $\phi_i(x)$ are nonzero, and at most $m + 1$ of the numbers $\phi_i(y)$ are nonzero. Thus the sum $\sum [\phi_i(x) - \phi_i(y)]\mathbf{z}_i$ has at most $2m + 2$ nonzero terms. Note that the sum of the coefficients vanishes since

$$\sum [\phi_i(x) - \phi_i(y)] = 1 - 1 = 0.$$

The points $\mathbf{z}_i$ are in general position in R^N, so that any subset of them having $N + 1$ or fewer elements is geometrically independent. And by hypothesis $N + 1 = 2m + 2$. (Aha!) Therefore, we conclude that

$$\phi_i(x) - \phi_i(y) = 0$$

for all i.

Now $\phi_i(x) > 0$ for some i, so that $x \in U_i$. Since $\phi_i(y) = \phi_i(x)$, we have $y \in U_i$ also, as asserted. □

To give some content to the imbedding theorem, we need some more examples of spaces that are finite dimensional. We prove the following theorem:

Theorem 9.7. Every compact subset of R^{N} *has topological dimension at most* N.

Proof. The proof is a generalization of the proof given in Example 3 above for R^2. Let ρ be the square metric on R^N.

Step 1. We begin by breaking R^N up into "unit cubes." Define $\mathcal{J}$ to be the following collection of open intervals in R:

$$\mathcal{J} = \{(n, n+1) \mid n \in Z\},$$

and define $\mathcal{K}$ to be the following collection of one-point sets in R:

$$\mathcal{K} = \{\{n\} \mid n \in Z\}.$$

If M is an integer such that $0 \leq M \leq N$, let $\mathcal{C}_M$ denote the set of all products

$$C = A_1 \times A_2 \times \cdots \times A_N,$$

where exactly M of the sets A_i belong to $\mathcal{J}$, and the remainder belong to $\mathcal{K}$. If $M > 0$, then C is homeomorphic to the product $(0, 1)^M$, and will be called an **M-cube**. If $M = 0$, then C consists of a single point, and will be called a **0-cube**.

Let $\mathcal{C} = \mathcal{C}_0 \cup \mathcal{C}_1 \cup \cdots \cup \mathcal{C}_N$. Note that each point $\mathbf{x}$ of R^N lies in precisely one element of $\mathcal{C}$, because each real number x_i lies in precisely one element of $\mathcal{J} \cup \mathcal{K}$. We shall expand each element C of $\mathcal{C}$ slightly to an open set $U(C)$ of R^N of diameter at most 2, in such a way that if C and C' are two different M-cubes, then $U(C)$ and $U(C')$ are disjoint.

First, note that if C and C' are two different M-cubes, then $\bar{C} \cap C' = \varnothing$. This is easily checked; let $C = A_1 \times \cdots \times A_N$ and $C' = A_1' \times \cdots \times A_N'$. Choose an index i so that $A_i \neq A_i'$. If $\bar{A}_i$ and A_i' are disjoint, then so are $\bar{C}$ and C'. Otherwise, A_i must be an interval $(n, n+1)$ and A_i' must equal $\{n\}$ or $\{n+1\}$. In that case, since both C and C' are M-cubes, there must be another index j for which A_j is a one-point set $\{k\}$ and A_j' is an interval $(l, l+1)$. Then $\bar{A}_j$ and A_j' are disjoint, so that $\bar{C}$ and C' are disjoint.

Second, note that the collection $\mathcal{C}_M$ of all M-cubes is locally finite. Indeed, if a is a point with integer coordinates, then the open set $B_\rho(\mathbf{a}, 1)$ intersects only 3^N different cubes of *all* dimensions.

Let $\mathbf{x} \in C$, where C is an M-cube. There is a neighborhood of $\mathbf{x}$ intersecting only finitely many M-cubes C' different from C. Since $C \cap \bar{C}' = \varnothing$ for each such C', we can choose $\epsilon(\mathbf{x})$ so that the $\epsilon(\mathbf{x})$-neighborhood of $\mathbf{x}$ does not intersect *any* M-cube C' different from C. Also, let $\epsilon(\mathbf{x}) < 1$. Then we define

$$U(C) = \bigcup_{\mathbf{x} \in C} B_\rho(\mathbf{x}, \tfrac{1}{2}\epsilon(\mathbf{x})).$$

It is straightforward to check that if C and C' are different M-cubes, $U(C)$ and $U(C')$ are disjoint.

Step 2. Given M with $0 \leq M \leq N$, define $\mathcal{A}_M$ to be the collection of all sets $U(C)$, where $C \in \mathcal{C}_M$. The elements of $\mathcal{A}_M$ are disjoint, and each has

diameter at most 2. (For C has diameter 1, and each point of $U(C)$ lies within $\frac{1}{2}\epsilon(\mathbf{x}) < \frac{1}{2}$ of a point $\mathbf{x}$ of C.)

The remainder of the proof is a copy of the proof given in Example 3 for R^2. □

Corollary 9.8. Every compact m-manifold has topological dimension at most m.

Corollary 9.9. Every compact m-manifold can be imbedded in R^{2m+1}.

Corollary 9.10. Let X *be a compact metrizable space. Then* X *can be imbedded in some euclidean space* R^N *if and only if* X *has finite topological dimension.*

As mentioned earlier, much of what we have proved holds without assumption of compactness. We ask you to prove the appropriate generalizations in the exercises that follow.

One thing we do *not* ask you to prove is the fact that the topological dimension of an m-manifold is precisely m. And for good reason; the proof requires the tools of algebraic topology, which we have not studied.

Nor do we ask you to prove that $N = 2m + 1$ is the smallest value of N such that every compact metrizable space of topological dimension m can be imbedded in R^N. The reason is the same. Even in the case of a linear graph, where $m = 1$, the proof is nontrivial, as we remarked in Example 2 above.

For further results in dimension theory, the reader is referred to the classical book of Hurewicz and Wallman [H-W].

Exercises

1. Show that the Cantor set has dimension 0.
2. Show that any connected Hausdorff space having more than one point has dimension at least 1.
3. Show that the comb space and the topologist's sine curve have dimension 1.
4. Show that the subspace

$$([0, 1] \times 0) \cup \{x \times y \mid y = x \sin (1/x) \quad \text{and} \quad 0 < x \leq 1\}$$

of R^2 has dimension 1.

5. Show that the points $\mathbf{0}, \epsilon_1, \epsilon_2, \epsilon_3$, and $(1, 1, 1)$ are in general position in R^3. Sketch the corresponding imbedding into R^3 of the complete graph on five vertices.
6. (a) Verify that $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$ is geometrically independent if and only if the vectors $\mathbf{x}_1 - \mathbf{x}_0, \ldots, \mathbf{x}_k - \mathbf{x}_0$ are linearly independent.

(b) Verify that the plane P determined by $\{\mathbf{x}_0, \ldots, \mathbf{x}_k\}$ is the set of all $\mathbf{x}$ such that $\mathbf{x} = \mathbf{x}_0 + \sum a_i(\mathbf{x}_i - \mathbf{x}_0)$.

(c) Prove that a k-dimensional subspace V^k of R^N has empty interior in R^N if $k < N$.

7. Complete the proof that every 2-complex can be imbedded in R^5.

8. Show that the points of the curve $x = t$, $y = t^2$, $z = t^3$ are in general position in R^3. Generalize to R^n. [*Hint:* Use the Vandermonde determinant.]

***9.** (a) Prove the following:

Theorem. Let X be a locally compact Hausdorff space having a countable basis. Suppose that every point of X has a neighborhood whose closure has topological dimension at most m. Then X has topological dimension at most m.

[*Hint:* Write $X = \bigcup C_n$, where C_n is compact and $C_n \subset \operatorname{Int} C_{n+1}$. Generalize the proof of Theorem 9.2 to show that $\dim X \leq m$.]

(b) Prove:

Corollary. Every m-manifold has topological dimension at most m.

(c) Prove:

Corollary. Every closed subset of R^N has topological dimension at most N.

***10.** *Theorem. Let X be a locally compact Hausdorff space with a countable basis, having topological dimension m. Then X can be imbedded as a closed subset of R^{2m+1}.*

Proof.

(a) Let $N = 2m + 1$. Let d be a metric on X. Given a compact subset C of X and given $\epsilon > 0$, define

$$U_\epsilon(C) = \{f \mid f \in \mathcal{C}(X, R^N) \quad \text{and} \quad \Delta(f \mid C) < \epsilon\}.$$

Show $U_\epsilon(C)$ is open and dense in $(\mathcal{C}(X, R^N), \bar{\rho})$, where $\bar{\rho}$ is the uniform metric. [*Hint:* Use the Tietze extension theorem.]

(b) Show that given $f \in \mathcal{C}(X, R^N)$ and given $\delta > 0$, there is a continuous injective map $g : X \to R^N$ such that $\bar{\rho}(f, g) < \delta$. [*Hint:* Write $X = \bigcup C_n$, where each C_n is compact and $C_n \subset C_{n+1}$ for each n. Consider $\bigcap U_{1/n}(C_n)$.]

(c) Let $f: X \to R^N$. The map f is said to be proper if for each compact set D in R^N, the set $f^{-1}(D)$ is compact. Let X^* and $(R^N)^*$ be the one-point compactifications of these two spaces; extend f to a map $F: X^* \to (R^N)^*$ by letting $F(\infty) = \infty$. Show that f is proper and continuous if and only if F is continuous.

(d) Show that if f is proper, g is continuous, and $\bar{\rho}(f, g) < 1$, then g is proper.

(e) Construct a proper continuous map $f: X \to R$; complete the proof of the theorem.

11. Assume the two preceding exercises.

(a) Prove:

Theorem. Every m-manifold can be imbedded in R^{2m+1} as a closed subset.

(b) Prove:

Theorem. A space X can be imbedded as a closed subset in R^N for some N if and only if X is locally compact and Hausdorff with a countable basis, and has finite topological dimension.

8. The Fundamental Group and Covering Spaces

One of the basic problems of topology is to determine whether two given topological spaces are homeomorphic or not. There is no method for solving this problem in general, but there do exist techniques that apply in particular cases.

Showing that two spaces *are* homeomorphic is a matter of constructing a continuous mapping from one to the other having a continuous inverse, and constructing continuous functions is a problem that we have developed techniques to handle.

Showing that two spaces are *not* homeomorphic is a different matter. For that one must show that a continuous function with continuous inverse does *not* exist. If one can find some topological property that holds for one space but not for the other, then the problem is solved—the spaces cannot be homeomorphic. The closed interval [0, 1] cannot be homeomorphic to the open interval (0, 1), for instance, because the first space is compact and the second one is not. And the real line R cannot be homeomorphic to the "long line" L, because R has a countable basis and L does not. Nor can the real line R be homeomorphic to the plane R^2; deleting a point from R^2 leaves a connected space remaining, and deleting a point from R does not.

But the topological properties we have studied up to now do not carry us

very far in solving the problem. For instance, how does one show that the plane R^2 is not homeomorphic to three-dimensional space R^3? As one goes down the list of topological properties—compactness, connectedness, local connectedness, metrizability, and so on—one can find no topological property that distinguishes between them. As another example, consider such surfaces as the 2-sphere S^2, the torus T (surface of a doughnut), and the double torus T_2 (surface of a two-holed doughnut). None of the topological properties we have studied up to now will distinguish between them.

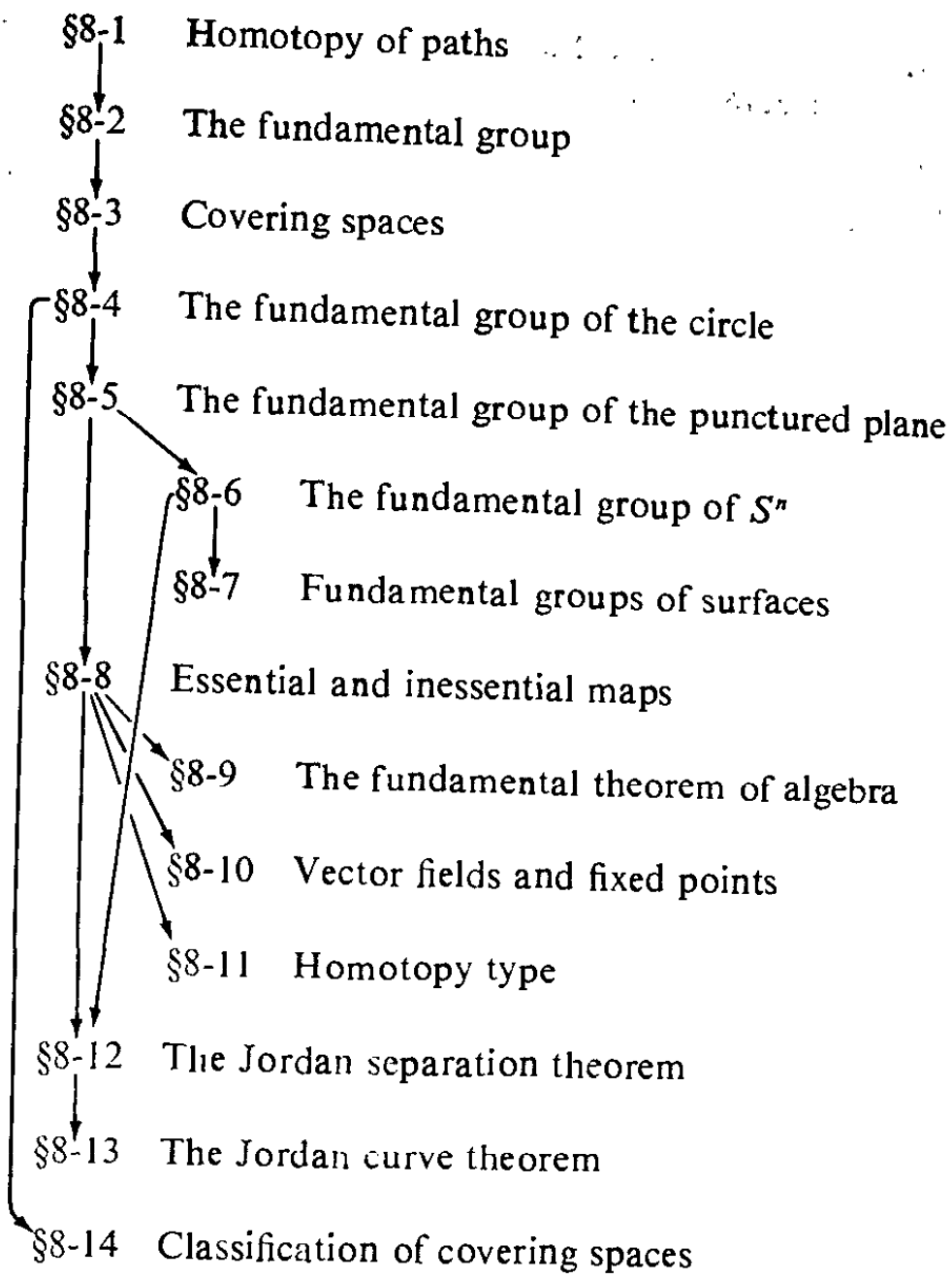
So we must introduce new properties and new techniques. One of the most natural such properties is that of *simple connectivity*. You probably have studied this notion already, when you studied line integrals in the plane. Roughly speaking, one says that a space X is simply connected if every closed curve in X can be shrunk to a point in X. (We shall make this more precise later.) The property of simple connectivity, it turns out, will distinguish between R^2 and R^3; deleting a point from R^3 leaves a simply connected space remaining, but deleting a point from R^2 does not. It will also distinguish between S^2 (which is simply connected) and the torus T (which is not). But it will not distinguish between T and T_2; neither of them is simply connected.

There is an idea more general than the idea of simple connectivity, an idea that includes simple connectivity as a special case. It involves a certain *group* that is called the *fundamental group* of the space. Two spaces that are homeomorphic have fundamental groups that are isomorphic. And the condition of simple connectivity is just the condition that the fundamental group of X be the trivial (one-element) group. Thus the proof that S^2 and T are not homeomorphic can be rephrased by saying that the fundamental group of S^2 is trivial and the fundamental group of T is not. The fundamental group will distinguish between more spaces than the condition of simple connectivity will. It can be used, for example, to show that T and T_2 are not homeomorphic; it turns out that T has an abelian fundamental group and T_2 does not.

In this chapter we define the fundamental group and study its properties. Then we apply it to a number of problems, including the problem of showing that various spaces, such as those already mentioned, are not homeomorphic.

Other applications include applications to vector fields and fixed points and antipode-preserving maps of the sphere, as well as the well-known *fundamental theorem of algebra*, which says that every polynomial equation with real or complex coefficients has a root. Finally, there is the famous *Jordan curve theorem*, to the effect that every simple closed curve C in the plane separates the plane into two components, of which C is the common boundary.

Some of the sections of this chapter are independent of one another. The dependence among them is expressed in the following diagram:



Throughout the chapter, we assume familiarity with components and local connectivity (§3-3 and §3-4). We also assume that the reader is familiar with the elements of group theory. In §8-12, we assume familiarity with local compactness (§3-8), and in §8-13, with quotient spaces (§2-11).

8-1 *Homotopy of Paths*

Before defining the fundamental group of a space X, we shall consider paths on X and an equivalence relation called *path homotopy* between them. And we shall define a certain operation on the collection of these equivalence classes that makes it into what is called in algebra a *groupoid.*

Definition. If f and f' are continuous maps of the space X into the space Y, we say that f is **homotopic** to f' if there is a continuous map $F: X \times I \rightarrow$

Y such that

$$F(x, 0) = f(x) \quad \text{and} \quad F(x, 1) = f'(x)$$

for each $x \in X$. (Here $I = [0, 1]$.) The map F is called a **homotopy** between f and f'. If f is homotopic to f', we write $f \simeq f'$.

We think of a homotopy as a continuous one-parameter family of maps from X to Y. If we imagine the parameter t as representing time, then the homotopy F represents a continuous "deforming" of the map f to the map f', as t goes from 0 to 1.

Now we consider the special case in which f is a path in X. Recall that if $f: [0, 1] \longrightarrow X$ is a continuous map such that $f(0) = x_0$ and $f(1) = x_1$, we say that f is a path in X *from* x_0 *to* x_1. We also say that x_0 is the **initial point**, and x_1 the **final point**, of the path f. In this chapter, we shall for convenience use the interval $I = [0, 1]$ as the domain for all paths.

If f and f' are two paths in X, there is a stronger relation between them than mere homotopy. It is defined as follows:

Definition. Two paths f and f', mapping the interval $I = [0, 1]$ into X, are said to be **path homotopic** if they have the same initial point x_0 and the same final point x_1, and if there is a continuous map $F: I \times I \longrightarrow X$ such that

$$F(s, 0) = f(s) \quad \text{and} \quad F(s, 1) = f'(s),$$
$$F(0, t) = x_0 \quad \text{and} \quad F(1, t) = x_1,$$

for each $s \in I$ and each $t \in I$. We call F a **path homotopy** between f and f'. See Figure 1. If f is path homotopic to f', we write $f \simeq_p f'$.

The first condition says simply that F is a homotopy between f and f', and the second says that for each t, the path

$$s \longrightarrow F(s, t)$$

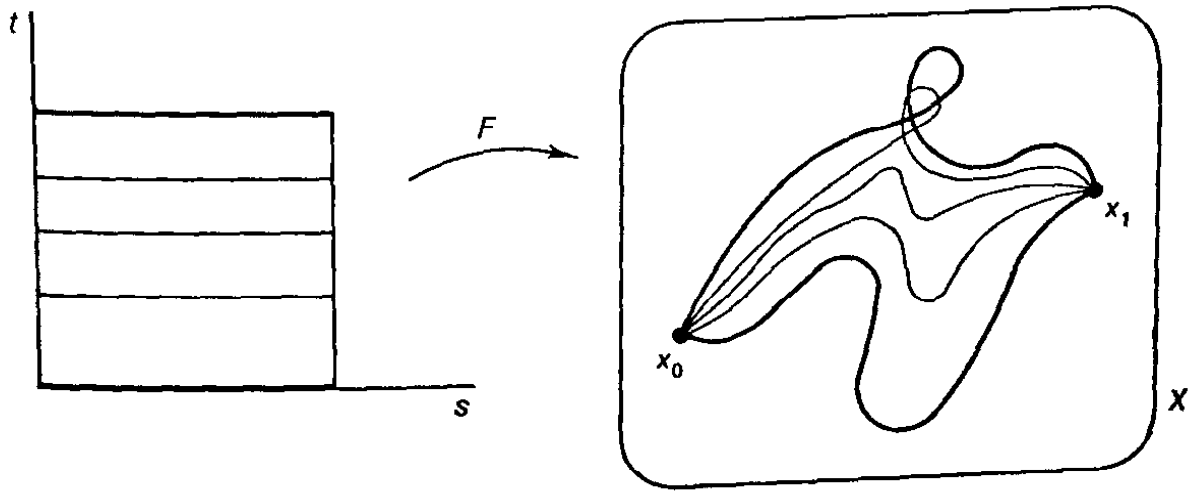


Figure 1

is a path from x_0 to x_1. Said differently, the first condition says that F represents a continuous way of deforming the path f to the path f', and the second condition says that the end points of the path remain fixed during the deformation.

Lemma 1.1. The relations $\simeq$ and $\simeq_p$ are equivalence relations.

If f is a path, we shall denote its path-homotopy equivalence class by $[f]$.

Proof. Let us verify the properties of an equivalence relation.

Given f, it is trivial that $f \simeq f$; the map $F(x, t) = f(x)$ is the required homotopy. If f is a path, F is a path homotopy.

Given $f \simeq f'$, we show that $f' \simeq f$. Let F be a homotopy between f and f'. Then $G(x, t) = F(x, 1 - t)$ is a homotopy between f' and f. If F is a path homotopy, so is G.

Suppose that $f \simeq f'$ and $f' \simeq f''$. We show that $f \simeq f''$. Let F be a homotopy between f and f', and let F' be a homotopy between f' and f''. Define $G : X \times I \to Y$ by the equation

$$G(x, t) = \begin{cases} F(x, 2t) & \text{for } t \in [0, \tfrac{1}{2}], \\ F'(x, 2t - 1) & \text{for } t \in [\tfrac{1}{2}, 1]. \end{cases}$$

The map G is well defined, since if $t = \frac{1}{2}$,

$$F(x, 2t) = F(x, 1) = f'(x) = F'(x, 0) = F'(x, 2t - 1).$$

Because G is continuous on the two closed subsets $X \times [0, \frac{1}{2}]$ and $X \times [\frac{1}{2}, 1]$ of $X \times I$, it is continuous on all of $X \times I$, by the pasting lemma. Thus G is the required homotopy between f and f''.

You can check that if F and F' are path homotopies, so is G. Figure 2 illustrates this situation. □

EXAMPLE 1. Let f and g be any two maps of a space X into R^2. It is easy to see that f and g are homotopic; the map

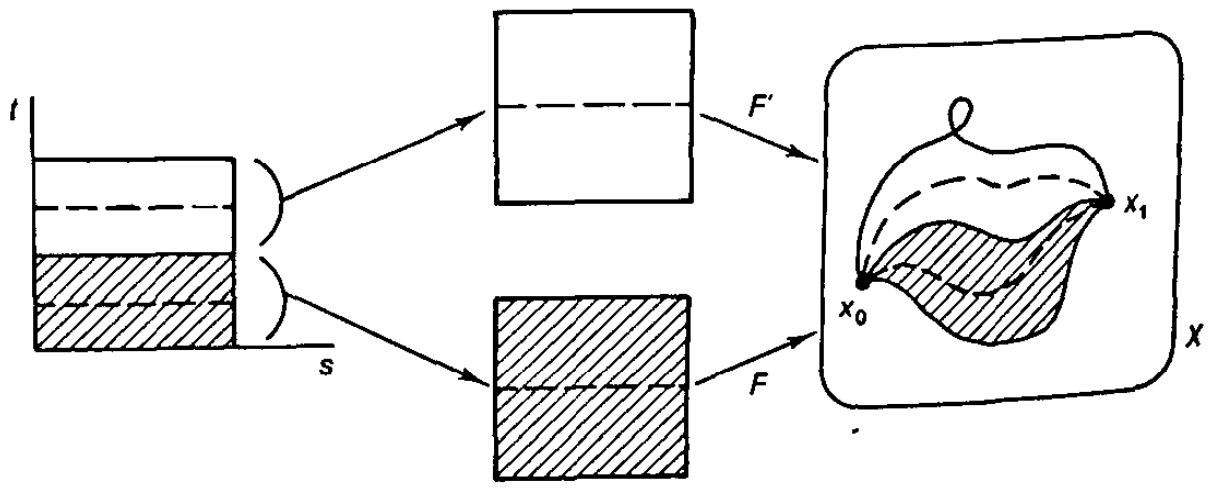


Figure 2

$$F(x, t) = (1 - t)f(x) + tg(x)$$

is a homotopy between them. It is called a **straight-line homotopy**, because it moves the point $f(x)$ to the point $g(x)$ along the straight-line segment joining them.

If f and g are paths from x_0 to x_1, then F will be a path homotopy, as you can check. This situation is pictured in Figure 3.

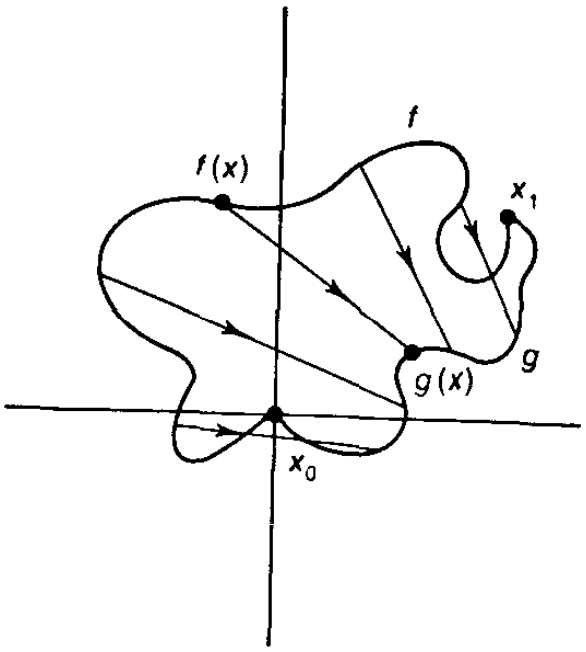


Figure 3

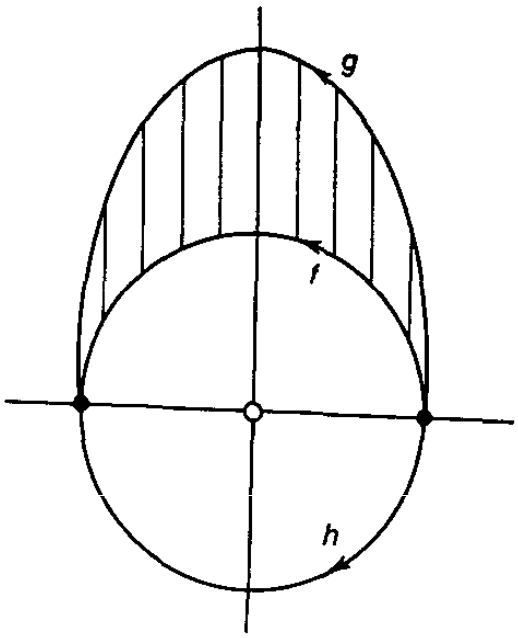


Figure 4

EXAMPLE 2. Let X denote the **punctured plane**, $R^2 - \{0\}$, which we shall denote by $R^2 - 0$ for short. The following paths in X,

$$f(s) = (\cos \pi s, \sin \pi s),$$

$$g(s) = (\cos \pi s, 2 \sin \pi s)$$

are path homotopic; the straight-line homotopy between them is an acceptable path homotopy. But the straight-line homotopy between f and the path

$$h(s) = (\cos \pi s, -\sin \pi s)$$

is not acceptable, for its image does not lie in the space $X = R^2 - 0$. See Figure 4.

Indeed, there exists *no* path homotopy in X between paths f and h. This result is hardly surprizing; it is intuitively clear that one cannot "deform f past the hole at 0" without introducing a discontinuity. But it takes some work to prove. We shall return to this example later.

This example illustrates the fact that you must know what the range space is before you can tell whether two paths are path homotopic or not. The paths f and h would be path homotopic if they were paths in R^2.

Now we introduce some algebra into this geometric situation. We define a certain operation on path-homotopy classes as follows:

Definition. If f is a path in X from x_0 to x_1, and if g is a path in X from x_1 to x_2, we define the **composition** $f * g$ of f and g to be the path h given by the equations

$$h(s) = \begin{cases} f(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ g(2s - 1) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

The function h is well-defined and continuous, by the pasting lemma; and it is a path in X from x_0 to x_2. We think of h as the path whose first half is the path f and whose second half is the path g.

We shall show that the operation of composition on paths induces a well-defined operation on path-homotopy classes, so that we can define

$$[f] * [g] = [f * g].$$

Furthermore, the operation $*$ on path-homotopy classes turns out to satisfy properties that look very much like the axioms for a group. They are called the *groupoid properties* of $*$. The only difference from the properties of a group is that $[f] * [g]$ is not defined for every pair of classes, but only for those pairs $[f]$, $[g]$ for which $f(1) = g(0)$.

Theorem 1.2. *The operation $*$ is well-defined on path-homotopy classes. It has the following properties:*

(1) (*Associativity*) *If* $[\mathrm{f}] * ([\mathrm{g}] * [\mathrm{h}])$ *is defined, so is* $([\mathrm{f}] * [\mathrm{g}]) * [\mathrm{h}]$ *and they are equal.*

(2) (*Right and left identities*) *Given* $\mathrm{x} \in \mathrm{X}$, *let* e_x *denote the constant path* $\mathrm{e}_x : \mathrm{I} \to \mathrm{X}$ *carrying all of* I *to the point* x. *If* f *is a path in* X *from* x_0 *to* x_1, *then*

$$[\mathrm{f}] * [\mathrm{e}_{x_1}] = [\mathrm{f}] \quad \textit{and} \quad [\mathrm{e}_{x_0}] * [\mathrm{f}] = [\mathrm{f}].$$

(3) (*Inverse*) *Given the path* f *in* X *from* x_0 *to* x_1, *let* $\bar{\mathrm{f}}$ *be the path defined by* $\bar{\mathrm{f}}(\mathrm{s}) = \mathrm{f}(1 - \mathrm{s})$. *It is called the* reverse *of* f. *Then*

$$[\mathrm{f}] * [\bar{\mathrm{f}}] = [\mathrm{e}_{x_0}] \quad \textit{and} \quad [\bar{\mathrm{f}}] * [\mathrm{f}] = [\mathrm{e}_{x_1}].$$

Proof. Each of the preceding statements has an elementary geometric proof. To show the operation well-defined, let F be a path homotopy between f and f' and let G be a path homotopy between g and g'. Define

$$H(s, t) = \begin{cases} F(2s, t) & \text{for } s \in [0, \tfrac{1}{2}], \\ G(2s - 1, t) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

Because $F(1, t) = x_1 = G(0, t)$ for all t, the map H is well-defined; it is continuous by the pasting lemma. You can check that H is the required path homotopy between $f * g$ and $f' * g'$. It is pictured in Figure 5.

(1) To prove associativity, we need to show that $f * (g * h) \simeq_p (f * g) * h$. What, exactly, do the paths $f * (g * h)$ and $(f * g) * h$ look like? Under $f * (g * h)$, the image of the point s traces out the image of f as s goes from 0 to $\frac{1}{2}$, it traces out the image of g as s goes from $\frac{1}{2}$ to $\frac{3}{4}$, and it traces out the image

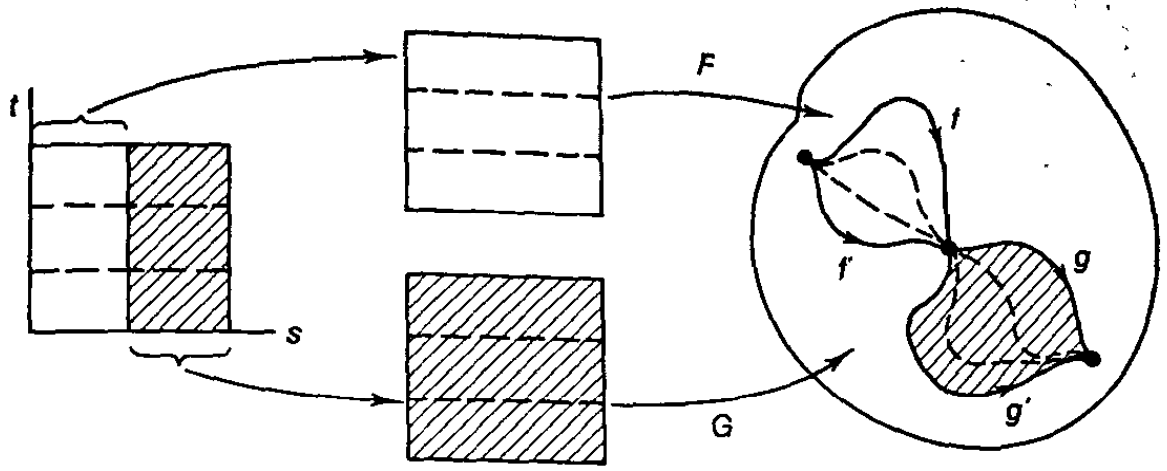


Figure 5

of h as s goes from $\frac{3}{4}$ to 1. Under $(f * g) * h$, the point traces out the same image, but at a different rate. It traces out the image of f as s goes from 0 to $\frac{1}{4}$, of g as s goes from $\frac{1}{4}$ to $\frac{1}{2}$, and of h as s goes from $\frac{1}{2}$ to 1.

We define the homotopy F as follows: First map the unit square I^2 onto I by the continuous map p which collapses each of the three quadrilaterals A, B, and C in Figure 6 onto its base. Follow p by the map $(f * g) * h$. The result will be the desired path homotopy F, as you can check mentally.

More formally, define

$$F(s,t) = \begin{cases} f(4s/(t+1)) & \text{for } s \in [0, (t+1)/4], \\ g(4s - t - 1) & \text{for } s \in [(t+1)/4, (t+2)/4], \\ h((4s - t - 2)/(2 - t)) & \text{for } s \in [(t+2)/4, 1]. \end{cases}$$

You can check that $4s/(t+1)$, $4s - t - 1$, and $(4s - t - 2)/(2 - t)$ lie in the domain $[0, 1]$ of f, g, and h when they are supposed to, so these formulas make sense, and that F is well-defined and hence continuous by the pasting lemma. It is straightforward to show that F is the required path homotopy.

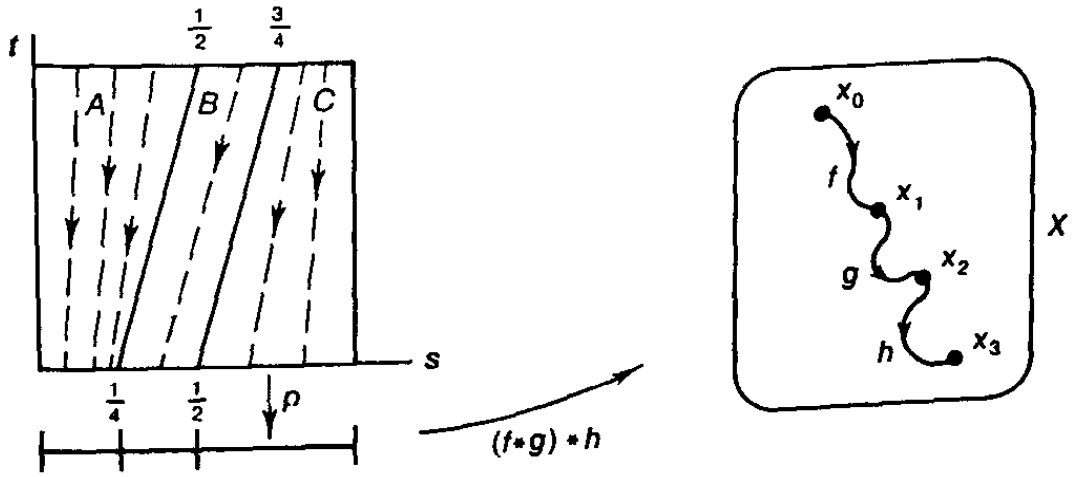


Figure 6

For instance,

$$(f*(g*h))(s) = \begin{cases} f(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ (g*h)(2s-1) & \text{for } s \in [\tfrac{1}{2}, 1] \text{ or } 2s-1 \in [0,1]. \end{cases}$$

Thus we have

$$(f*(g*h))(s) = \begin{cases} f(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ g(2(2s-1)) & \text{for } 2s-1 \in [0, \tfrac{1}{2}] \text{ or } s \in [\tfrac{1}{2}, \tfrac{3}{4}], \\ h(2(2s-1)-1) & \text{for } 2s-1 \in [\tfrac{1}{2}, 1] \text{ or } s \in [\tfrac{3}{4}, 1]. \end{cases}$$

And this is just $F(s, 1)$, as you can check. Similarly, $((f*g)*h)(s) = F(s, 0)$.

(2) We prove that $f \simeq_p f * e_{x_1}$. The desired path homotopy is constructed as follows: First map the unit square I^2 onto I by a continuous map p that collapses the quadrilateral A of Figure 7 onto its base, and collapses the

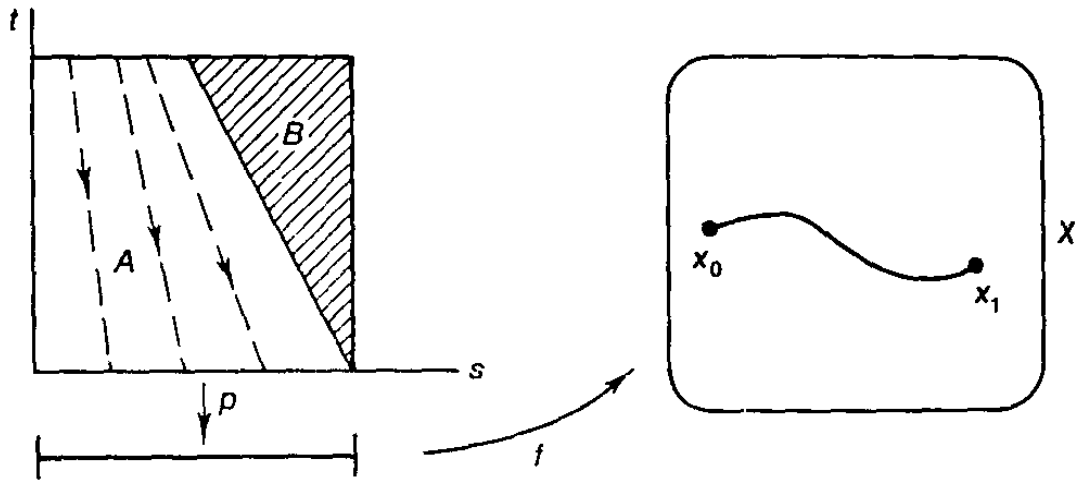


Figure 7

triangle B to its bottom vertex. Follow p by the map f. The result is the desired path homotopy G, as you can check mentally.

Formally, define

$$G(s,t) = \begin{cases} f(2s/(2-t)) & \text{for } s \in [0, (2-t)/2], \\ x_1 & \text{for } s \in [(2-t)/2, 1]. \end{cases}$$

You can check that G is well defined and continuous, and is the desired path homotopy.

The proof that $e_{x_0} * f \simeq_p f$ is similar.

(3) Now we show that $f * \bar{f} \simeq_p e_{x_0}$. The intuitive idea of the proof is very simple. The path $f * \bar{f}$ is a path that goes from x_0 to x_1 and then comes back to x_0 along the same route. Suppose that, for parameter value t, we let α_t be the path which goes out from x_0 partway along the route of f, as far as the point $f(t)$. Then the homotopy we want is just the path $\alpha_t * \bar{\alpha}_t$. It is an acceptable path homotopy, because it leaves both end points fixed at x_0. See Figure 8.

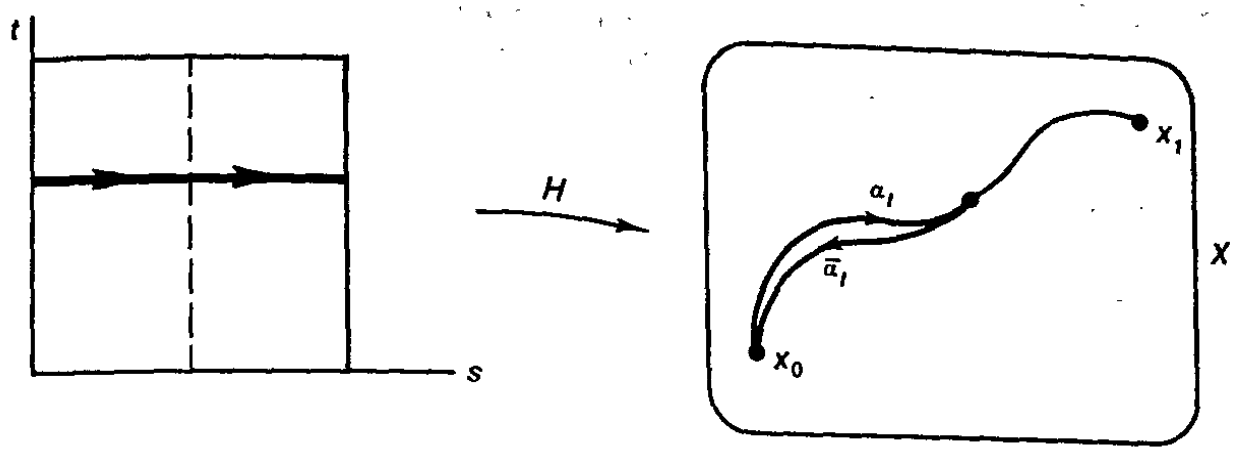


Figure 8

Formally, we define

$$H(s, t) = \begin{cases} f(2ts) & \text{for } s \in [0, \tfrac{1}{2}], \\ f(2t(1 - s)) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

It is easy to check that $2ts$ and $2t(1 - s)$ lie in the domain of f when they are supposed to, so the formulas make sense, that H is well-defined (and hence continuous by the pasting lemma), and that H is the required path homotopy between e_{x_0} and $f * \bar{f}$.

A similar argument could be used to show that $\bar{f} * f \simeq_p e_{x_1}$. But better yet, note the following: We have shown that for any path g, we have $g * \bar{g} \simeq_p e_x$, where x is the initial point of g. In particular, $\bar{f} * \bar{\bar{f}} \simeq_p e_{x_1}$, where $\bar{\bar{f}}$ is the reverse of $\bar{f}$. But the reverse of $\bar{f}$ is just f! Thus $\bar{f} * f \simeq_p e_{x_1}$, as desired. □

Exercises

1. Show that if $h, h' : X \longrightarrow Y$ are homotopic and $k, k' : Y \longrightarrow Z$ are homotopic, then $k \circ h$ and $k' \circ h'$ are homotopic.
2. Suppose that X is a **convex set in** R^n; that is, for every pair x, y of points of X, the line segment joining them lies in X. Show that any two paths in X having the same end points are path homotopic.
3. Consider the proof of Theorem 1.2.
   (a) Check that $F(s, 0) = ((f * g) * h)(s)$ and $F(0, t) = x_0$ and $F(1, t) = x_3$.
   (b) Check the statements made about the maps G and H in (2) and (3).
4. Given spaces X and Y, let $[X, Y]$ denote the set of homotopy classes of maps of X into Y.
   (a) Let $I = [0, 1]$. Show that for any X, the set $[X, I]$ has a single element.
   (b) Show that if Y is path connected, the set $[I, Y]$ has a single element.
5. A space X is said to be **contractible** if the identity map $i_X : X \longrightarrow X$ is homotopic to a constant map.
   (a) Show that I and R are contractible.

(b) Show that a contractible space is path connected.
(c) Show that if Y is contractible, then for any X, the set $[X, Y]$ has a single element.
(d) Show that if X is contractible and Y is path connected, then $[X, Y]$ has a single element.

8-2 *The Fundamental Group*

The set of path-homotopy classes of paths in a space X does not form a group under the operation $*$, only a groupoid. But suppose we pick out a point x_0 of X to serve as a "base point" and restrict ourselves to those paths that begin and end at x_0. The set of these path-homotopy classes does form a group under $*$. It will be called the *fundamental group* of X.

In this section we shall prove several properties of the fundamental group. In particular, we shall show that the group is, up to isomorphism, independent of the choice of base point (provided that X is path connected). We shall also show that the group is a topological invariant of the space X, the fact that is of crucial importance in using it to study homeomorphism problems.

Definition. Let X be a space; let x_0 be a point of X. A path in X that begins and ends at x_0 is called a **loop** based at x_0. The set of path homotopy classes of loops based at x_0, with the operation $*$, is called the **fundamental group** of X relative to the **base point** x_0. It is denoted by $\pi_1(X, x_0)$.

It follows from the preceding theorem that the operation $*$, when restricted to this set, satisfies the axioms for a group. Given two loops f and g based at x_0, the composition $f * g$ is always defined and is a loop based at x_0. Associativity, the existence of an identity element $[e_{x_0}]$, and the existence of an inverse $[\bar{f}]$ for $[f]$ are immediate.

Sometimes this group is called the **first homotopy group** of X, which term implies that there is a second homotopy group. There are indeed groups $\pi_n(X, x_0)$ for all $n \in Z_+$, but we shall not study them in this book. They are part of the general subject called *homotopy theory*.

EXAMPLE 1. Let R^n denote euclidean n-space. Then $\pi_1(R^n, x_0)$ is the trivial group (the group consisting of the identity alone). For if f is a loop in R^n based at x_0, the straight-line homotopy

$$F(s, t) = tx_0 + (1 - t)f(s)$$

is a path homotopy between f and the constant loop e_{x_0}.

EXAMPLE 2. More generally, if X is any *convex* subset of R^n, then $\pi_1(X, x_0)$ is the trivial group. The straight-line homotopy will work once again, for convexity of X means that for any two points x and y of X, the straight-line segment

$$\{tx + (1-t)y \mid 0 \leq t \leq 1\}$$

between them lies in X. In particular, the *unit ball* B^n in R^n,

$$B^n = \{x \mid x_1^2 + \cdots + x_n^2 \leq 1\},$$

has trivial fundamental group.

To find a space with a nontrivial fundamental group is more difficult; this we do in §8-4.

An immediate question one asks is the extent to which the fundamental group depends on the base point. The answer is given in Corollary 2.2, which follows.

Definition. Let α be a path in X from x_0 to x_1. We define a map

$$\hat{\alpha} : \pi_1(X, x_0) \longrightarrow \pi_1(X, x_1)$$

by the equation

$$\hat{\alpha}([f]) = [\bar{\alpha}] * [f] * [\alpha].$$

The map $\hat{\alpha}$ is pictured in Figure 9. It is well-defined because the operation $*$ is well-defined. If f is a loop based at x_0, then $\bar{\alpha} * (f * \alpha)$ is a loop based at x_1. Hence $\hat{\alpha}$ maps $\pi_1(X, x_0)$ into $\pi_1(X, x_1)$, as desired.

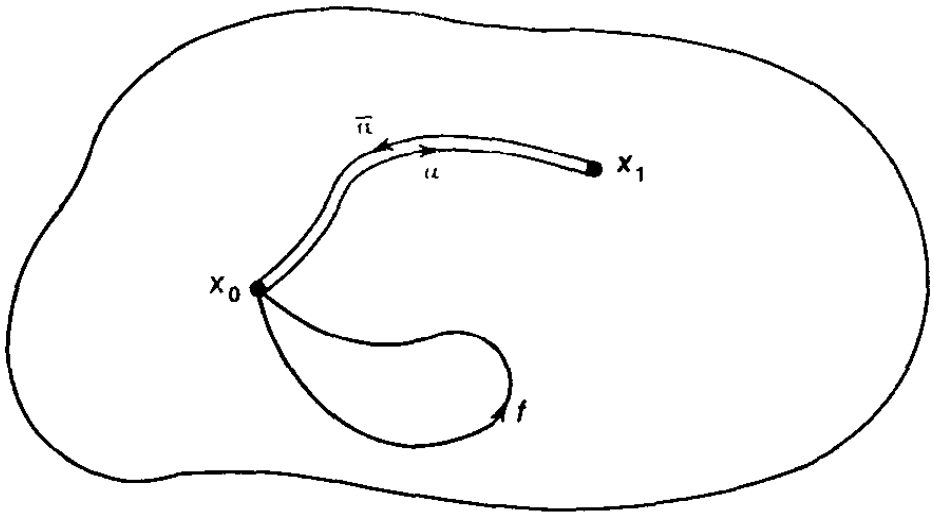


Figure 9

Theorem 2.1. *The map $\hat{\alpha}$ is a group isomorphism.*

Proof. To show that $\hat{\alpha}$ is a homomorphism, we compute

$$\begin{aligned}\hat{\alpha}([f]) * \hat{\alpha}([g]) &= ([\bar{\alpha}] * [f] * [\alpha]) * ([\bar{\alpha}] * [g] * [\alpha]) \\ &= [\bar{\alpha}] * [f] * [g] * [\alpha] \\ &= \hat{\alpha}([f] * [g]).\end{aligned}$$

This proof uses the groupoid properties of $*$, which we proved in Theorem 1.2.

To show that $\hat{\alpha}$ is an isomorphism, we show that if β denotes the path $\bar{\alpha}$,

which is the reverse of α, then $\hat{\beta}$ is an inverse for $\hat{\alpha}$. We compute, for each element $[h]$ of $\pi_1(X, x_1)$,

$$\hat{\beta}([h]) = [\bar{\beta}] * [h] * [\beta] = [\alpha] * [h] * [\bar{\alpha}],$$
$$\hat{\alpha}(\hat{\beta}([h])) = [\bar{\alpha}] * ([\alpha] * [h] * [\bar{\alpha}]) * [\alpha] = [h].$$

A similar computation shows that $\hat{\beta}(\hat{\alpha}([f])) = [f]$ for each $[f] \in \pi_1(X, x_0)$. □

Corollary 2.2. If X *is path connected and* x_0 *and* x_1 *are two points of* X, *then* $\pi_1(X, x_0)$ *is isomorphic to* $\pi_1(X, x_1)$.

Suppose that X is a topological space. Let C be the path component of X containing x_0. It is easy to see that $\pi_1(C, x_0) = \pi_1(X, x_0)$, since all loops and homotopies in X that are based at x_0 must lie in the subspace C. Thus $\pi_1(X, x_0)$ depends only on the path component of X containing x_0, and gives us no information whatever about the rest of X. For this reason, it is usual to deal only with path-connected spaces when studying the fundamental group.

If X is path connected, all the groups $\pi_1(X, x)$ are isomorphic, so it is tempting to try to "identify" all these groups with one another, and to speak simply of the fundamental group of the space X, without reference to base point. The difficulty with this approach is that there is no *natural* way of identifying $\pi_1(X, x_0)$ with $\pi_1(X, x_1)$; different paths α and β from x_0 to x_1 may give rise to different isomorphisms between these groups. For this reason, omitting the base point can lead to error.

It turns out that the isomorphism of $\pi_1(X, x_0)$ with $\pi_1(X, x_1)$ is independent of path if and only if the fundamental group is abelian. (See Exercise 2.) This is a stringent requirement on the space X.

Definition. A space X is said to be simply connected if it is a path-connected space and if $\pi_1(X, x_0)$ is the trivial (one-element) group for some $x_0 \in X$, and hence for every $x_0 \in X$. We often express the fact that $\pi_1(X, x_0)$ is the trivial group by writing $\pi_1(X, x_0) = 0$.

Lemma 2.3. In a simply connected space X, *any two paths having the same initial and final points are path homotopic.*

Proof. Let f and g be two paths from x_0 to x_1. Then $f * \bar{g}$ is defined and is a loop on X based at x_0. Since X is simply connected, $f * \bar{g} \simeq_p e_{x_0}$. Applying the groupoid properties, we see that

$$[(f * \bar{g}) * g] = [e_{x_0} * g] = [g].$$

But

$$[(f * \bar{g}) * g] = [f * (\bar{g} * g)] = [f * e_{x_1}] = [f].$$

Thus f and g are path homotopic. □

It is intuitively clear that the fundamental group is a topological invariant of the space X. A convenient way to prove this fact formally is to introduce the notion of the "homomorphism induced by a continuous map."

Suppose that $h : X \longrightarrow Y$ is a continuous map that carries the point x_0 of X to the point y_0 of Y. We often denote this fact by writing

$$h : (X, x_0) \longrightarrow (Y, y_0).$$

If f is a loop in X based at x_0, then the composite $h \circ f : I \longrightarrow Y$ is a loop in Y based at y_0. The correspondence $f \longrightarrow h \circ f$ thus gives rise to a map carrying $\pi_1(X, x_0)$ into $\pi_1(Y, y_0)$. We define it formally as follows:

Definition. Let $h : (X, x_0) \longrightarrow (Y, y_0)$ be a continuous map. Define

$$h_* : \pi_1(X, x_0) \longrightarrow \pi_1(Y, y_0)$$

by the equation

$$h_*([f]) = [h \circ f].$$

The map h_* is called the **homomorphism induced by** h, relative to the base point x_0.

It is easy to see that h_* is well-defined. If f and f' are path homotopic, let $F : I \times I \longrightarrow X$ be the path homotopy between them. Then $h \circ F$ is a path homotopy between the loops $h \circ f$ and $h \circ f'$. It is also easy to check that h_* is a homomorphism. For

$$(f * g)(s) = \begin{cases} f(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ g(2s - 1) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

It follows that

$$h((f * g)(s)) = \begin{cases} h(f(2s)) & \text{for } s \in [0, \tfrac{1}{2}], \\ h(g(2s - 1)) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

Thus $h \circ (f * g)$ equals the composition $(h \circ f) * (h \circ g)$. It follows that

$$h_*([f] * [g]) = h_*([f]) * h_*([g]),$$

so that h_* is a homomorphism.

The induced homomorphism h_* depends not only on the map $h : X \longrightarrow Y$ but also on the choice of the base point x_0. (Once x_0 is chosen, y_0 is determined by h.) So some notational difficulty will arise if we want to consider several different base points for X. If x_0 and x_1 are two different points of X, we cannot use the same symbol h_* to stand for two different homomorphisms, one having domain $\pi_1(X, x_0)$ and the other having domain $\pi_1(X, x_1)$. Even if X is path connected, so these groups are isomorphic, they are still not the same group. In such a case, we shall use the notation

$$(h_{x_0})_* : \pi_1(X, x_0) \longrightarrow \pi_1(Y, y_0)$$

for the first homomorphism and $(h_{x_1})_*$ for the second. If there is only one base point under consideration, we shall omit mention of the base point and denote the induced homomorphism merely by h_*.

The induced homomorphism has two properties that are crucial in the applications. They are called the "functorial properties" of the induced homomorphism, and are given in the following theorem:

Theorem 2.4. If $h : (X, x_0) \longrightarrow (Y, y_0)$ *and* $k : (Y, y_0) \longrightarrow (Z, z_0)$, *then* $(k \circ h)_* = k_* \circ h_*$. *If* $i : (X, x_0) \longrightarrow (X, x_0)$ *is the identity map, then* i_* *is the identity homomorphism.*

Proof. The proof is a triviality. By definition,

$$(k \circ h)_*([f]) = [(k \circ h) \circ f],$$

$$(k_* \circ h_*)([f]) = k_*(h_*([f])) = k_*([h \circ f]) = [k \circ (h \circ f)].$$

Similarly, $i_*([f]) = [i \circ f] = [f]$. □

Corollary 2.5. If $h : (X, x_0) \longrightarrow (Y, y_0)$ *is a homeomorphism of* X *with* Y, *then* h_* *is an isomorphism of* $\pi_1(X, x_0)$ *with* $\pi_1(Y, y_0)$.

Proof. Let $k : (Y, y_0) \longrightarrow (X, x_0)$ be the inverse of h. Then $k_* \circ h_* = (k \circ h)_* = i_*$, where i is the identity map of (X, x_0); and $h_* \circ k_* = (h \circ k)_* = j_*$, where j is the identity map of (Y, y_0). Since i_* and j_* are the identity homomorphisms of the groups $\pi_1(X, x_0)$ and $\pi_1(Y, y_0)$, respectively, k_* is the inverse of h_*. □

Exercises

1. A subset A of R^n is said to be star convex if for some point a_0 of A, all the line segments joining a_0 to other points of A lie in A.
   (a) Find a star convex set that is not convex.
   (b) Show that if A is star convex, A is simply connected.
   (c) Show that if A is star convex, any two paths in A having the same initial and final points are path homotopic.

2. Let x_0 and x_1 be two given points of the path-connected space X. Show that $\pi_1(X, x_0)$ is abelian if and only if for every pair α and β of paths from x_0 to x_1, we have $\hat{\alpha} = \hat{\beta}$.

3. Let $A \subset X$ and let $r : X \longrightarrow A$ be a retraction (see Exercise 7 of §4-3). Given $a_0 \in A$, show that

$$r_* : \pi_1(X, a_0) \longrightarrow \pi_1(A, a_0)$$

   is surjective. [*Hint*: Consider the inclusion $j : (A, a_0) \longrightarrow (X, a_0)$.]

4. Let A be a subset of R^n; let $h : (A, a_0) \longrightarrow (Y, y_0)$. Show that if h is extendable to a continuous map of R^n into Y, then h_* is the **zero homomorphism** (the trivial homomorphism that maps everything to the identity element).

5. Let $h : X \longrightarrow Y$ be continuous. Show that if X is path connected, the homomorphism induced by h is independent of base point, up to isomorphisms of the groups involved. More precisely, if $h(x_0) = y_0$ and $h(x_1) = y_1$, show that there are isomorphisms ϕ and ψ such that the following diagram of maps "commutes":

$$\begin{array}{ccc} \pi_1(X, x_0) & \xrightarrow{(h_{x_0})_*} & \pi_1(Y, y_0) \\ \downarrow \phi & & \downarrow \psi \\ \pi_1(X, x_1) & \xrightarrow{(h_{x_1})_*} & \pi_1(Y, y_1) \end{array}$$

that is, such that $\psi \circ (h_{x_0})_* = (h_{x_1})_* \circ \phi$. Conclude that if $(h_{x_0})_*$ is the zero homomorphism (or injective, or surjective), then so is $(h_{x_1})_*$.

6. Let G be a topological group with operation $\cdot$ and identity element x_0. (See the Supplementary Exercises for Chapter 2.) Let $\Omega(G, x_0)$ denote the set of all loops in G based at x_0. If $f, g \in \Omega(G, x_0)$, let us define a loop $f \circledast g$ by the rule

$$(f \circledast g)(s) = f(s) \cdot g(s).$$

(a) Show that this operation makes the set $\Omega(G, x_0)$ into a group.
(b) Show that this operation induces a group operation $\circledast$ on $\pi_1(G, x_0)$.
(c) Show that the two group operations $*$ and $\circledast$ on $\pi_1(G, x_0)$ are the same. [*Hint:* Compute $(f * e_{x_0}) \circledast (e_{x_0} * g)$.]
(d) Show that $\pi_1(G, x_0)$ is abelian.

8-3 Covering Spaces

We have not yet computed a nontrivial fundamental group. We shall remedy that deficiency in the next section, when we compute the fundamental group of the circle S^1. We need first to introduce the notion of a covering space. For us, covering spaces will serve mainly as a tool for computing fundamental groups. But they are important in their own right, particularly in the study of Riemann surfaces and complex manifolds. (See [A-S].)

Definition. Let $p: E \longrightarrow B$ be a continuous surjective map. The open set U of B is said to be **evenly covered** by p if the inverse image $p^{-1}(U)$ can be written as the union of disjoint open sets V_α in E such that for each α, the restriction of p to V_α is a homeomorphism of V_α onto U. The collection $\{V_\alpha\}$ will be called a partition of $p^{-1}(U)$ into **slices**.

If U is an open set that is evenly covered by p, we often picture the set $p^{-1}(U)$ as a "stack of pancakes," each having the same size and shape as U, floating in the air above U; the map p squashes them all down onto U. See Figure 10.

Definition. Let $p: E \longrightarrow B$ be continuous and surjective. If every point b of B has a neighborhood U that is evenly covered by p, then p is called a **covering map**, and E is said to be a **covering space** of B.

Note that if $p: E \longrightarrow B$ is a covering map, then for each $b \in B$ the subset $p^{-1}(b)$ of E necessarily has the discrete topology. For each slice V_α is open in

p^{-1} U)

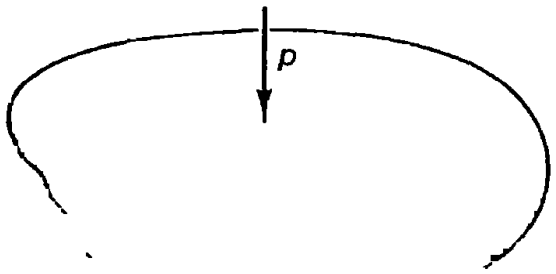


Figure 10

E and intersects the set $p^{-1}(b)$ in a single point; therefore this point is open in the subspace topology on $p^{-1}(b)$.

EXAMPLE 1. Let X be any space; let $i: X \longrightarrow X$ be the identity map. Then i is a covering map (of the most trivial sort). More generally, let E be the space $X \times \{1, \ldots, n\}$ consisting of n disjoint copies of X. The map $p: E \longrightarrow X$ given by $p(x, i) = x$ for all i is again a (rather trivial) covering map. In this case, we can picture the entire space E as a stack of pancakes over X.

To eliminate trivial coverings of the pancake-stack variety, one often requires the space E to be *connected.* A useful example of a covering of this sort is given in the following theorem:

Theorem 3.1. *The map* $\mathrm{p}: \mathrm{R} \longrightarrow \mathrm{S}^1$ *given by the equation*

$$\mathrm{p(x)} = (\cos 2\pi \mathrm{x}, \sin 2\pi \mathrm{x})$$

is a covering map.

One can picture p as a function that wraps the real line R around the circle S^1, and in the process maps each interval $[n, n + 1]$ onto S^1.

Proof. The fact that p is a covering map comes from elementary properties of the sine and cosine functions. Consider, for example, the subset U of S^1 consisting of those points having positive first coordinate. The set $p^{-1}(U)$ consists of those points x for which $\cos 2\pi x$ is positive; that is, it is the union of the intervals

$$V_n = (n - \tfrac{1}{4}, n + \tfrac{1}{4}),$$

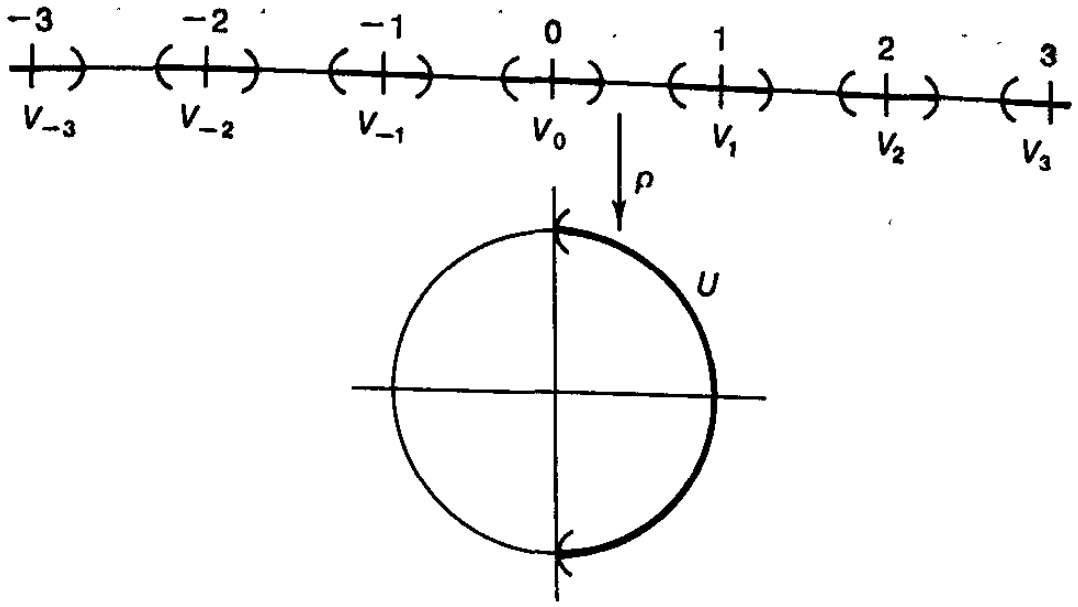


Figure 11

for all $n \in Z$. See Figure 11. Now, restricted to any closed interval $\bar{V}_n$, the map p is injective, because $\sin 2\pi x$ is strictly monotonic on such an interval. Furthermore, p carries $\bar{V}_n$ surjectively onto $\bar{U}$, and V_n to U, by the intermediate value theorem. Since $\bar{V}_n$ is compact, $p \mid \bar{V}_n$ is a homeomorphism of $\bar{V}_n$ with $\bar{U}$. In particular, $p \mid V_n$ is a homeomorphism of V_n with U.

Similar arguments can be applied to the intersections of S^1 with the upper and lower open half-planes, and with the open left-hand half-plane. These open sets cover S^1, and each of them is evenly covered by p. Hence $p : R \longrightarrow S^1$ is a covering map. □

EXAMPLE 2. Consider the space $T = S^1 \times S^1$; it is called the **torus.** It is a general fact that the product of two covering maps is a covering map. (See Exercise 2.) Therefore, as pictured in Figure 12, the product

$$p \times p : R \times R \longrightarrow S^1 \times S^1$$

is a covering of the torus by the plane R^2, where p denotes the covering map of

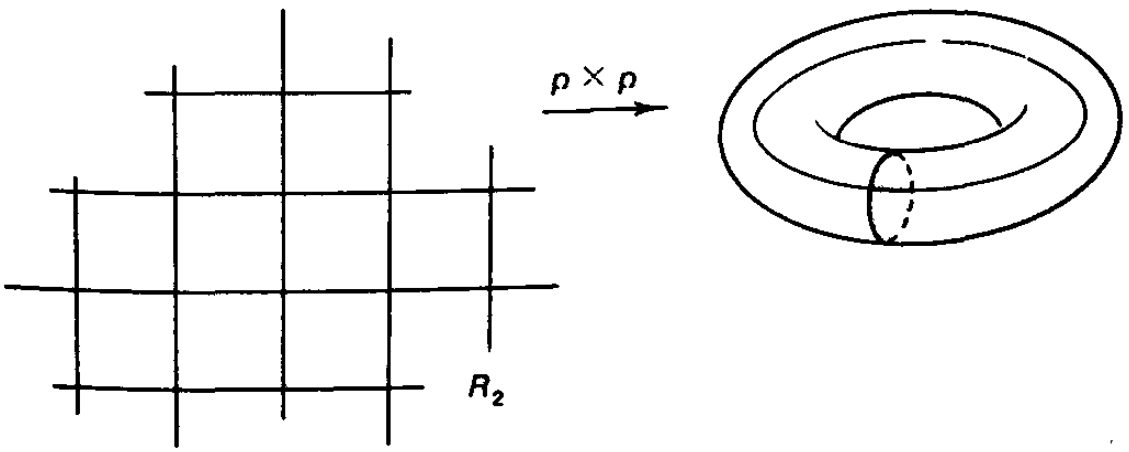


Figure 12

Theorem 3.1. Each of the unit squares $[n, n+1] \times [m, m+1]$ gets wrapped by $p \times p$ entirely around the torus.

In this figure we have pictured the torus not as the product $S^1 \times S^1$, which is a subspace of R^4 and therefore hard to visualize, but as the familiar doughnut-shaped surface D in R^3 obtained by rotating the circle C_1 in the xz-plane of radius $\frac{1}{3}$ centered at $(1, 0, 0)$ about the z-axis. It is not hard to see that $S^1 \times S^1$ is homeomorphic with the surface D. Let C_2 be the circle of radius 1 in the xy-plane centered at the origin. Then let us map $C_1 \times C_2$ into D by defining $f(a \times b)$ to be that point into which a is carried when one rotates the circle C_1 about the z-axis until its center hits the point b. See Figure 13. The map f will be a homeomorphism of $C_1 \times C_2$ with D, as you can check mentally. If you wish, you can write equations for f and check continuity, injectivity, and surjectivity directly. (Continuity of f^{-1} will follow from compactness of $C_1 \times C_2$.)

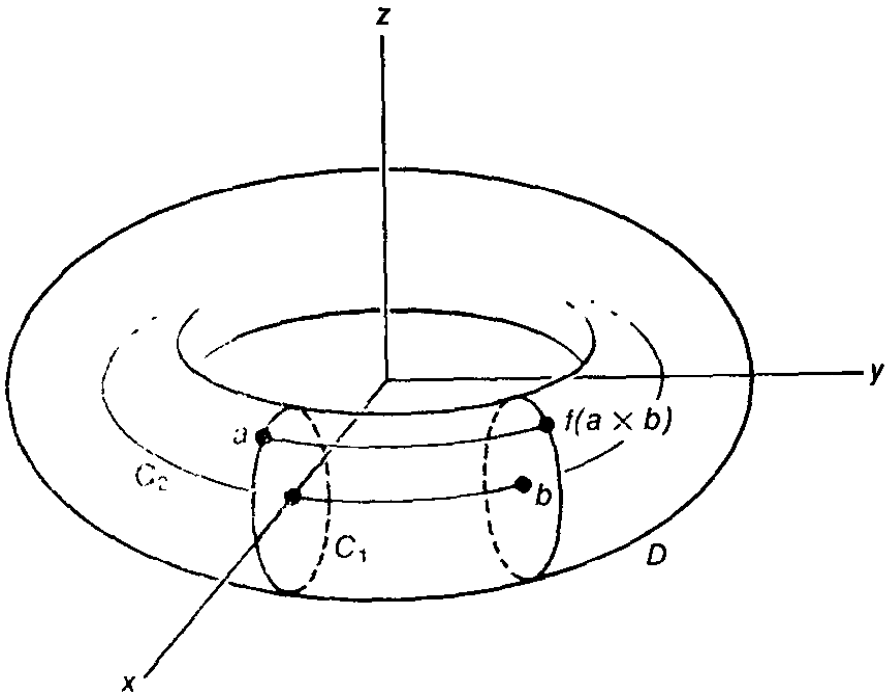


Figure 13

EXAMPLE 3. Consider the covering map

$$p \times i : R \times R_+ \longrightarrow S^1 \times R_+,$$

where i is the identity map of R_+ and p is the map of Theorem 3.1. If we take the standard homeomorphism of $S^1 \times R_+$ with $R^2 - \mathbf{0}$, sending $x \times t$ to tx, the composite gives us a covering

$$R \times R_+ \longrightarrow R^2 - \mathbf{0}$$

of the punctured plane by the open upper half-plane. It is pictured in Figure 14. This covering map appears in the study of complex variables as the *Riemann surface* corresponding to the complex logarithm function.

If $p : E \longrightarrow B$ is a covering map, then p is a local homeomorphism of E with B. That is, each point e of E has a neighborhood that is mapped homeomorphically by p onto an open subset of B. The condition that p be a local homeomorphism does not suffice however to ensure that p is a covering map, as the following example shows.

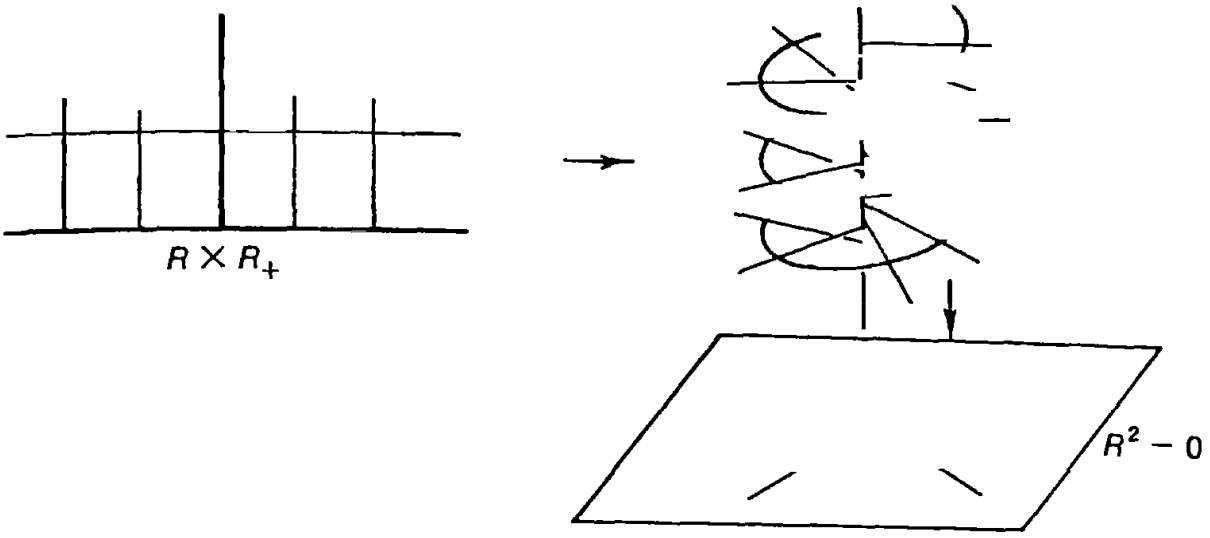


Figure 14

EXAMPLE 4. The map $p : R_+ \longrightarrow S^1$ given by the equation

$$p(x) = (\cos 2\pi x, \sin 2\pi x)$$

is surjective, and it is a local homeomorphism. See Figure 15. But it is not a covering map, for the point $b_0 = (1, 0)$ has no neighborhood U that is evenly covered by p. The typical neighborhood U of b_0 has an inverse image consisting of small neighborhoods V_n of each integer n for $n > 0$, along with a small interval V_0 of the form $(0, \epsilon)$. Each of the intervals V_n for $n > 0$ is mapped homeomorphically onto U by the map p, but the interval V_0 is only imbedded in U by p.

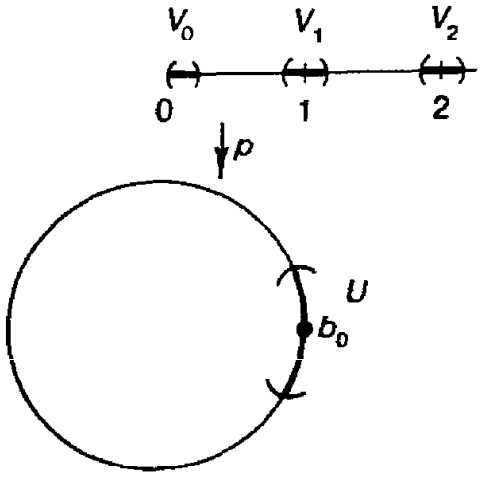


Figure 15

EXAMPLE 5. The preceding example might lead you to think that the real line R is the only connected covering space of the circle S^1. This is not so. Consider, for example, the map $p : S^1 \longrightarrow S^1$ given in equations by

$$p(z) = z^2.$$

[Here we consider S^1 as the subset of the complex plane C consisting of those complex numbers z with $|z| = 1$.] We leave it to you to check that p is a covering map.

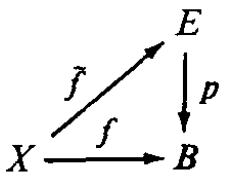
Exercises

1. Show that if $p: E \longrightarrow B$ is a covering map, then p is an open map.
2. Show that if $p: E \longrightarrow B$ and $p': E' \longrightarrow B'$ are covering maps, then $p \times p': E \times E' \longrightarrow B \times B'$ is a covering map.
3. Let Y have the discrete topology. Show that if $p: X \times Y \longrightarrow X$ is projection onto the first coordinate, then p is a covering map.
4. Show that the map p of Example 5 is a covering map. Generalize to the map $p(z) = z^n$.
5. Let $p: E \longrightarrow B$ be a covering map; let B be connected. Show that if $p^{-1}(b_0)$ has k elements for some $b_0 \in B$, then $p^{-1}(b)$ has k elements for every $b \in B$. In such a case, E is called a **k-fold covering** of B.
6. Let $p: X \longrightarrow Y$ and $q: Y \longrightarrow Z$ be covering maps.
   (a) Show that if $q^{-1}(z)$ is finite for each $z \in Z$, then $q \circ p: X \longrightarrow Z$ is a covering map.
   *(b) Show that the theorem fails if $q^{-1}(z)$ is not finite.
7. Let $p: E \longrightarrow B$ be a covering map. Assume that B is connected and locally connected. Show that if C is a component of E, then $p|C: C \longrightarrow B$ is a covering map.
8. Write equations for the map $f: C_1 \times C_2 \longrightarrow D$ of Example 2 and check that it is a homeomorphism.

8-4 The Fundamental Group of the Circle

The study of covering spaces of a space X is intimately related to the study of the fundamental group of X. In this section, we establish the crucial links between the two concepts, and compute the fundamental group of the circle.

Definition. Let $p: E \longrightarrow B$ be a map. If f is a continuous mapping of some space X into B, a **lifting** of f is a map $\tilde{f}: X \longrightarrow E$ such that $p \circ \tilde{f} = f$.



The existence of liftings when p is a covering map is an important tool in studying covering spaces and the fundamental group. First, we show that for a covering space, paths can be lifted; and then we show that path homotopies can be lifted as well. First, an example:

EXAMPLE 1. Consider the covering $p : R \longrightarrow S^1$ of Theorem 3.1. The path $f : [0, 1] \longrightarrow S^1$ beginning at $b_0 = (1, 0)$ given by $f(s) = (\cos \pi s, \sin \pi s)$ lifts to the path $\tilde{f}(s) = s/2$ beginning at 0 and ending at $\frac{1}{2}$. The path $g(s) = (\cos \pi s, -\sin \pi s)$ lifts to the path $\tilde{g}(s) = -s/2$ beginning at 0 and ending at $-\frac{1}{2}$. The path $h(s) = (\cos 4\pi s, \sin 4\pi s)$ lifts to the path $\tilde{h}(s) = 2s$ beginning at 0 and ending at 2. Intuitively, h wraps the interval [0, 1] around the circle twice; this is reflected in the fact that the lifted path $\tilde{h}$ begins at zero and ends at the number 2. These paths are pictured in Figure 16.

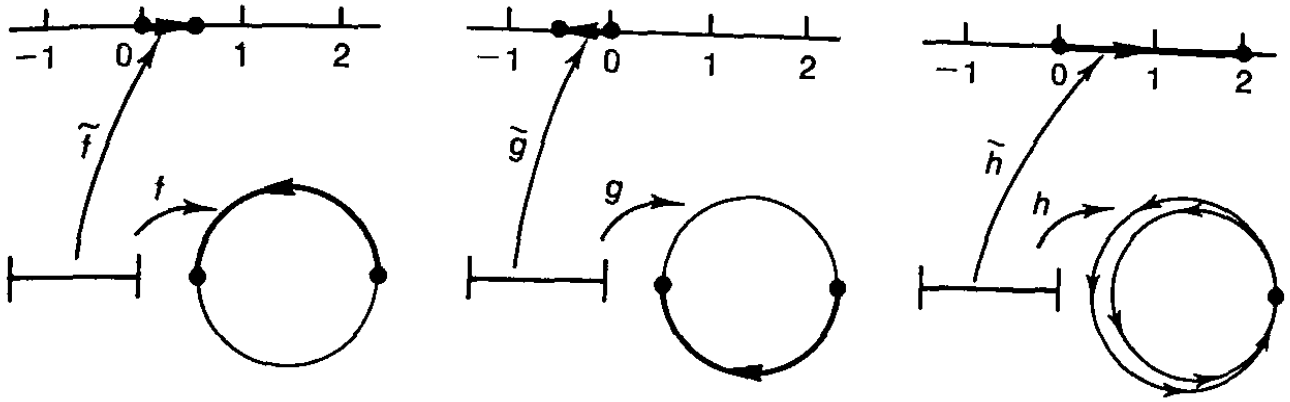


Figure 16

Lemma 4.1. *Let* $p : E \longrightarrow B$ *be a covering map; let* $p(e_0) = b_0$. *Any path* $f : [0, 1] \longrightarrow B$ *beginning at* b_0 *has a unique lifting to a path* $\tilde{f}$ *in* E *beginning at* e_0.

Proof. Cover B by open sets U each of which is evenly covered by p. Find a subdivision of [0, 1], say $s_0, \ldots, s_n$, such that for each i the set $f([s_i, s_{i+1}])$ lies in such an open set U. (Here we use the Lebesgue number lemma.) We define the lifting $\tilde{f}$ step by step.

First, define $\tilde{f}(0) = e_0$. Then, supposing $\tilde{f}(s)$ is defined for $0 \leq s \leq s_i$, we define $\tilde{f}$ on $[s_i, s_{i+1}]$ as follows: The set $f([s_i, s_{i+1}])$ lies in some open set U that is evenly covered by p. Let $\{V_\alpha\}$ be a partition of $p^{-1}(U)$ into slices; each set V_α is mapped homeomorphically onto U by p. Now $\tilde{f}(s_i)$ lies in one of these sets, say in V_0. Define $\tilde{f}(s)$ for $s \in [s_i, s_{i+1}]$ by the equation

$$\tilde{f}(s) = (p \mid V_0)^{-1}(f(s)).$$

Because $p \mid V_0 : V_0 \longrightarrow U$ is a homeomorphism, $\tilde{f}$ will be continuous on $[s_i, s_{i+1}]$.

Continuing in this way, we define $\tilde{f}$ on all of [0, 1]. Continuity of $\tilde{f}$ follows from the pasting lemma of §2-7; the fact that $p \circ \tilde{f} = f$ is immediate from the definition of $\tilde{f}$.

The uniqueness of $\tilde{f}$ is also proved step by step. Suppose that $\tilde{\tilde{f}}$ is another lifting of f beginning at e_0. Then $\tilde{\tilde{f}}(0) = e_0 = \tilde{f}(0)$. Suppose that $\tilde{\tilde{f}}(s) = \tilde{f}(s)$ for all s such that $0 \leq s \leq s_i$. Let V_0 be as in the preceding paragraph; then for $s \in [s_i, s_{i+1}]$, $\tilde{f}(s)$ is defined as $(p \mid V_0)^{-1}(f(s))$. What can $\tilde{\tilde{f}}(s)$ equal? Since $\tilde{\tilde{f}}$ is a lifting of f, it must carry the interval $[s_i, s_{i+1}]$ into the set $p^{-1}(U) = \bigcup V_\alpha$.

The slices V_α are open and disjoint; because the set $\tilde{f}([s_i, s_{i+1}])$ is connected, it must lie entirely in one of the sets V_α. Because $\tilde{f}(s_i) = \tilde{f}(s_i)$, which is in V_0, $\tilde{f}$ must carry all of $[s_i, s_{i+1}]$ into the set V_0. Thus for s in $[s_i, s_{i+1}]$, $\tilde{f}(s)$ must equal some point y of V_0 lying in $p^{-1}(f(s))$. But there is only *one* such point y, namely, $(p|V_0)^{-1}(f(s))$. Hence $\tilde{f}(s) = \tilde{f}(s)$ for $s \in [s_i, s_{i+1}]$. □

Lemma 4.2. *Let* $\mathrm{p}: \mathrm{E} \longrightarrow \mathrm{B}$ *be a covering map; let* $\mathrm{p}(\mathrm{e}_0) = \mathrm{b}_0$. *Let the map* $\mathrm{F}: \mathrm{I} \times \mathrm{I} \longrightarrow \mathrm{B}$ *be continuous, with* $\mathrm{F}(0, 0) = \mathrm{b}_0$. *There is a lifting of* F *to a continuous map*

$$\tilde{\mathrm{F}}: \mathrm{I} \times \mathrm{I} \longrightarrow \mathrm{E}$$

such that $\tilde{\mathrm{F}}(0, 0) = \mathrm{e}_0$. *If* F *is a path homotopy, then* $\tilde{\mathrm{F}}$ *is a path homotopy.*

The lifting $\tilde{F}$ of F is, in fact, unique, but this we shall not prove.

Proof. Given F, we first define $\tilde{F}(0, 0) = e_0$. Next, we use the preceding lemma to extend $\tilde{F}$ to the left-hand edge $0 \times I$ and the bottom edge $I \times 0$ of $I \times I$. Then we extend $\tilde{F}$ to all of $I \times I$ as follows:

Choose subdivisions

$$s_0 < s_1 < \cdots < s_m,$$

$$t_0 < t_1 < \cdots < t_n$$

of I fine enough that each rectangle

$$I_i \times J_j = [s_{i-1}, s_i] \times [t_{j-1}, t_j]$$

is mapped by F into an open set of B that is evenly covered by p. (Use the Lebesgue number lemma.) We define the lifting $\tilde{F}$ step by step, beginning with the rectangle $I_1 \times J_1$, continuing with the other rectangles $I_i \times J_1$ in the "bottom row"; then with the rectangles $I_i \times J_2$ in the next row, and so on.

In general, given i_0 and j_0, assume that $\tilde{F}$ is defined on the set A which is the union of $0 \times I$ and $I \times 0$ and all the rectangles "previous" to $I_{i_0} \times J_{j_0}$ (those rectangles $I_i \times J_j$ for which $j < j_0$ and those for which $j = j_0$ and $i < i_0$). Assume also that $\tilde{F}$ is a continuous lifting of $F|A$. We define $\tilde{F}$ on $I_{i_0} \times J_{j_0}$. Choose an open set U of B that is evenly covered by p and contains the set $F(I_{i_0} \times J_{j_0})$. Let $\{V_\alpha\}$ be a partition of $p^{-1}(U)$ into slices; each set V_α is mapped homeomorphically onto U by p. Now $\tilde{F}$ is already defined on the set $C = A \cap (I_{i_0} \times J_{j_0})$. This set is the union of the left and bottom edges of the rectangle $I_{i_0} \times J_{j_0}$, so *it is connected.* Therefore, $\tilde{F}(C)$ is connected, and must lie entirely within one of the sets V_α. Suppose it lies in V_0. Then the situation is as pictured in Figure 17.

Let $p_0 : V_0 \longrightarrow U$ denote the restriction of p to V_0. Since $\tilde{F}$ is a lifting of $F|A$, we know that for $x \in C$,

$$p_0(\tilde{F}(x)) = p(\tilde{F}(x)) = F(x),$$

so that $\tilde{F}(x) = p_0^{-1}(F(x))$. Hence we may extend F by defining

$$\tilde{F}(x) = p_0^{-1}(F(x))$$

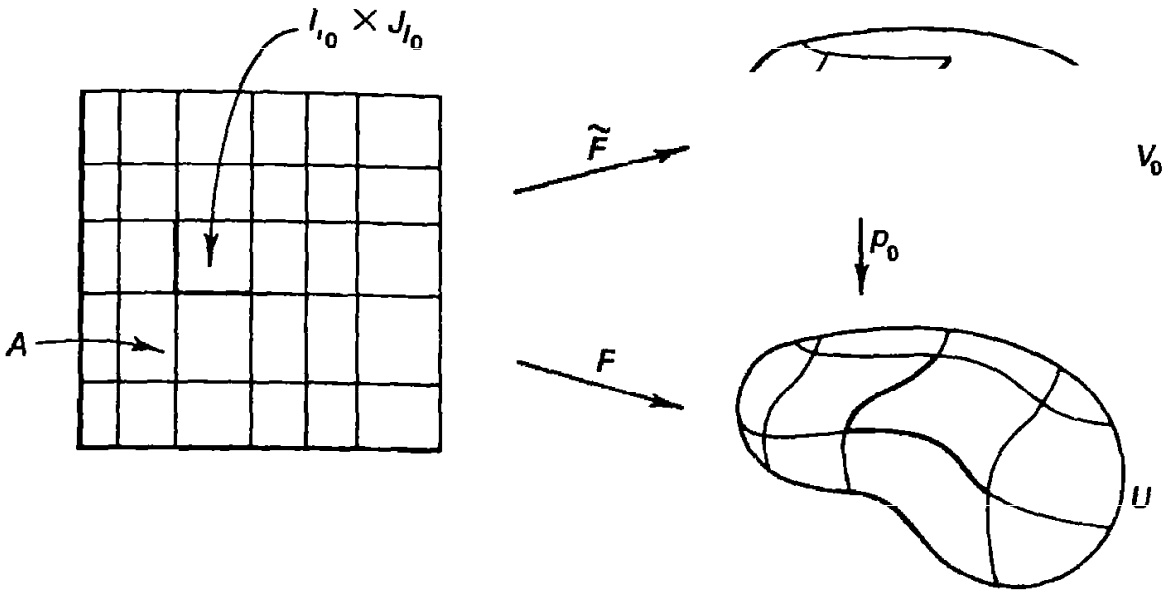


Figure 17

for $x \in I_{i_0} \times J_{j_0}$. The extended map will be continuous by the pasting lemma.

Continuing in this way, we define $\tilde{F}$ on all of I^2.

Now suppose that F is a path homotopy. We wish to show that $\tilde{F}$ is a path homotopy. The map F carries the entire left edge $0 \times I$ of I^2 into a single point b_0 of B. Because $\tilde{F}$ is a lifting of F, it carries this edge into the set $p^{-1}(b_0)$. But this set has the discrete topology as a subspace of E. Since $0 \times I$ is connected and $\tilde{F}$ is continuous, $\tilde{F}(0 \times I)$ is connected and thus must equal a one-point set. Similarly, $\tilde{F}(1 \times I)$ must be a one-point set. Thus $\tilde{F}$ is a path homotopy. □

The following theorem establishes the crucial link between covering spaces and the fundamental group:

Theorem 4.3. *Let* $\mathrm{p}: \mathrm{E} \longrightarrow \mathrm{B}$ *be a covering map; let* $\mathrm{p}(\mathrm{e}_0) = \mathrm{b}_0$. *Let* f *and* g *be two paths in* B *from* b_0 *to* b_1; *let* $\tilde{\mathrm{f}}$ *and* $\tilde{\mathrm{g}}$ *be their respective liftings to paths in* E *beginning at* e_0. *If* f *and* g *are path homotopic, then* $\tilde{\mathrm{f}}$ *and* $\tilde{\mathrm{g}}$ *end at the same point of* E *and are path homotopic.*

Proof. Let $F: I \times I \longrightarrow B$ be the path homotopy between f and g. Then $F(0, 0) = b_0$. Let $\tilde{F}: I \times I \longrightarrow E$ be a lifting of F to E such that $\tilde{F}(0, 0) = e_0$. By the preceding lemma, $\tilde{F}$ is a path homotopy, so that $\tilde{F}(0 \times I) = \{e_0\}$ and $\tilde{F}(1 \times I)$ is a one-point set $\{e_1\}$.

The restriction $\tilde{F} \mid I \times 0$ of $\tilde{F}$ to the bottom edge of $I \times I$ is a path on E beginning at e_0 that is a lifting of $F \mid I \times 0$. By uniqueness of path liftings, we must have $\tilde{F}(s, 0) = \tilde{f}(s)$. Similarly, $\tilde{F} \mid I \times 1$ is a path on E that is a lifting of $F \mid I \times 1$, and it begins at e_0 because $F(0 \times I) = \{e_0\}$. By uniqueness of path liftings, $\tilde{F}(s, 1) = \tilde{g}(s)$. Therefore, both $\tilde{f}$ and $\tilde{g}$ end at e_1, and $\tilde{F}$ is a path homotopy between them. □

Now we apply this theorem to the case of S^1.

Theorem 4.4. *The fundamental group of the circle is infinite cyclic.*

Proof. Let b_0 be the point $(1, 0)$ of S^1. We shall construct an isomorphism of the group $\pi_1(S^1, b_0)$ with the group $(Z, +)$ of integers.

For this purpose, consider the covering map $p : R \longrightarrow S^1$ given by the equation $p(x) = (\cos 2\pi x, \sin 2\pi x)$. If f is a loop on S^1 based at b_0, let $\tilde{f}$ be the lifting of f to a path on R beginning at 0. The point $\tilde{f}(1)$ must be a point of the set $p^{-1}(b_0)$; that is, $\tilde{f}(1)$ must equal some integer n. The preceding theorem tells us that this integer depends only on the path homotopy class of f. Therefore, we can define

$$\phi : \pi_1(S^1, b_0) \longrightarrow Z$$

by letting $\phi([f])$ be this integer. We assert that ϕ is a group isomorphism.

The map ϕ is surjective. Let n be a point of $p^{-1}(b_0)$. Because R is path connected, we can choose a path $\tilde{f} : [0, 1] \longrightarrow R$ in R from 0 to n. Define $f = p \circ \tilde{f}$. Then f is a loop in S^1 based at b_0, and $\tilde{f}$ is its lifting to a path in R beginning at 0. By definition, $\phi([f]) = n$.

The map ϕ is injective. Assume that $\phi([f]) = n = \phi([g])$; we shall prove that $[f] = [g]$. Let $\tilde{f}$ and $\tilde{g}$ be the liftings of f and g, respectively, to paths on R beginning at 0; both $\tilde{f}$ and $\tilde{g}$ end at n, by hypothesis. Because R is simply connected, $\tilde{f}$ and $\tilde{g}$ are path homotopic; let $\tilde{F}$ be the path homotopy between them. The map $F = p \circ \tilde{F}$ will be a path homotopy between f and g, as you can check.

The map ϕ is a homomorphism. Let f and g be two loops in S^1 based at b_0; let $\tilde{f}$ and $\tilde{g}$ be their liftings, respectively, to paths on R beginning at 0. Let $\tilde{f}(1) = n$ and $\tilde{g}(1) = m$. Define a path h on R by the equations

$$h(s) = \begin{cases} \tilde{f}(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ n + \tilde{g}(2s - 1) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

Then h is a path on R beginning at 0. We assert that h is a lifting of $f * g$. Note first that for all x, we have $p(n + x) = p(x)$, because the functions sine and cosine have period 2π. Then

$$p(h(s)) = \begin{cases} p(\tilde{f}(2s)) = f(2s) & \text{for } s \in [0, \tfrac{1}{2}], \\ p(n + \tilde{g}(2s - 1)) = p(\tilde{g}(2s - 1)) = g(2s - 1) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

Thus $p \circ h = f * g$, so that h is the lifting of $f * g$ which begins at 0. By definition, $\phi([f * g])$ is $h(1)$, which equals $n + m$. Therefore,

$$\phi([f * g]) = \phi([f]) + \phi([g]),$$

as desired. □

Most of the proof of the preceding theorem generalizes to arbitrary simply connected covering spaces. The only part special to the covering $p : R \longrightarrow S^1$

was the existence of the addition operation in R; that operation enabled us to show that ϕ was a homomorphism. In a general covering space, we have no convenient addition operation; nevertheless, we can still get a good bit of information about the fundamental group.

Theorem 4.5. *Let* $p : (E, e_0) \longrightarrow (B, b_0)$ *be a covering map. If* E *is path connected, then there is a surjection*

$$\phi : \pi_1(B, b_0) \longrightarrow p^{-1}(b_0).$$

If E *is simply connected,* ϕ *is a bijection.*

Proof. The proof is almost a copy of the proof of Theorem 4.4; we leave the details to you. □

Definition. If E is a simply connected space, and if $p : E \longrightarrow B$ is a covering map, then we say that E is a **universal covering space** of B.

If B has a universal covering space E, then E is uniquely determined up to homeomorphism (provided B is locally path connected). See Exercise 12. The reason we call E a *universal* covering space is given in the same exercise.

We shall study in §8-14 conditions under which a space B possesses a universal covering space.

Exercises

1. What goes wrong with the "path-lifting lemma" (Lemma 4.1) for the local homeomorphism of Example 4 of §8-3?
2. In defining the map $\tilde{F}$ in the proof of Lemma 4.2, why were we so careful about the order in which we considered the small rectangles?
3. Let $p : E \longrightarrow B$ be a covering map. Let α and β be paths in B with $\alpha(1) = \beta(0)$; let $\tilde{\alpha}$ and $\tilde{\beta}$ be liftings of them such that $\tilde{\alpha}(1) = \tilde{\beta}(0)$. Show that $\tilde{\alpha} * \tilde{\beta}$ is a lifting of $\alpha * \beta$.
4. Consider the covering map $p : R \times R_+ \longrightarrow R^2 - \mathbf{0}$ of Example 3, §8-3. Find liftings of the paths

$$f(t) = (2 - t, 0),$$
$$g(t) = ((1 + t)\cos 2\pi t, (1 + t)\sin 2\pi t)$$
$$h(t) = f * g.$$

   Sketch these paths and their liftings.
5. Let $n \in Z_+$. Consider the maps $g, h : S^1 \longrightarrow S^1$ given by $g(z) = z^n$ and $h(z) = 1/z^n$. [Here we represent S^1 as the set of complex numbers z of absolute value 1.] Compute the induced homomorphisms g_*, h_* of the infinite cyclic group $\pi_1(S^1, b_0)$ into itself.

6. Show there is no retraction $r : B^2 \longrightarrow S^1$, where B^2 is the unit ball in R^2.

7. Prove Theorem 4.5.

8. Show that if B is simply connected, then any covering map $p : E \longrightarrow B$ for which E is path connected is a homeomorphism.

9. Consider the covering map $p \times p : R \times R \longrightarrow S^1 \times S^1$ of Example 2, §8-3. Consider the path

$$f(t) = (\cos 2\pi t, \sin 2\pi t) \times (\cos 4\pi t, \sin 4\pi t).$$

in $S^1 \times S^1$. Sketch what f looks like when $S^1 \times S^1$ is identified with the doughnut surface D. Find a lifting $\tilde{f}$ of f to $R \times R$, and sketch it. Formulate a conjecture about the fundamental group of the torus, and prove it.

10. Prove the following generalization of Theorem 4.5:

Theorem. Let $p : E \longrightarrow B$ be a covering map; let E be path connected; let $p(e_0) = b_0$. Then

(a) *$p_* : \pi_1(E, e_0) \longrightarrow \pi_1(B, b_0)$ is injective.*

(b) *There is a bijection*

$$\phi : \pi_1(B, b_0)/H \longrightarrow p^{-1}(b_0),$$

where $H = p_*(\pi_1(E, e_0))$ and $\pi_1(B, b_0)/H$ is the collection of right cosets of H in $\pi_1(B, b_0)$.

11. Assume the hypotheses of the preceding theorem. If $\pi_1(B, b_0)$ is infinite cyclic, what can you say about $\pi_1(E, e_0)$? What if $\pi_1(B, b_0)$ is finite and p is a k-fold covering?

*12. (a) *Lemma (A lifting lemma). Let $p : E \longrightarrow B$ be a covering map; let $p(e_0) = b_0$. Let $f : (Y, y_0) \longrightarrow (B, b_0)$ be a continuous map. If Y is locally path connected and simply connected, then f can be lifted uniquely to a continuous map $\tilde{f} : Y \longrightarrow E$ such that $f(y_0) = e_0$.*

[*Hint:* Given $y \in Y$, choose a path α in Y from y_0 to y; lift $f \circ \alpha$ to a path in E beginning at e_0, and define $\tilde{f}(y)$ to be the end point of this path.]

(b) *Theorem. Let B be locally path connected. Suppose $p : E \longrightarrow B$ is a simply connected covering space of B. If $p' : E' \longrightarrow B$ is any path-connected covering space of B, then there is a covering map $q : E \longrightarrow E'$ such that $p = p' \circ q$.*

[*Hint:* Lift p to E'.] This theorem shows why E is called a *universal* covering space of B; it covers every other path-connected covering space of B.

(c) *Theorem (Uniqueness of the universal covering space). If B is locally path connected, and if $p : E \longrightarrow B$ and $p' : E' \longrightarrow B$ are two simply connected covering spaces of B, then there is a homeomorphism $h : E \longrightarrow E'$ such that $p' \circ h = p$.*

*13. *Theorem. Let G be a topological group with operation $\cdot$; suppose $p : \tilde{G} \longrightarrow G$ is a simply connected covering space of G. If G is locally path connected, then there is a multiplication $\odot$ on $\tilde{G}$ relative to which $\tilde{G}$ is a topological group and p is a homomorphism.*

[*Hint:* Let e be the identity element of G; choose $\tilde{e}$ in $p^{-1}(e)$. Given $\tilde{x}$ and $\tilde{y}$

in $\tilde{G}$, choose paths $\tilde{\alpha}$ and $\tilde{\beta}$ from $\tilde{e}$ to $\tilde{x}$ and $\tilde{y}$, respectively. Let $\alpha(s) = p(\tilde{\alpha}(s))$ and $\beta(s) = p(\tilde{\beta}(s))$. Take the path $\gamma(s) = \alpha(s) \cdot \beta(s)$ in G, lift it to a path $\tilde{\gamma}$ in $\tilde{G}$ that begins at $\tilde{e}$, and define $\tilde{x} \odot \tilde{y}$ to be the end point of $\tilde{\gamma}$.]

The group $\tilde{G}$ is called the **universal covering group** of G.

8-5 The Fundamental Group of the Punctured Plane

In this section, we shall prove that the fundamental group of the punctured plane $R^2 - \mathbf{0}$ is infinite cyclic. We shall need this fact in several of the applications. The proof will lead us into a short discussion of deformation retracts and their relation to the fundamental group.

Theorem 5.1. *Let* $x_0 \in S^1$. *The inclusion mapping*

$$j : (S^1, x_0) \longrightarrow (R^2 - \mathbf{0}, x_0)$$

induces an isomorphism of fundamental groups.

Proof. Let $r : R^2 - \mathbf{0} \longrightarrow S^1$ be the continuous map defined by $r(x) = x/\|x\|$, where $\|x\|$ denotes the distance of x from the origin $\mathbf{0}$ in the euclidean metric. The map r can be pictured as collapsing each radial ray in $R^2 - \mathbf{0}$ onto the point where the ray intersects S^1; it maps each point x of S^1 to itself.

We claim that r_* is an inverse for j_*. First consider the composite map

$$(S^1, x_0) \xrightarrow{j} (R^2 - \mathbf{0}, x_0) \xrightarrow{r} (S^1, x_0);$$

this map equals the identity map of S^1. Therefore, by the functorial properties of the induced homomorphism, $r_* \circ j_*$ is the identity isomorphism of $\pi_1(S^1, x_0)$.

To show that $j_* \circ r_*$ is the identity isomorphism of $\pi_1(R^2 - \mathbf{0}, x_0)$ with itself, let us take a loop f in $R^2 - \mathbf{0}$ based at x_0. Then $j_*(r_*[f])$ is the homotopy class of the loop $g = j \circ r \circ f : I \longrightarrow R^2 - \mathbf{0}$, which is given in equations by

$$g(s) = \frac{f(s)}{\|f(s)\|}.$$

We need to show that $g \simeq_p f$; but this is easy. As indicated in Figure 18, we just move the point $f(s)$ gradually along its radial ray until it hits the point $g(s)$.

More formally, define $F : I \times I \longrightarrow R^2 - \mathbf{0}$ by the equation

$$F(s, t) = t \frac{f(s)}{\|f(s)\|} + (1 - t) f(s).$$

It is clear that $F(s, t)$ is never equal to $\mathbf{0}$, for

$$\frac{t}{\|f(s)\|} + (1 - t) \neq 0 \quad \text{and} \quad f(s) \neq \mathbf{0}.$$

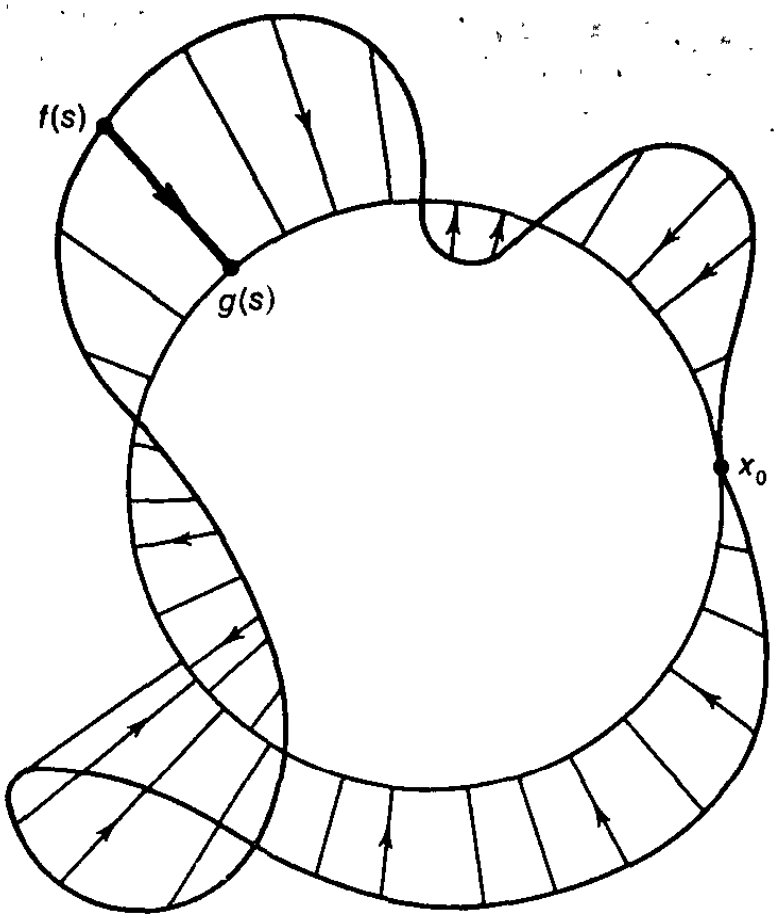


Figure 18

Also, $F(0, t) = F(1, t) = x_0$ for all t. Thus F is the required path homotopy between f and g. □

There is nothing special about the circle and the plane in the preceding proof. Exactly the same proof applies to prove the following theorem:

Theorem 5.2. *If* $x_0 \in S^{n-1}$, *the inclusion*

$$j : (S^{n-1}, x_0) \longrightarrow (R^n - 0, x_0)$$

induces an isomorphism of fundamental groups.

Of course, we have not yet computed the fundamental group of S^{n-1}, so this does not tell us much.

What made the preceding proof work? Roughly speaking, it worked because we had a natural way of deforming the path f, which lies in $R^2 - 0$, into the path $r \circ f$, which lies in S^1. Another way of looking at this proof is to note that we can deform the *entire space* $R^2 - 0$ into the circle S^1 if we like, by gradually collapsing each radial line to the point where it intersects S^1. This deformation is the one that deforms the path f into the path $r \circ f$; it is called a *strong deformation retraction* of $R^2 - 0$ onto S^1.

Analyzing the proof in this way leads to a generalization of Theorem 5.1. Although we shall not have occasion to use it in this book, it is a theorem that is very useful in computing homotopy groups.

Recall that if A is a subspace of X, we say that A is a *retract* of X if there is a continuous map $r : X \to A$ such that $r(a) = a$ for every $a \in A$. The map r is called a *retraction* of X onto A.

Definition. Let A be a subspace of X. Then A is said to be a **strong deformation retract** of X if there is a continuous map $H : X \times I \to X$ such that

$$H(x, 0) = x \quad \text{for } x \in X,$$
$$H(x, 1) \in A \quad \text{for } x \in X,$$
$$H(a, t) = a \quad \text{for } a \in A \quad \text{and} \quad t \in I.$$

The map H is called a **strong deformation retraction.**

Said differently, the space A is a strong deformation retract of X if X can be deformed *gradually* into A, with each point of A remaining fixed during the deformation. At the end of the deformation, we have a retraction of X onto A, mapping x into $H(x, 1)$.

EXAMPLE 1. The map $H : (R^n - 0) \times I \to (R^n - 0)$ defined by

$$H(x, t) = t\frac{x}{\|x\|} + (1 - t)x$$

is a strong deformation retraction of $R^n - 0$ onto S^{n-1}; it gradually collapses each radial line into the point where it intersects S^{n-1}.

Theorem 5.3. *Let* A *be a strong deformation retract of* X. *Let* $a_0 \in$ A. *Then the inclusion map*

$$j : (A, a_0) \to (X, a_0)$$

induces an isomorphism of fundamental groups.

The proof is similar to that of Theorem 5.1; we leave it to you. This theorem can be used to compute fundamental groups for a number of spaces, by reducing the spaces involved to more familiar ones. Typical examples follow.

EXAMPLE 2. Let B denote the z-axis in R^3. Consider the space $R^3 - B$. It has the punctured xy-plane $(R^2 - 0) \times 0$ as a strong deformation retract. The map H defined by the equation

$$H(x, y, z, t) = (x, y)(1 - t)z$$

is a strong deformation retraction; it gradually collapses each line parallel to the z-axis into the point where the line intersects the xy-plane. We conclude that the space $R^3 - B$ has an infinite cyclic fundamental group.

EXAMPLE 3. Consider $R^2 - p - q$, the **doubly punctured plane.** We assert it has the "figure eight" space as a strong deformation retract. Rather than writing equations, we merely sketch the deformation retraction; it is the three-stage deformation indicated in Figure 19.

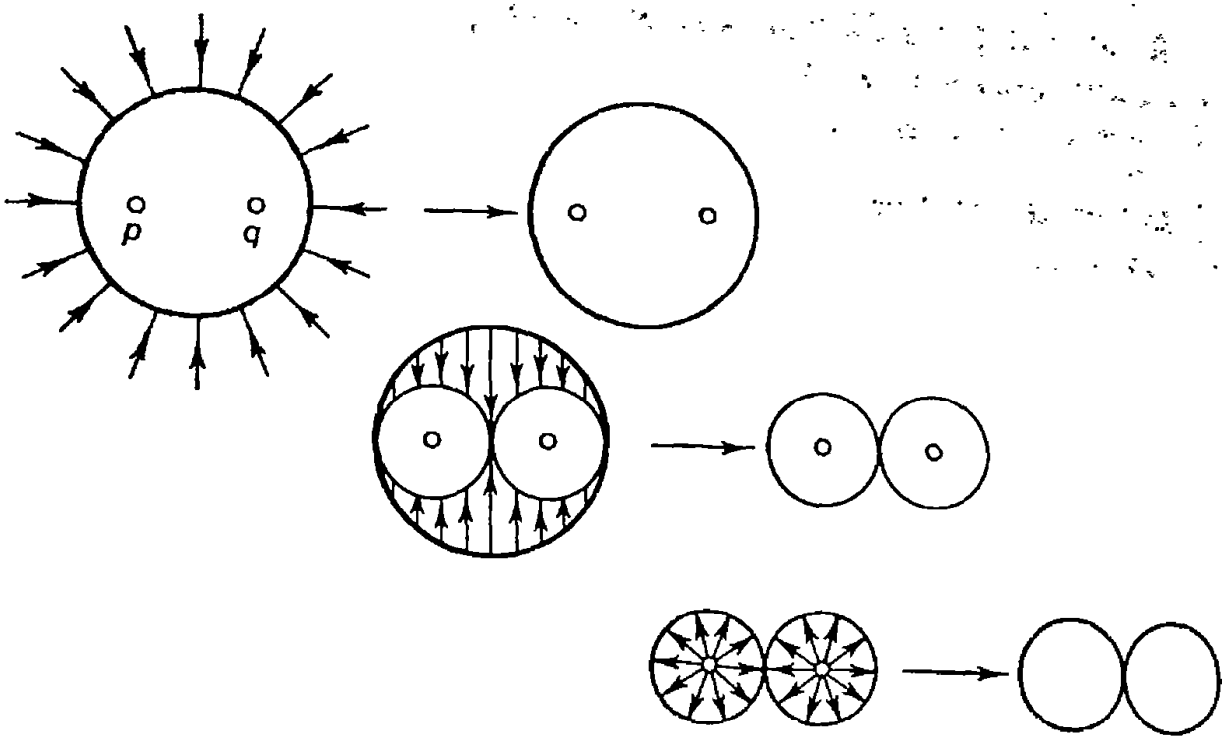


Figure 19

EXAMPLE 4. Another strong deformation retract of $R^2 - p - q$ is the "theta space"

$$\theta = S^1 \cup (0 \times [-1, 1]);$$

we leave it to you to sketch the maps involved. As a result, the figure eight (8) and the theta space (θ) have isomorphic fundamental groups, even though neither is a strong deformation retract of the other.

Of course, we do not know anything about the fundamental group of the figure eight as yet. But we shall.

We should remark that Theorem 5.3 remains true if one replaces the condition

$$H(a, t) = a \quad \text{for } a \in A \text{ and } t \in I$$

by either of the weaker conditions

$$H(a, t) \in A \quad \text{for } a \in A \text{ and } t \in I,$$

or

$$H(a, 1) = a \quad \text{for } a \in A.$$

But the proof in either of these cases is more difficult. (See Exercise 5 of §8-11.) In practice, the theorem one usually applies is Theorem 5.3 anyway.

Exercises

1. Show that if A is a strong deformation retract of X, and B is a strong deformation retract of A, then B is a strong deformation retract of X.

2. Show that the paths f and h in Example 2 of §8-1 are not path homotopic.
3. Prove Theorem 5.2.
4. Prove Theorem 5.3.
5. For each of the following spaces, the fundamental group is either trivial, infinite cyclic, or isomorphic to the fundamental group of the figure eight. Determine for each space which of the three alternatives holds.
   (a) The "solid torus," $B^2 \times S^1$.
   (b) The torus T with a point removed.
   (c) The cylinder $S^1 \times I$.
   (d) The infinite cylinder $S^1 \times R$.
   (e) R^3 with the nonnegative x, y, and z axes deleted.

   The following subsets of R^2:
   (f) $\{x \mid \|x\| > 1\}$
   (g) $\{x \mid \|x\| \geq 1\}$
   (h) $\{x \mid \|x\| < 1\}$
   (i) $S^1 \cup (R_+ \times 0)$
   (j) $S^1 \cup (R_+ \times R)$
   (k) $S^1 \cup (R \times 0)$
   (l) $R^2 - (R_+ \times 0)$
6. Let $\mathbf{C}$ denote the complex plane; let f be a loop in $\mathbf{C} - \mathbf{0}$ based at x_0.
   (a) Let $p : R \longrightarrow S^1$ be the standard covering map. Consider the loop $h(s) = f(s)/\|f(s)\|$ in S^1; lift it to a path $\tilde{h}$ in R. Then $\tilde{h}(1) = \tilde{h}(0) + n$ for some integer n. Show that the path homotopy class of f determines n uniquely, and conversely. The integer n is called the **winding number** of f with respect to $\mathbf{0}$.
   (b) *Theorem. Suppose f is a piecewise-differentiable loop in $\mathbf{C} - \mathbf{0}$. Then the winding number of f with respect to $\mathbf{0}$ equals the integral*

$$\frac{1}{2\pi i} \int_f \frac{dz}{z}.$$

   [*Hint:* If $z = g(s)$ is an arbitrary piecewise-differentiable path in $\mathbf{C} - \mathbf{0}$, consider the path $s \longrightarrow g(s)/\|g(s)\|$ in S^1 and let θ be a lifting of it to the covering space R. Then $g(s) = \|g(s)\| e^{2\pi i \theta(s)}$. Compute $\int_g dz/z$.]
   (c) Let $b_0 = 1 + 0i$. There are standard isomorphisms $\pi_1(\mathbf{C} - \mathbf{0}, x_0) \longrightarrow \pi_1(\mathbf{C} - \mathbf{0}, b_0) \longrightarrow \pi_1(S^1, b_0) \longrightarrow Z$. Show that the winding number of f equals the image of $[f]$ under these isomorphisms.
7. (a) Show that $R^3 - \mathbf{0}$ is simply connected. [*Hint:* Given a loop f in $R^3 - \mathbf{0}$ based at x_0, first show that f is path homotopic to a loop made up of finitely many straight-line segments, none of which are coplanar with the line segment joining x_0 and the origin.]
   (b) Show that S^2 is simply connected.

8-6 The Fundamental Group of S^n

Now we compute the fundamental group of S^n, showing that S^n is simply connected if $n \geq 2$. A proof for the case $n = 2$ was outlined in Exercise 7 of the preceding section. Here is a different proof, which uses a theorem we shall need later anyway.

Theorem 6.1 (The special Van Kampen theorem). *Let* $X = U \cup V$, *where* U *and* V *are open in* X *and* $U \cap V$ *is path connected. Let* x_0 *be a point of* $U \cap V$. *If both inclusions*

$$i : (U, x_0) \longrightarrow (X, x_0) \quad \textit{and} \quad j : (V, x_0) \longrightarrow (X, x_0)$$

induce zero homomorphisms of fundamental groups, then $\pi_1(X, x_0) = 0$.

Note that both i_* and j_* are necessarily zero homomorphisms if U and V are simply connected. This is the case we need for the following theorem. We shall need the more general result in proving the Jordan separation theorem (§8-12).

Proof. Let $f : I \longrightarrow X$ be a loop based at x_0. We wish to show that f is path homotopic to a constant loop.

Step 1. By the Lebesgue number lemma, there is a subdivision

$$0 = a_0 < a_1 < \cdots < a_n = 1$$

of the interval [0, 1] such that for each i, the set $f([a_{i-1}, a_i])$ lies entirely in one of the open sets U or V. Among all such subdivisions, choose one for which the number n of subintervals is *minimal.* Then it follows that for each i, the point $f(a_i)$ lies in $U \cap V$.

Suppose that $f(a_i) \notin U$, for instance. Then neither $f([a_{i-1}, a_i])$ nor $f([a_i, a_{i+1}])$ lies entirely in U. Therefore, both of them must lie entirely in V. We can then discard a_i from the subdivision, and still have a subdivision of [0, 1] for which the image of each subinterval lies either in U or in V. This contradicts minimality. Hence $f(a_i)$ must belong to U.

Step 2. Consider the restriction of f to the interval $[a_{i-1}, a_i]$. Reparametrize it so that it is based on the interval [0, 1], defining

$$f_i(s) = f((1 - s)a_{i-1} + sa_i) \quad \text{for } s \in [0, 1].$$

We show that f_i is path homotopic to a path that lies entirely in U.

Of course, if f_i itself lies in U, nothing needs to be done. We define F_i to be the trivial path homotopy $F_i(s, t) = f_i(s)$ of f_i to itself.

If f_i does not lie in U, then f_i must lie entirely in V. Using path connectivity of $U \cap V$, let us choose paths g and h in $U \cap V$ from the base point x_0 to the

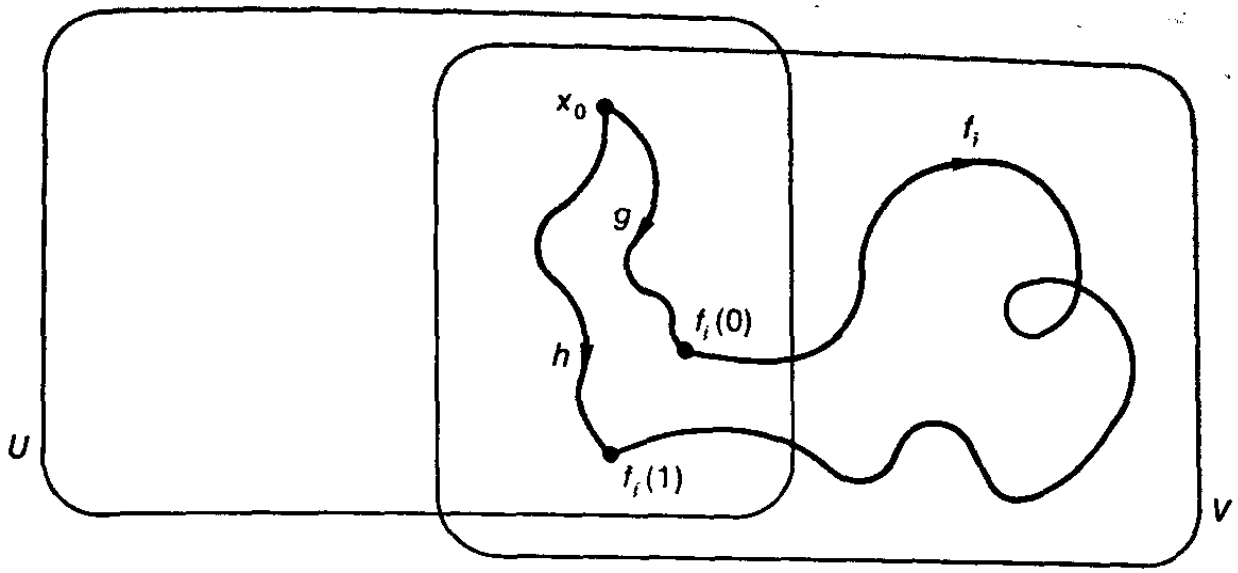


Figure 20

points $f_i(0)$ and $f_i(1)$, respectively. (See Figure 20.) Consider the composition $(g * f_i) * \bar{h}$; it is a loop in V based at x_0. Under the inclusion map j, this loop goes into a loop in X based at x_0. Since j_* is the zero homomorphism, this loop is path homotopic, in X, to the constant loop e_{x_0}. Then f_i is path homotopic (in X) to the path $\bar{g} * h$ (using the groupoid properties of $*$ proved in Theorem 1.2). Let F_i be this path homotopy; it is a path homotopy (in X) between f_i and a path lying in $U \cap V$.

Now we reparametrize each path homotopy F_i so as to get a map of $[a_{i-1}, a_i] \times I$ into X, and then we paste the pieces together to get a path homotopy F between f and a path that lies entirely in U. Specifically, define $F: I \times I \longrightarrow X$ by the equation

$$F(s, t) = F_i\left(\frac{s - a_{i-1}}{a_i - a_{i-1}}, t\right) \quad \text{for } s \in [a_{i-1}, a_i].$$

Because each path homotopy F_i leaves the end points fixed, F is well-defined; the pasting lemma shows it is continuous. Let $f'(s) = F(s, 1)$.

Step 3. We have shown f is path homotopic (in X) to a loop f' that lies entirely in U. Because

$$i_*: \pi_1(U, x_0) \longrightarrow \pi_1(X, x_0)$$

is the zero homomorphism, the loop f' must be path homotopic (in X) to the constant loop e_{x_0}. Thus $f \simeq_p f' \simeq_p e_{x_0}$, as loops in X. This is what wished to prove. □

This theorem is a special case of a famous theorem of topology called the *Van Kampen theorem*, which expresses the fundamental group of the space X quite generally in terms of the fundamental groups of U and V and the homomorphisms induced by the inclusions of $U \cap V$ into U and V, provided that $U \cap V$ is path connected. (See [M].)

Theorem 6.2. For $n \geq 2$, *the* n*-sphere* S^n *is simply connected.*

Proof. Let $p = (0, \ldots, 0, 1) \in R^{n+1}$ be the "north pole" of S^n; let $q = (0, \ldots, 0, -1)$ be the "south pole."

Step 1. First, we show that $S^n - p$ is homeomorphic with R^n.

Let $x = (x_1, \ldots, x_{n+1})$ be a point of S^n different from p. Suppose that we take the straight line in R^{n+1} determined by x and p and intersect it with the plane $x_{n+1} = 0$; let $f(x)$ denote this point of intersection. The resulting map $f: S^n - p \longrightarrow R^n$ is called **stereographic projection.**

More precisely, we define f by the equation

$$f(x) = \frac{1}{1 - x_{n+1}}(x_1, \ldots, x_n).$$

Obviously f is continuous. To show that f is a homeomorphism, one checks that the map $g : R^n \longrightarrow S^n - p$ given by

$$g(y_1, \ldots, y_n) = (ty_1, ty_2, \ldots, ty_n, 1 - t),$$

where $t = 2/(1 + (y_1)^2 + \cdots + (y_n)^2)$, is an inverse for f.

Since $S^n - q$ is homeomorphic to $S^n - p$ under the reflection map

$$\rho(x_1, \ldots, x_n, x_{n+1}) = (x_1, \ldots, x_n, -x_{n+1}),$$

$S^n - q$ is also homeomorphic to R^n.

Step 2. Now let $U = S^n - p$ and $V = S^n - q$. The space S^n equals the union of the open sets U and V. Furthermore, the sets U and V are simply connected, being homeomorphic with R^n. Once we show that $U \cap V$ is path connected, we can conclude from Theorem 6.1 that the space S^n is simply connected.

The intersection of U and V is the set $S^n - p - q$. To show that this set is path connected, note first that it is homeomorphic with $R^n - \mathbf{0}$ under stereographic projection. It is easy to show that $R^n - \mathbf{0}$ is path connected: Any point x of $R^n - \mathbf{0}$ can be joined to the point $x_0 = (1, 0, \ldots, 0)$ by the straight-line path in $R^n - \mathbf{0}$, except for points x of the form $(a, 0, \ldots, 0)$, where $a < 0$. In that case, we can take the straight-line path from x to $x_1 = (0, 1, 0, \ldots, 0)$, followed by the straight-line path from x_1 to x_0. (Here we use, of course, the fact that $n > 1$.) □

Corollary 6.3. $\mathrm{R}^n - \mathbf{0}$ *is simply connected if* $n > 2$.

Proof. By Theorem 5.2, $R^n - \mathbf{0}$ and S^{n-1} have isomorphic fundamental groups. □

Corollary 6.4. R^n *and* R^2 *are not homeomorphic for* $n > 2$.

Proof. Deleting a point from R^n leaves a simply connected space, while deleting a point from R^2 does not. □

This corollary has a generalization to higher dimensions; R^n and R^m are not homeomorphic if $n \neq m$. But the proof requires more tools of algebraic topology than we have developed.

Exercises

1. Let X be the union of two copies of S^2 having a point in common. What is the fundamental group of X? Prove that your answer is correct. [Be careful! The union of two simply connected spaces having a point in common is not necessarily simply connected. See [S], p. 59.]
2. Let X be the union of two open sets U and V; let $U \cap V$ be path connected; let $x_0 \in U \cap V$. Let i and j be the inclusion maps, as in Theorem 6.1. If you are given that j_* is the zero homomorphism, what can you say about i_*?
3. Criticize the following "proof" that S^2 is simply connected: Let f be a loop in S^2 based at x_0. Choose a point p of S^2 not lying in the image of f. Since $S^2 - p$ is homeomorphic with R^2, and R^2 is simply connected, the loop f is path homotopic to the constant loop.

8-7 Fundamental Groups of Surfaces

Recall that a *surface* is a Hausdorff space with a countable basis, every point of which has a neighborhood that is homeomorphic with an open subset of R^2. Surfaces are of interest in various parts of mathematics, including geometry, topology, and complex analysis. What we do here is to consider four familiar surfaces—the sphere S^2, the projective plane P^2, the torus T, and the double torus T_2—and show by comparing their fundamental groups that these surfaces are not homeomorphic.

One can, in fact, use the fundamental group to classify all compact surfaces completely; but this we shall not attempt. The interested reader is referred to Chapter 4 of [M].

First we consider the torus. To compute its fundamental group, we shall need a theorem about the fundamental group of a product space.

Theorem 7.1. $\pi_1(X \times Y, x_0 \times y_0)$ *is isomorphic with* $\pi_1(X, x_0) \times \pi_1(Y, y_0)$.

Proof. Recall that if A and B are groups with operation $\cdot$, then the cartesian product $A \times B$ is given a group structure by using the operation

$$(a \times b)\cdot(a' \times b') = (a \cdot a') \times (b \cdot b').$$

Recall also that if $h : C \to A$ and $k : C \to B$ are group homomorphisms, then the map $\Phi : C \to A \times B$ defined by $\Phi(c) = h(c) \times k(c)$ is a group homomorphism.

Now let $p : X \times Y \to X$ and $q : X \times Y \to Y$ be the projection mappings. If we use the base points indicated in the statement of the theorem, we

have induced homomorphisms

$$p_*: \pi_1(X \times Y, x_0 \times y_0) \longrightarrow \pi_1(X, x_0),$$
$$q_*: \pi_1(X \times Y, x_0 \times y_0) \longrightarrow \pi_1(Y, y_0).$$

We define a homomorphism

$$\Phi: \pi_1(X \times Y, x_0 \times y_0) \longrightarrow \pi_1(X, x_0) \times \pi_1(Y, y_0)$$

by the equation

$$\Phi([f]) = p_*([f]) \times q_*([f]) = [p \circ f] \times [q \circ f].$$

We shall show that Φ is an isomorphism.

The map Φ is surjective. Let $g: I \longrightarrow X$ be a loop based at x_0; let $h: I \longrightarrow Y$ be a loop based at y_0. We wish to show that the element $[g] \times [h]$ lies in the image of Φ. Define $f: I \longrightarrow X \times Y$ by the equation

$$f(s) = g(s) \times h(s).$$

Then f is a loop in $X \times Y$ based at $x_0 \times y_0$, and

$$\Phi([f]) = [p \circ f] \times [q \circ f] = [g] \times [h],$$

as desired.

The kernel of Φ vanishes. Suppose that $f: I \longrightarrow X \times Y$ is a loop in $X \times Y$ based at $x_0 \times y_0$, and $\Phi([f]) = [p \circ f] \times [q \circ f]$ is the identity element. This means that $p \circ f \simeq_p e_{x_0}$ and $q \circ f \simeq_p e_{y_0}$; let G and H be the respective path homotopies. Then the map $F: I \times I \longrightarrow X \times Y$ defined by

$$F(s, t) = G(s, t) \times H(s, t)$$

is a path homotopy between f and the constant loop based at $x_0 \times y_0$. □

Corollary 7.2. The fundamental group of the torus $\mathrm{T} = \mathrm{S}^1 \times \mathrm{S}^1$ is isomorphic to the group $\mathrm{Z} \times \mathrm{Z}$.

Now we define the projective plane and compute its fundamental group.

Definition. The **projective plane** P^2 is the space obtained from S^2 by identifying each point x of S^2 with its antipodal point $-x$.

Formally, define an equivalence relation on S^2 by setting $x \sim x$ and $x \sim (-x)$; then P^2 is the set of equivalence classes. If $p: S^2 \longrightarrow P^2$ maps each point x to its equivalence class, we topologize P^2 by defining V to be open in P^2 if and only if $p^{-1}(V)$ is open in S^2.

The projective plane is the fundamental object of study in projective goemetry, just as the euclidean plane R^2 is in ordinary euclidean geometry. Topologists are primarily interested in it as an example of a surface.

Theorem 7.3. *The projective plane P^2 is a surface, and the map $\mathrm{p}: \mathrm{S}^2 \longrightarrow \mathrm{P}^2$ is a covering map.*

Proof. First we show that p is an open map. Let U be open in S^2. Now the antipodal map $a : S^2 \longrightarrow S^2$ given by $a(x) = -x$ is a homeomorphism of S^2; hence $a(U)$ is open in S^2. Since

$$p^{-1}(p(U)) = U \cup a(U),$$

this set also is open in S^2. Therefore, by definition, $p(U)$ is open in P^2.

Now we show that p is a covering map. Given a point y of P^2, choose $x \in p^{-1}(y)$. Then choose an ϵ-neighborhood U of x in S^2 for some $\epsilon < 1$, using the euclidean metric d of R^3. Then U contains no pair $\{z, a(z)\}$ of antipodal points of S^2, since $d(z, a(z)) = 2$. As a result, the map

$$p : U \longrightarrow p(U)$$

is bijective. Being continuous and open, it is a homeomorphism. Similarly,

$$p : a(U) \longrightarrow p(a(U)) = p(U)$$

is a homeomorphism. The set $p^{-1}(p(U))$ is thus the union of the two disjoint open sets U and $a(U)$, each of which is mapped homeomorphically by p onto $p(U)$.

Then $p(U)$ is a neighborhood of $p(x) = y$ that is evenly covered by p. Hence p is a covering map.

Since S^2 has a countable basis $\{U_n\}$, the space P^2 has a countable basis $\{p(U_n)\}$.

We show P^2 is Hausdorff. Let y_1 and y_2 be two points of P^2. The set $p^{-1}(y_1) \cup p^{-1}(y_2)$ consists of four points; let 2ϵ be the minimum distance between them. Let U_1 be the ϵ-neighborhood of one of the points of $p^{-1}(y_1)$, and let U_2 be the ϵ-neighborhood of one of the points of $p^{-1}(y_2)$. Then

$$U_1 \cup a(U_1) \quad \text{and} \quad U_2 \cup a(U_2)$$

are disjoint. It follows that $p(U_1)$ and $p(U_2)$ are disjoint neighborhoods of y_1 and y_2, respectively, in P^2.

Since S^2 is a surface and every point of P^2 has a neighborhood homeomorphic with an open subset of S^2, the space P^2 is also a surface. □

Corollary 7.4. $\pi_1(P^2, y)$ *is a group of order* 2.

Proof. The projection $p : S^2 \longrightarrow P^2$ is a covering map. Since S^2 is simply connected, we can apply Theorem 4.5, which tells us there is a bijective correspondence between $\pi_1(P^2, y)$ and the set $p^{-1}(y)$. Since this set is a two-element set, $\pi_1(P^2, y)$ is a group of order 2.

Any group of order 2 is isomorphic to Z_2, the integers mod 2, of course. □

One can proceed similarly to define P^n, for any $n \in Z_+$, as the space obtained from S^n by identifying each point x with its antipode $-x$; it is called **projective n-space**. The proof of Theorem 7.3 goes through without change to prove that the projection $p : S^n \longrightarrow P^n$ is a covering map. Then because S^n is

simply connected for $n \geq 2$, it follows that $\pi_1(P^n, y)$ is a two-element group for $n \geq 2$. We leave it to you to figure out what happens when $n = 1$.

We have not yet found any space whose fundamental group is not abelian. Let us remedy that deficiency now.

Lemma 7.5. The fundamental group of the figure eight is not abelian.

Proof. The figure eight is the union of two circles A and B with a point x_0 in common. We now describe a certain covering space E for the figure eight.

The space E is the subspace of the plane consisting of the x-axis and the y-axis, along with tiny circles tangent to these axes, one circle tangent to the x-axis at each nonzero integer point and one circle tangent to the y-axis at each nonzero integer point. See Figure 21.

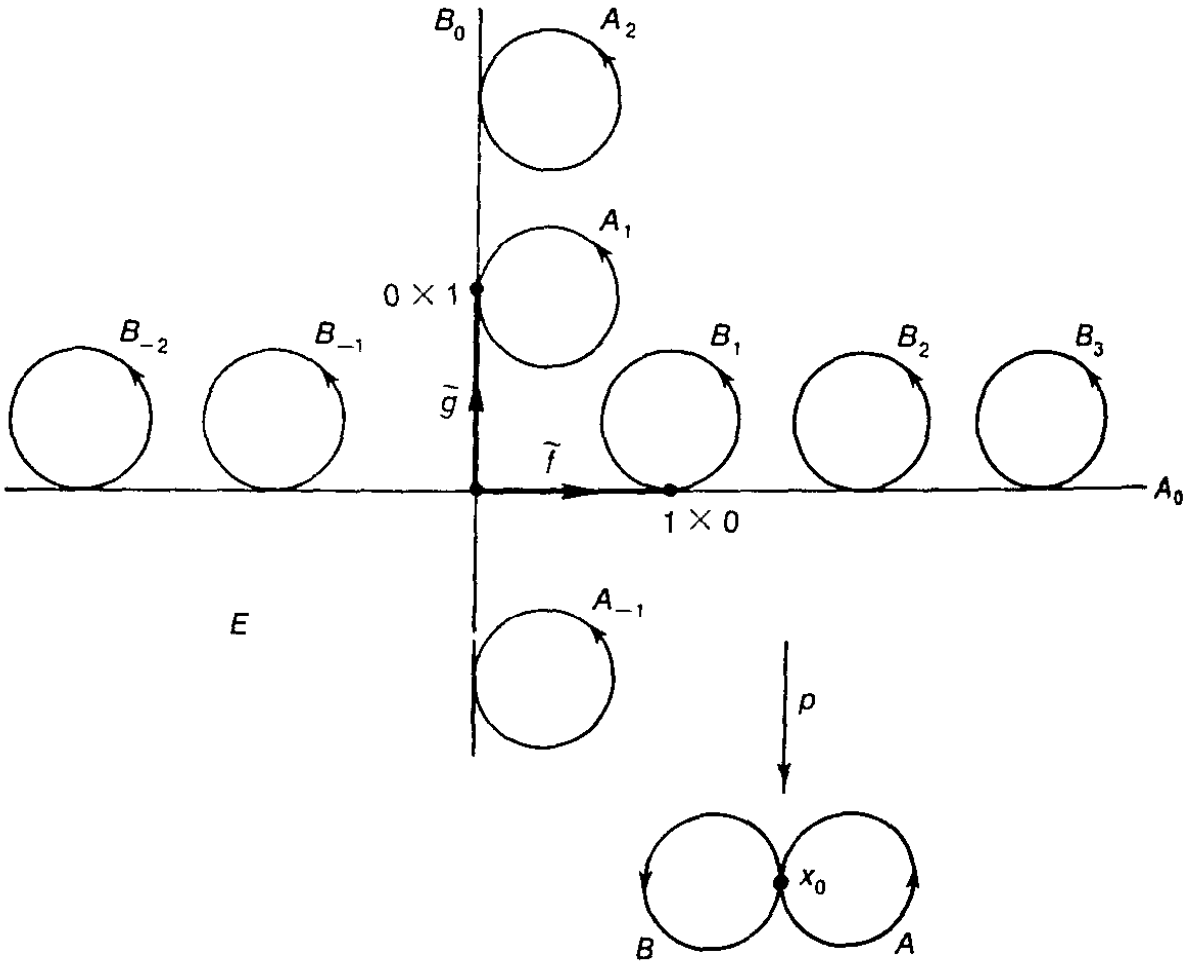


Figure 21

The projection map p wraps the x-axis around the circle A and wraps the y-axis around the other circle B; in each case the integer points are mapped by p into the base point x_0. Each circle tangent to an integer point on the x-axis is mapped homeomorphically by p onto B, while each circle tangent to an integer point on the y-axis is mapped homeomorphically onto A; in each

case the point of tangency is mapped onto the point x_0. We leave it to you to check mentally that the map p is indeed a covering map.

We could write this description down in equations if we wished, but the informal description seems to us easier to follow.

Now let $\tilde{f}: I \longrightarrow E$ be the path $\tilde{f}(s) = s \times 0$, going along the x-axis from the origin to the point 1×0. Let $\tilde{g}: I \longrightarrow E$ be the path $\tilde{g}(s) = 0 \times s$, going along the y-axis from the origin to the point 0×1. Let $f = p \circ \tilde{f}$ and $g = p \circ \tilde{g}$; then f and g are loops in the figure eight based at x_0, going around the circles A and B, respectively. We assert that $f * g$ and $g * f$ are not path homotopic, so that the fundamental group of the figure eight is not abelian.

To prove this assertion, let us lift each of these to a path in E beginning at the origin. The path $f * g$ lifts to a path that goes along the x-axis from the origin to 1×0, and then goes once around the circle tangent to the x-axis at 1×0. On the other hand, the path $g * f$ lifts to a path in E that goes along the y-axis from the origin to 0×1, and then goes once around the circle tangent to the y-axis at 0×1. Since the lifted paths do not end at the same point, $f * g$ and $g * f$ cannot be path homotopic. □

The fundamental group of the figure eight is, in fact, the group that algebraists call the "free group on two generators." But this we shall not prove.

Theorem 7.6. *The fundamental group of the double torus T_2 is not abelian.*

Proof. The double torus T_2 is the surface obtained by taking two copies of the torus, deleting a small open disc from each of them, and pasting the remaining pieces together along their edges. We assert that the figure eight is a retract of T_2. This fact implies that inclusion j induces an injective homomorphism

$$j_*: \pi_1(8, x_0) \longrightarrow \pi_1(T_2, x_0),$$

so that $\pi_1(T_2, x_0)$ is not abelian.

One can write equations for the retraction $r: T_2 \longrightarrow 8$, but it is simpler to indicate it in pictures, as we have done in Figure 22. First one collapses the

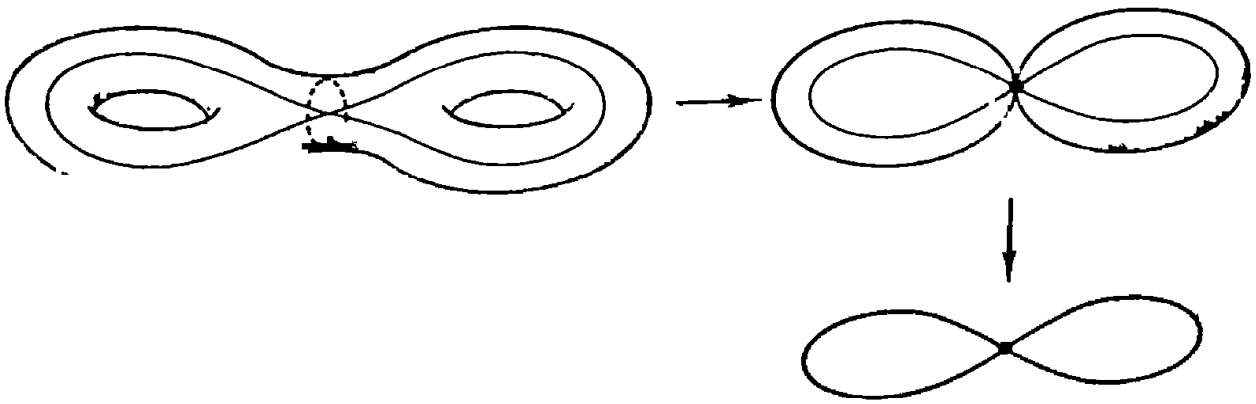


Figure 22

dotted circle to a point, obtaining two tori with a point in common. Then one collapses each of the two tori onto their "meridianal circles." □

Corollary 7.7. The surfaces S^2, P^2, T, *and* T_2 *are topologically distinct.*

Exercises

1. Compute the fundamental groups of the "solid torus" $S^1 \times B^2$ and the product space $S^1 \times S^2$.
2. Let X be the quotient space obtained from B^2 by identifying each point x of S^1 with its antipode $-x$. Show that X is homeomorphic to the projective plane P^2.
3. Let $p : E \longrightarrow 8$ be the map constructed in the proof of Lemma 7.5. Let E' be the subspace of E which is the union of the x-axis and the y-axis. Show that $p|E'$ is not a covering map.
4. Consider the covering map indicated in Figure 23. Here p wraps A_1 around A twice and wraps B_1 around B twice; p maps A_0 and B_0 homeomorphically onto A and B, respectively. Use this covering space to show that the fundamental group of the figure eight is not abelian.

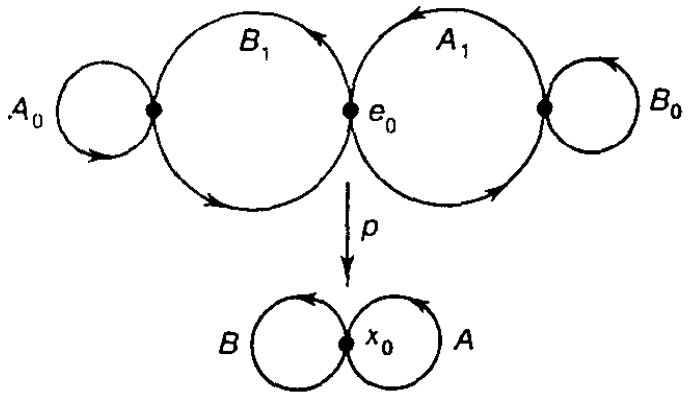


Figure 23

5. Given a group G and a space X, an **action** of G on X is a function assigning, to each element α of G, a continuous map

$$h_\alpha : X \longrightarrow X$$

in such a way that:

(i) If e is the identity element of G, then h_e is the identity map of X.

(ii) If $\alpha = \beta \cdot \gamma$, then $h_\alpha = h_\beta \circ h_\gamma$.

(a) Show that defining $h_n : R \longrightarrow R$ by the equation $h_n(x) = x + n$ defines an action of the integers Z on R.

(b) Show that rotation of S^2 about the z-axis defines an action of the group S^1 of complex numbers of absolute value 1 on the sphere S^2.

(c) Show that, given n, rotation of S^2 about the z-axis through the various angles which are multiples of $2\pi/n$ defines an action of Z_n (the additive group of integers mod n) on S^2.

(d) Show that the antipodal map of S^2 defines an action of the group Z_2 on S^2.

6. Given an action of G on X, define the **orbit space** X/G to be the quotient space of X determined by the following equivalence relation: $x \sim x'$ if $x' = h_\alpha(x)$ for some $\alpha \in G$. The orbit spaces under the actions defined in (a)–(d) of Exercise 5 are homeomorphic to familiar spaces. What are they?

7. An action of G on X is said to be **fixed point free** if the only map h_α that possesses a fixed point is the map h_e. Which of the actions defined in Exercise 5 are fixed point free?

8. *Theorem. Let there be given an action of the finite group G on the path-connected Hausdorff space X; assume that the action is fixed point free. If X is simply connected, then $\pi_1(X/G, x_0)$ is isomorphic to G.*

9. Consider S^3 as the space of all pairs of complex numbers (z_1, z_2) satisfying the equation $|z_1|^2 + |z_2|^2 = 1$. Given relatively prime positive integers n and k, define $h: S^3 \longrightarrow S^3$ by the equation

$$h(z_1, z_2) = (z_1 e^{2\pi l/n}, z_2 e^{2\pi i k/n}).$$

(a) Show that the maps h, $h^2 = h \circ h$, $h^3, \ldots$ can be used to define an action of Z_n on S^3 that is fixed point free. The orbit space $L(n, k)$ is called a **lens space**.
(b) Show that $L(n, k)$ is a compact 3-manifold.
(c) Assume Exercise 8. Show that if $L(n, k)$ and $L(n', k')$ are homeomorphic, then $n = n'$. (In fact, $L(n, k)$ is homeomorphic to $L(n', k')$ if and only if $n = n'$ and either $k \equiv k' \pmod n$ or $kk' \equiv 1 \pmod n$. The proof is decidedly nontrivial.)

8-8 Essential and Inessential Maps

We have already used the fundamental group to study the problem of determining whether or not two given spaces are homeomorphic. Now we apply it to a second problem of topology, that of determining whether or not two given maps of one space into another are homotopic.

The simplest such problem is to determine whether or not a given map $h: X \longrightarrow Y$ is homotopic to a constant map. We shall prove that if h is homotopic to a constant map, then the induced homomorphism h_* of fundamental groups is trivial. (The converse holds if X happens to be the circle S^1; see Exercise 2.)

We defer the more general problem of determining whether two arbitrary maps $h, k: X \longrightarrow Y$ are homotopic to a later section (§8-11).

Definition. A map $h: X \longrightarrow Y$ is said to be **inessential** if h is homotopic to a constant map. Otherwise, it is said to be **essential**.

Lemma 8.1. *Let* $\mathrm{h}: \mathrm{S}^1 \longrightarrow \mathrm{Y}$. *Then the following are equivalent:*

(1) h *is inessential.*

(2) h *can be extended to a continuous map* $\mathrm{g}: \mathrm{B}^2 \longrightarrow \mathrm{Y}$.

Proof. (1) $\Rightarrow$ (2). Suppose that $H: S^1 \times I \longrightarrow Y$ is a homotopy between h and the constant map k which carries S^1 to y_0. Let $\pi: S^1 \times I \longrightarrow B^2$ be the map

$$\pi(x, t) = (1 - t)x.$$

Then π carries $S^1 \times [0, 1)$ bijectively onto $B^2 - \mathbf{0}$, and it maps $S^1 \times 1$ to the point $\mathbf{0}$. (You can check this.) Because $S^1 \times I$ is compact and B^2 is Hausdorff, π is a closed map; that is, it carries closed sets of $S^1 \times I$ to closed sets of B^2.

Now the map $H: S^1 \times I \longrightarrow Y$ is constant on the set $S^1 \times 1$. Therefore, H induces a map $g: B^2 \longrightarrow Y$ such that $g \circ \pi = H$. [It is defined by letting $g(x) = H(\pi^{-1}(x))$ if $x = o$ and $g(x) = y_0$ if $x = o$.] The map g is continuous; this follows from the fact that π is a quotient map. [Or one can prove it directly by noting that if C is closed in Y, then $H^{-1}(C)$ is closed in $S^1 \times I$, so that $\pi(H^{-1}(C)) = g^{-1}(C)$ is closed in B^2.] The map g is the desired extension of h; for if $x \in S^1$, then

$$g(x) = g(\pi(x, 0)) = H(x, 0) = h(x).$$

See Figure 24.

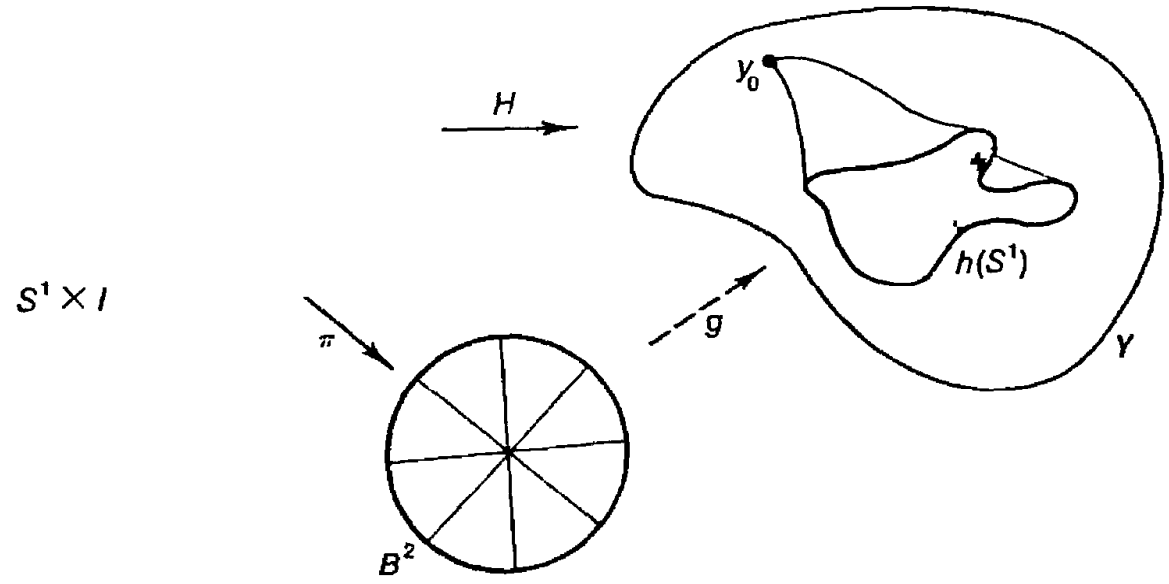


Figure 24

(2) $\Rightarrow$ (1). Let $g: B^2 \longrightarrow Y$ be a continuous extension of h. Let us define $F: S^1 \times I \longrightarrow Y$ by the equation $F(x, t) = g((1 - t)x)$. Then F is a homotopy between h and a constant map. □

Theorem 8.2. *Let* $\mathrm{h}: \mathrm{X} \longrightarrow \mathrm{Y}$. *If* h *is inessential, then* h_* *is the zero homomorphism.*

Proof. Consider first the case $X = S^1$. We apply the preceding lemma. Let $g: B^2 \to Y$ be an extension of h; let $j: S^1 \to B^2$ be the inclusion mapping. Then $g \circ j = h$. Let b_1 be a point of S^1 and let $y_1 = h(b_1)$. Consider the induced homomorphisms

$$\begin{array}{ccc} \pi_1(S^1, b_1) & \xrightarrow{h_*} & \pi_1(Y, y_1) \\ & {\scriptstyle j_*} \searrow \quad \nearrow {\scriptstyle g_*} & \\ & \pi_1(B^2, b_1) & \end{array}$$

By the functorial properties of the induced homomorphism, $h_* = g_* \circ j_*$. But j_* is the zero homomorphism, because its range is the trivial group. Therefore, h_* is the zero homomorphism.

Now we consider the general case. Let $f: I \to X$ be a loop in X based at x_0. Let $\phi: I \to S^1$ be the standard loop

$$\phi(s) = (\cos 2\pi s, \sin 2\pi s).$$

Now f induces a map $k: S^1 \to X$, defined by the equation $k(a) = f(\phi^{-1}(a))$.

$$\begin{array}{ccccc} I & \xrightarrow{f} & X & \xrightarrow{h} & Y \\ & {\scriptstyle \phi} \searrow \quad \nearrow {\scriptstyle k} & & & \\ & S^1 & & & \end{array}$$

It is easy to check that k is continuous. By hypothesis, h is homotopic to a constant; let $H: X \times I \to Y$ be the homotopy. Then $h \circ k$ is also homotopic to a constant; the map $H'(a, t) = H(k(a), t)$ is the required homotopy. It follows from the preceding paragraph that $(h \circ k)_*$ is the zero homomorphism. In particular, $(h \circ k)_*([\phi]) = 0$. But

$$(h \circ k)_*([\phi]) = [h \circ k \circ \phi] = [h \circ f] = h_*([f]). \quad \square$$

As an application, we prove the theorem we used in §7-9, when we computed the topological dimension of a triangular region.

Corollary 8.3. *Let* T *be a closed triangular region in* R^2*; let* Bd T *denote the union of the edges of* T*. There is no continuous map* $\mathrm{f}: \mathrm{T} \to \mathrm{Bd\ T}$ *that maps each edge of* T *into itself.*

Proof. We suppose there is such a continuous map $f: T \to \mathrm{Bd}\ T$, and derive a contradiction.

Let $g: \mathrm{Bd}\ T \to \mathrm{Bd}\ T$ be the restriction of f. Because g maps each edge of T into itself, g is homotopic to the identity map. Indeed, the map

$$G(x, t) = tx + (1 - t)g(x)$$

maps $\mathrm{Bd}\ T \times I$ into Bd T; for if x belongs to Bd T, then x and $g(x)$ lie in the

same edge of T, so that the line segment joining them lies in this edge as well. Thus G is a homotopy between g and the identity.

Now we use the fact that Bd T is homeomorphic to S^1. The identity map of Bd T does not induce the zero homomorphism of the fundamental group, so that it cannot be homotopic to a constant, by the preceding theorem. Therefore, g is not homotopic to a constant.

On the other hand, g is extendable to the continuous map f of T into Bd T, so that g *is* homotopic to a constant. □

Exercises

1. Here is an easy "proof" that if $h: X \longrightarrow Y$ is inessential, then h_* is the zero homomorphism. Where is the fallacy? Let $H: X \times I \longrightarrow Y$ be a homotopy between h and a constant map. Given a loop f in X based at x_0, define a map $F: I \times I \longrightarrow Y$ by the equation $F(s, t) = H(f(s), t)$. The map F is a homotopy between $h \circ f$ and a constant path, as desired.

2. Let $h: S^1 \longrightarrow Y$. Show that if h_* is the zero homomorphism, then h is inessential. [*Hint:* Let $\phi: I \longrightarrow S^1$ be the standard loop; let $f = h \circ \phi$. Show there is a path homotopy F between f and e_{y_0}, and that F induces a map $H: S^1 \times I \longrightarrow Y$ such that $H \circ (\phi \times i_I) = F$. See Figure 25.]

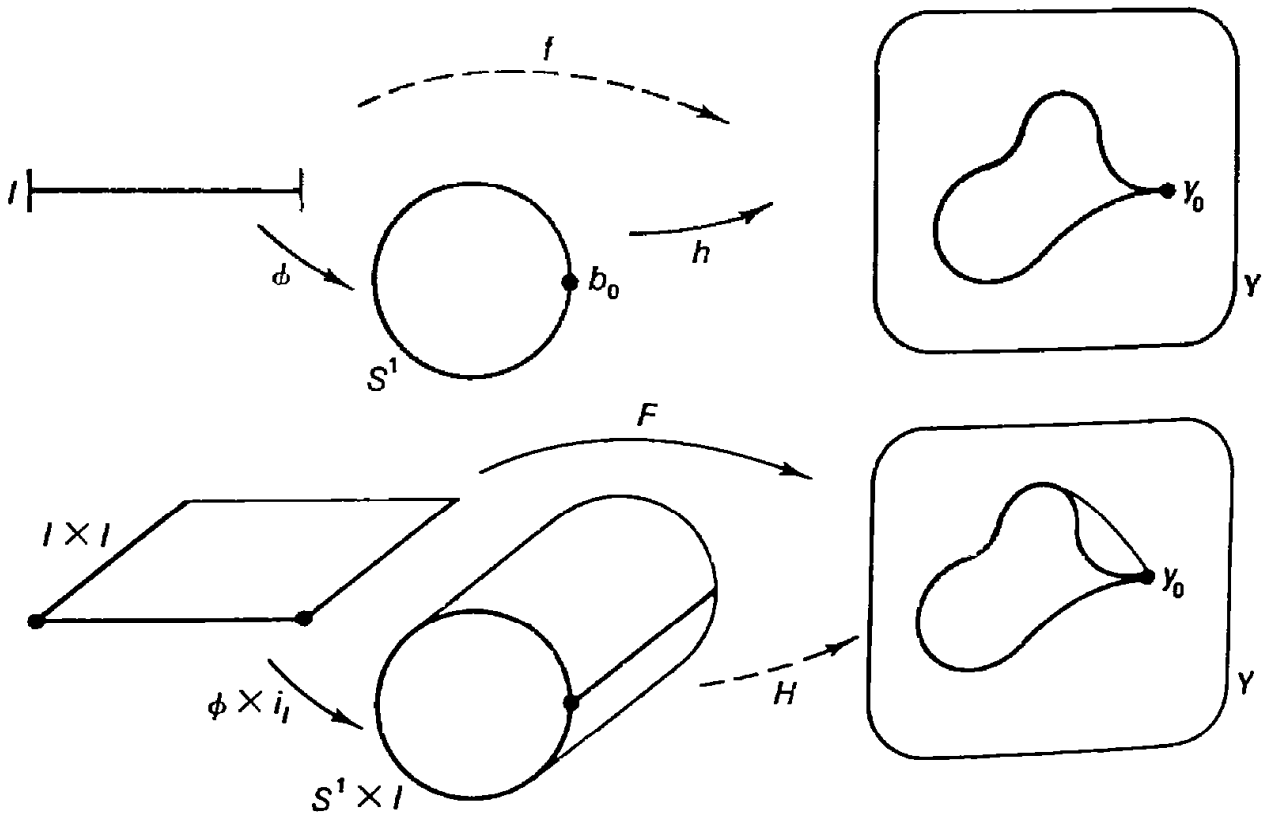


Figure 25

3. Let Y be path connected. Show that $\pi_1(Y, y_0) = 0$ if and only if every map $h: S^1 \longrightarrow Y$ is inessential.

4. Recall that a space Y is said to be *contractible* if the identity map $i_Y : Y \longrightarrow Y$ is inessential. Show that if Y is contractible, then Y is simply connected.

5. A map $h : S^n \longrightarrow S^m$ is said to be **antipode-preserving** if $h(-x) = -h(x)$ for every $x \in S^n$.

Theorem. If $h : S^1 \longrightarrow S^1$ is antipode-preserving and continuous, then h is essential.

(a) Let $p : S^1 \longrightarrow S^1$ be the map $p(z) = z^2$, where z is a complex number. Show that h induces a continuous map $g : S^1 \longrightarrow S^1$ such that $p \circ h = g \circ p$.

(b) Show that if f is any path in S^1 from a point x to its antipode $-x$, then $p \circ f$ is a loop in S^1 that is not path homotopic to a constant.

(c) Show that both p and g induce monomorphisms (injective homomorphisms) of the fundamental groups.

(d) Conclude that h is essential.

6. Assume Exercise 5.

(a) Prove the following:

Theorem (Borsuk–Ulam theorem for S^2). There is no continuous antipode-preserving map $f : S^2 \longrightarrow S^1$.

[*Hint:* Consider the equator in S^2.] It is in general true that there is no continuous antipode-preserving map $f : S^n \longrightarrow S^m$ if $m < n$. But the proof requires more tools than we now possess.

(b) Prove:

Theorem. Given a continuous map $f : S^2 \longrightarrow R^2$, there is a point x of S^2 such that $f(x) = f(-x)$.

[*Hint:* Consider $(f(x) - f(-x))/\|f(x) - f(-x)\|$.]

(c) Prove:

Theorem (A "theorem of meteorology"). At any given moment in time, there exists a pair of antipodal points on the surface of the earth at which both the temperature and the barometric pressure are equal.

(d) Prove:

Theorem. If $g : S^2 \longrightarrow S^2$ is continuous and $g(x) \neq g(-x)$ for all x, then g is surjective.

*7. Generalize Exercise 5 as follows: Assume that $h : S^1 \longrightarrow S^1$ is antipode-preserving; let $h(x_0) = x_1$. Then h_* maps a generator of $\pi_1(S^1, x_0)$ to an *odd* multiple of a generator of $\pi_1(S^1, x_1)$.

8-9 The Fundamental Theorem of Algebra

It is a basic fact about the complex numbers that every polynomial equation

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

of degree n with real or complex coefficients has n roots (if the roots are counted according to their multiplicities). You probably first were told this fact in high school algebra, although it is doubtful that it was proved for you at that time.

The proof is, in fact, rather hard; the most difficult part is to prove that every polynomial equation of positive degree has *at least one* root. There are various ways of doing this. One can use only techniques of algebra; this proof is long and arduous. Or one can develop the theory of analytic functions of a complex variable to the point where it becomes a trivial corollary of Liouville's theorem. Or one can prove it as a relatively easy corollary of our computation of the fundamental group of the circle; this we do now.

Theorem 9.1 (The fundamental theorem of algebra). *A polynomial equation*

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0$$

of degree $n > 0$ *with real or complex coefficients has at least one (real or complex) root.*

Proof. Step 1. Consider the map $h : S^1 \longrightarrow S^1$ given by $h(z) = z^n$, where z is a complex number. We show that the induced homomorphism

$$h_* : \pi_1(S^1, b_0) \longrightarrow \pi_1(S^1, b_0)$$

carries a generator of this infinite cyclic group to n times itself.

Let $\phi : I \longrightarrow S^1$ be the standard loop

$$\phi(s) = (\cos 2\pi s, \sin 2\pi s) = e^{2\pi i s}$$

in S^1. Its image under h_* is the loop

$$h(\phi(s)) = (e^{2\pi i s})^n = e^{2\pi i n s} = (\cos 2\pi n s, \sin 2\pi n s).$$

This loop lifts to the path $s \longrightarrow ns$ in the covering space R. Therefore, the loop $h \circ \phi$ corresponds to the integer n under the standard isomorphism of $\pi_1(S^1, b_0)$ with the integers, whereas ϕ corresponds to the number 1.

Step 2. Now we prove a special case of the theorem. Given a polynomial equation

$$z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0 = 0,$$

we assume that

$$|a_{n-1}| + \cdots + |a_1| + |a_0| < 1$$

and show that the equation has a root lying in the unit ball B^2.

Assume that the equation has no root in B^2. Then we can define a map $g : B^2 \longrightarrow R^2 - 0$ by the equation

$$g(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0.$$

Let $f : S^1 \longrightarrow R^2 - 0$ be the restriction of g to S^1. Because f is extendable to the map g of B^2 into $R^2 - 0$, the map f is inessential, by Lemma 8.1.

On the other hand, f is homotopic to the map $k : S^1 \longrightarrow R^2 - 0$ defined by $k(z) = z^n$. For the homotopy $F : S^1 \times I \longrightarrow R^2 - 0$ defined by

$$F(z, t) = z^n + t(a_{n-1}z^{n-1} + \cdots + a_1z + a_0)$$

is the required homotopy; $F(z, t)$ never vanishes because

$$\begin{aligned} |F(z,t)| &\geq |z^n| - |t(a_{n-1}z^{n-1} + \cdots + a_0)| \\ &\geq 1 - t(|a_{n-1}z^{n-1}| + \cdots + |a_0|) \\ &= 1 - t(|a_{n-1}| + \cdots + |a_0|) > 0. \end{aligned}$$

Furthermore, the map k is essential. For k equals the composite of the map $h: S^1 \to S^1$ of Step 1, given by $h(z) = z^n$, and the inclusion map $j: S^1 \to R^2 - 0$. Since h_* is "multiplication by n" and j_* is an isomorphism, k_* is not the zero homomorphism. Therefore, k must be essential.

Since f is homotopic to k, the map f also must be essential. Thus we reach a contradiction.

Step 3. Now we prove the general case. Given a polynomial equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0;$$

let us choose a real number $c > 0$ and substitute $x = cy$. We obtain the equation

$$(cy)^n + a_{n-1}(cy)^{n-1} + \cdots + a_1(cy) + a_0 = 0$$

or

$$y^n + \frac{a_{n-1}}{c}y^{n-1} + \cdots + \frac{a_1}{c^{n-1}}y + \frac{a_0}{c^n} = 0.$$

If this equation has the root $y = y_0$, then the original equation has the root $x_0 = cy_0$. So we need merely choose c large enough that

$$\left|\frac{a_{n-1}}{c}\right| + \left|\frac{a_{n-2}}{c^2}\right| + \cdots + \left|\frac{a_1}{c^{n-1}}\right| + \left|\frac{a_0}{c^n}\right| < 1$$

to reduce the theorem to the special case considered in Step 2. □

Exercises

1. Given a polynomial equation

$$x^n + a_{n-1}x^{n-1} + \cdots + a_1x + a_0 = 0$$

with real or complex coefficients. Show that if $|a_{n-1}| + \cdots + |a_1| + |a_0| < 1$, then *all* the roots of the equation lie interior to the unit ball B^2. [*Hint:* Let $g(x) = 1 + a_{n-1}x + \cdots + a_1x^{n-1} + a_0x^n$, and show that $g(x) \neq 0$ for $x \in B^2$.]

2. Find a circle about the origin containing all the roots of the polynomial equation $x^7 + x^2 + 1 = 0$.

8-10 Vector Fields and Fixed Points

In this section we apply the fundamental group to two problems of geometry. One problem concerns the existence of vector fields tangent to given surfaces. The second deals with the "fixed-point problem": Given X, does every continuous map $f: X \to X$ necessarily have a fixed point?

We shall obtain only some of the simpler results. Deeper theorems, including some of current research interest, demand much more of the machinery of algebraic topology than we have studied.

Theorem 10.1 Given a nonvanishing vector field on B^2, *there exists a point of* S^1 *where the vector field points directly inward and a point of* S^1 *where it points directly outward.*

Proof. A **vector field** on B^2 is an ordered pair $(x, v(x))$, where x is in B^2 and v is a continuous map of B^2 into R^2. In calculus, one often uses the notation

$$\mathbf{v}(x) = v_1(x)\mathbf{i} + v_2(x)\mathbf{j}$$

for the function v, where $\mathbf{i}$ and $\mathbf{j}$ are the standard unit basis vectors in R^2. But we shall stick with simple functional notation. To say that a vector field is *nonvanishing* means that $v(x) \neq \mathbf{0}$ for every x; in such a case v actually maps B^2 into $R^2 - \mathbf{0}$.

We show first that given v, it must point directly inward at some point of S^1.

Consider the map $w: S^1 \to R^2 - \mathbf{0}$ obtained by restricting v to S^1. If there is no point x of S^1 at which the vector field points directly inward, then for no x in S^1 is $w(x)$ equal to a negative multiple of x. It follows that w is homotopic to the inclusion map $j: S^1 \to R^2 - \mathbf{0}$, for the map $F: S^1 \times I \to R^2 - \mathbf{0}$ given by the equation

$$F(x, t) = tx + (1 - t)w(x)$$

is the required homotopy. It is pictured in Figure 26. Obviously, F is continuous. To show that F never vanishes, note that if $F(x, t) = \mathbf{0}$, then

$$(1 - t)w(x) = -tx.$$

This equation is clearly false for $t = 0$ or $t = 1$, since $x \in S^1$ and $w(x) \neq \mathbf{0}$. For $0 < t < 1$, it says that $w(x) = -tx/(1 - t)$, so that $w(x)$ equals a negative multiple of x, which is forbidden.

Since w is homotopic to the inclusion map $j: S^1 \to R^2 - \mathbf{0}$, it must be essential. On the other hand, w is extendable to the continuous map $v: B^2 \to R^2 - \mathbf{0}$, so that it is inessential. Thus we arrive at a contradiction. Therefore, v points directly inward at some point of S^1.

Consider now the nonvanishing vector field $(x, -v(x))$. By the result just

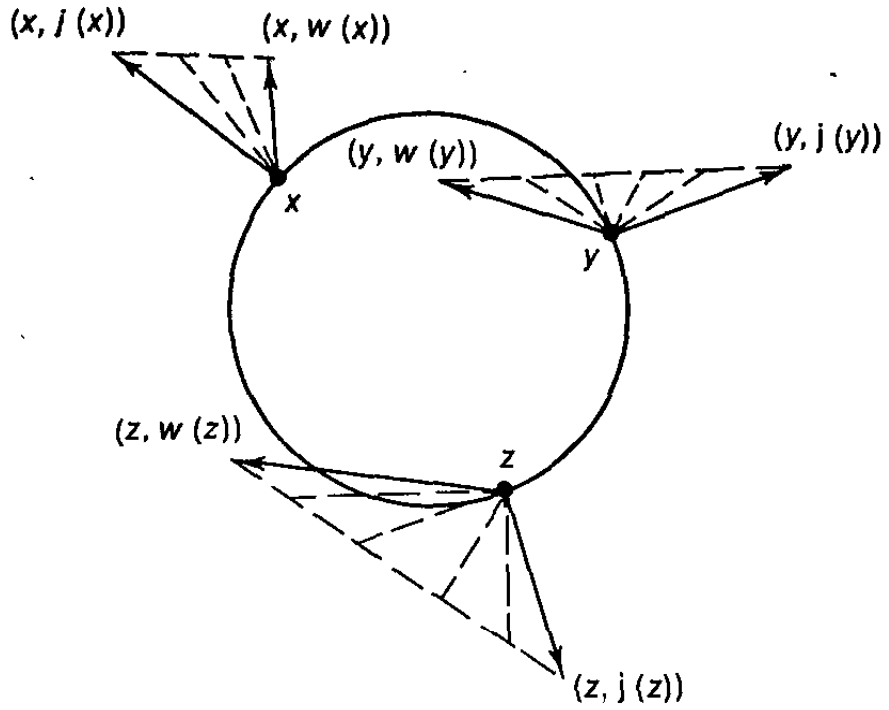


Figure 26

proved, it points directly inward at some point of S^1. Then v points directly outward at that point. □

We have already seen that every continuous map $f: [0, 1] \longrightarrow [0, 1]$ necessarily has a fixed point (see Exercise 3 of §3-2). The same is true for the ball B^2, although the proof is deeper:

Theorem 10.2 (Brouwer fixed-point theorem for the disc). If $f : B^2 \longrightarrow B^2$ *is continuous, then there exists a point* $x \in B^2$ *such that* $f(x) = x$.

Proof. We proceed by contradiction. Suppose that $f(x) \neq x$ for every x in B^2. Then defining $v(x) = f(x) - x$ gives us a nonvanishing vector field $(x, v(x))$ on B^2. But the vector field v cannot point directly outward at any point x of S^1, for that would mean

$$f(x) - x = ax$$

for some *positive* real number a, whence $f(x) = (1 + a)x$ lies outside the unit ball B^2. We thus arrive at a contradiction. □

One might well wonder why fixed-point theorems are of interest in mathematics. It turns out that many problems, such as problems concerning existence of solutions for systems of equations, for instance, can be formulated as fixed-point problems. Here is one example, a classical theorem of Frobenius. We assume some knowledge of linear algebra at this point.

Corollary 10.3. Let A *be a* 3 *by* 3 *matrix of positive real numbers. Then* A *has a positive real eigenvalue (characteristic value).*

Proof. Let $T: R^3 \longrightarrow R^3$ be the linear transformation whose matrix (relative to the standard basis for R^3) is A. Let B be the intersection of the

2-sphere S^2 with the first octant

$$\{(x_1, x_2, x_3) \mid x_1 \geq 0 \text{ and } x_2 \geq 0 \text{ and } x_3 \geq 0\}$$

of R^3. It is easy to show that B is homeomorphic to the ball B^2, so that the fixed-point theorem holds for continuous maps of B into itself.

Now if $x = (x_1, x_2, x_3)$ is in B, then all the components of x are nonnegative and at least one is positive. Because all entries of A are positive, the vector $T(x)$ is a vector all of whose components are positive. As a result, the map $x \longrightarrow T(x)/\|T(x)\|$ is a continuous map of B to itself, which therefore has a fixed point x_0. Then

$$T(x_0) = \|T(x_0)\| x_0,$$

so that T (and therefore the matrix A) has the positive real eigenvalue $\|T(x_0)\|$. □

Before leaving the subject of vector fields, we should mention one of the most interesting problems to which they lead:

Given a surface S, does it possess a nonvanishing tangent vector field?

Seemingly a problem of differential geometry, this is, in fact, a problem of topology, for the answer depends only on the topological type of S. It turns out that among compact surfaces, only the torus T and the Klein bottle K (if you know what that is) have nonzero tangent vector fields. They are illustrated in Figure 27. [The torus, in fact, has *two* linearly independent vector fields.]

The proof of this fact is outside the scope of this book. (See [G-P] for a proof.) But we can proceed one step toward a solution by showing that the sphere S^2, at least, possesses no nonvanishing vector field. We shall sketch the proof geometrically, rather than carry it out formally using equations.

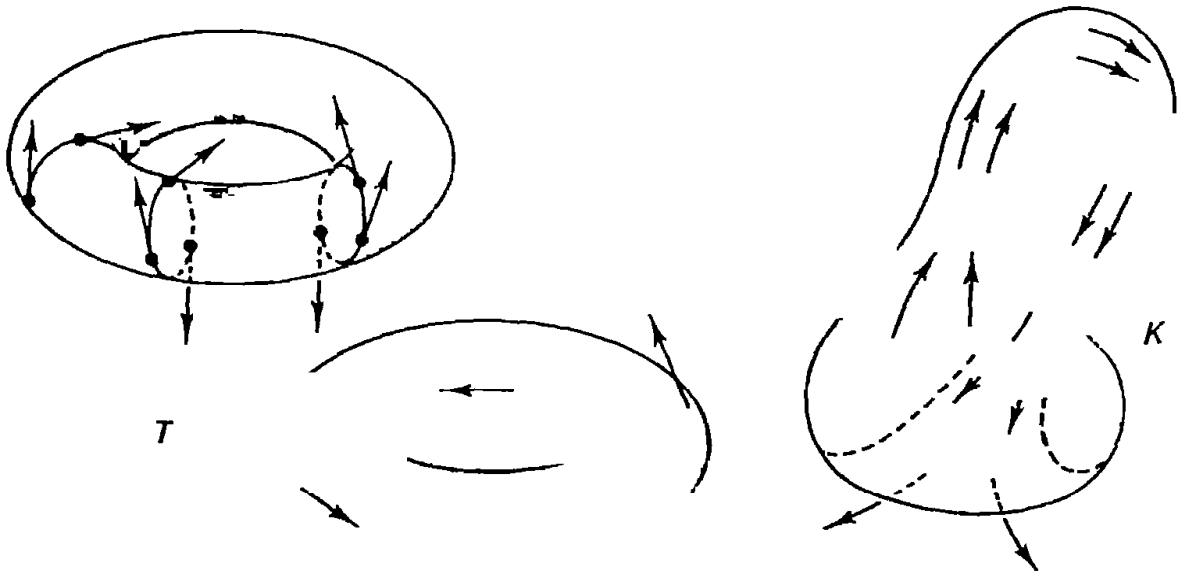


Figure 27

Theorem 10.4. *The sphere S^2 possesses no nonvanishing tangent vector field.*

Proof. Suppose that S^2 had a nonvanishing vector field $(x, v(x))$. Consider the north pole $p = (0, 0, 1)$ of S^2; we shall assume for convenience that at p the vector field is parallel to the y-axis. Take a small open ball U in S^2 centered at p, so small that on the ball U the vector field does not vary more than a few degrees from being parallel to the y-axis.

Now consider the map $f: S^2 - p \longrightarrow R^2$ given by "stereographic projection." (See Step 1 of the proof of Theorem 6.2.)

The map f is, in fact, a homeomorphism of $S^2 - p$ with R^2. More than that, it carries *tangent vectors* to S^2 continuously into *tangent vectors* to R^2. How? The easiest way to see what happens is to take a given tangent vector v at a point x, and to find a curve C in S^2 having that vector as its velocity vector at x. The map f carries the curve C into a curve in R^2, and we let it take the vector v into the velocity vector of the image curve $f(C)$ at the point $f(x)$. Let us denote the image of (x, v) by (y, w). You can check by direct computation that f is *smooth* (w is a well-defined, continuous function of (x, v)) and *nonsingular* (w is nonzero if v is nonzero).

Now we ask the question: What happens to the nonvanishing vector field (x, v) under this map? It is carried into a nonvanishing vector field (y, w) on R^2. In particular, consider the subspace $S^2 - U$ of S^2, which the

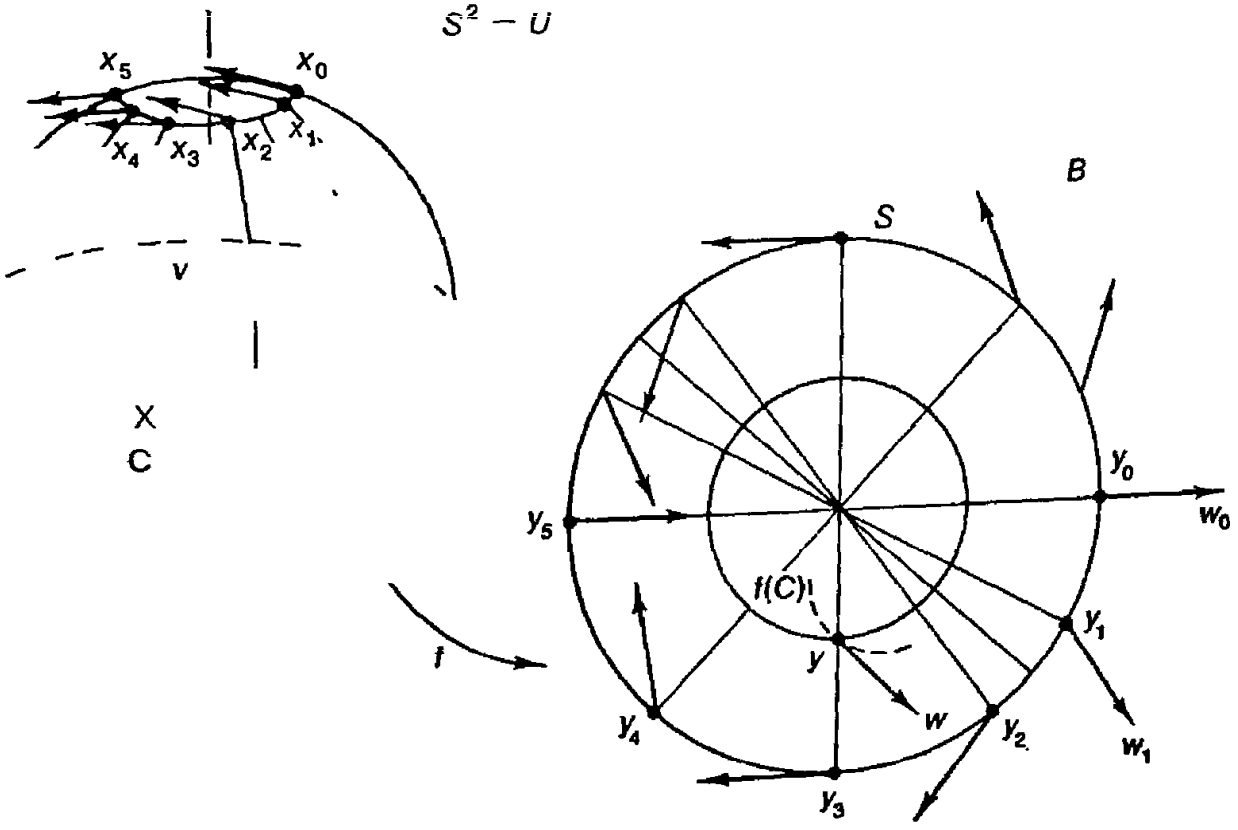


Figure 28

map f carries onto a ball B in R^2 of large radius about the origin. What does the image of the vector field look like? We have sketched in Figure 28 what it looks like on the large circle S that is the boundary of B.

So far so good. But let us examine this vector field $(y, w(y))$ on B more closely, particularly its restriction to the circle S. You can see from the picture that the map $h: S \longrightarrow R^2 - 0$ defined by

$$h(y) = w(y)$$

carries a generator of $\pi_1(S, x_0)$ to twice a generator of $\pi_1(R^2 - 0, x_0)$. Intuitively, as y goes around the circle S once, the point $h(y)$ goes around the origin twice.

Now comes the contradiction. We know that h is extendable to a map of B into $R^2 - 0$, because $(y, w(y))$ is a nonvanishing vector field on B. Thus h_* must be the zero homomorphism of fundamental groups. On the other hand, we know both fundamental groups in question are infinite cyclic groups, and that h_* carries a generator of the first to twice a generator of the second. In particular, h_* is *not* the zero homomorphism. □

These theorems have generalizations to higher dimensions, which are discussed in Exercises 7 and 8.

Exercises

1. Show that if A is a retract of B^2, then every continuous map $f: A \longrightarrow A$ has a fixed point.
2. Show that if A is a nonsingular 3 by 3 matrix having nonnegative entries, then A has a positive real eigenvalue.
3. Show that the set B of Corollary 10.3 is homeomorphic to B^2.
*4. Try to give an algebraic proof of Corollary 10.3. (This exercise is for those students who are tempted by the thought that this result is trivial! You might try the 2 by 2 case first.)
5. Show that if $f: S^1 \longrightarrow S^1$ is inessential, then f has a fixed point and f carries some point x to its antipode $-x$.
6. Show that given a continuous map $f: S^2 \longrightarrow S^2$, either f has a fixed point or f carries some point x to its antipode $-x$. [*Hint:* Use Theorem 10.4.]
7. Suppose that you are given the fact that for each n, there is no retraction $r: B^{n+1} \longrightarrow S^n$. (This result can be proved using techniques of algebraic topology). Derive the following corollaries:
   (a) The identity map $i: S^n \longrightarrow S^n$ is essential.
   (b) The antipodal map $a: S^n \longrightarrow S^n$ is essential.
   (c) The inclusion map $j: S^n \longrightarrow R^{n+1} - 0$ is essential.

(d) Every nonvanishing vector field on B^{n+1} points directly outward at some point of S^n, and directly inward at some point of S^n.
(e) Every continuous map $f: B^n \longrightarrow B^n$ has a fixed point.
(f) Every n by n matrix with positive real entries has a positive real eigenvalue.
(g) If $f: S^n \longrightarrow S^n$ is inessential, then f has a fixed point and f carries some point to its antipode.

8. It is a theorem that S^n has a nonvanishing tangent vector field if and only if n is odd. The "if" part is easy. Show that if $n = 2m - 1$, then

$$v(x) = (-x_2, x_1, -x_4, x_3, \ldots, -x_{2m}, x_{2m-1})$$

is a nonvanishing tangent vector field on S^n. The "only if" part is more difficult; see Exercise 7 of §8-11.

8-11 Homotopy Type

In §8-8 we considered the problem of determining whether or not a given map $h: X \longrightarrow Y$ was homotopic to a constant map. Now we consider the more general problem of determining whether or not two given maps $h, k: X \longrightarrow Y$ are homotopic. We prove that a necessary condition for h and k to be homotopic is that, up to an isomorphism of the groups involved, h_* and k_* are equal.

It turns out that this theorem has an application to the problem of computing fundamental groups. We have already noted, in §8-5, that two spaces have isomorphic fundamental groups if one is a strong deformation retract of the other, and we have used this fact in computing some fundamental groups. But there is a relation more general than this, called *homotopy equivalence*, which also implies that the spaces in question have isomorphic fundamental groups. We study that relation in this section.

First, let us consider what happens when two maps are homotopic.

Theorem 11.1. *Let* $\mathrm{h}, \mathrm{k}: \mathrm{X} \longrightarrow \mathrm{Y}$. *Let* $\mathrm{h}(\mathrm{x}_0) = \mathrm{y}_0$ *and* $\mathrm{k}(\mathrm{x}_0) = \mathrm{y}_1$. *If* h *and* k *are homotopic, then there is a path* α *in* Y *from* y_0 *to* y_1 *such that* $\mathrm{k}_* = \hat{\alpha} \circ \mathrm{h}_*$. *If* $\mathrm{y}_0 = \mathrm{y}_1$, *and if the base point* x_0 *remains fixed during the homotopy, then* $\mathrm{k}_* = \mathrm{h}_*$.

$$\begin{array}{ccc} \pi_1(\mathrm{X}, \mathrm{x}_0) & \xrightarrow{\mathrm{h}_*} & \pi_1(\mathrm{Y}, \mathrm{y}_0) \\ & {\scriptstyle \mathrm{k}_*} \searrow & \downarrow {\scriptstyle \hat{\alpha}} \\ & & \pi_1(\mathrm{Y}, \mathrm{y}_1) \end{array}$$

Proof. Let $H: X \times I \longrightarrow Y$ be the homotopy between h and k; then $H(x, 0) = h(x)$ and $H(x, 1) = k(x)$ for $x \in X$. Let $\alpha: I \longrightarrow Y$ be defined by the equation $\alpha(t) = H(x_0, t)$; then α is a path in Y from y_0 to y_1. We assert that $k_* = \hat{\alpha} \circ h_*$.

That is, we assert that for every loop $f: I \longrightarrow X$ in X based at x_0, we have

$$k_*([f]) = \hat{\alpha}(h_*([f])).$$

We must prove that $[k \circ f] = [\bar{\alpha}] * [h \circ f] * [\alpha]$, or, equivalently, that

$$(*) \qquad [\alpha] * [k \circ f] * [\bar{\alpha}] = [h \circ f].$$

Checking this equation requires the construction of a path homotopy G.

We describe G geometrically as follows: For given parameter value t, let α_t be the path that goes partway along α, from y_0 to $\alpha(t)$, reparametrized so as to be based on the interval [0, 1]. And let β_t be the loop $\beta_t(s) = H(f(s), t)$

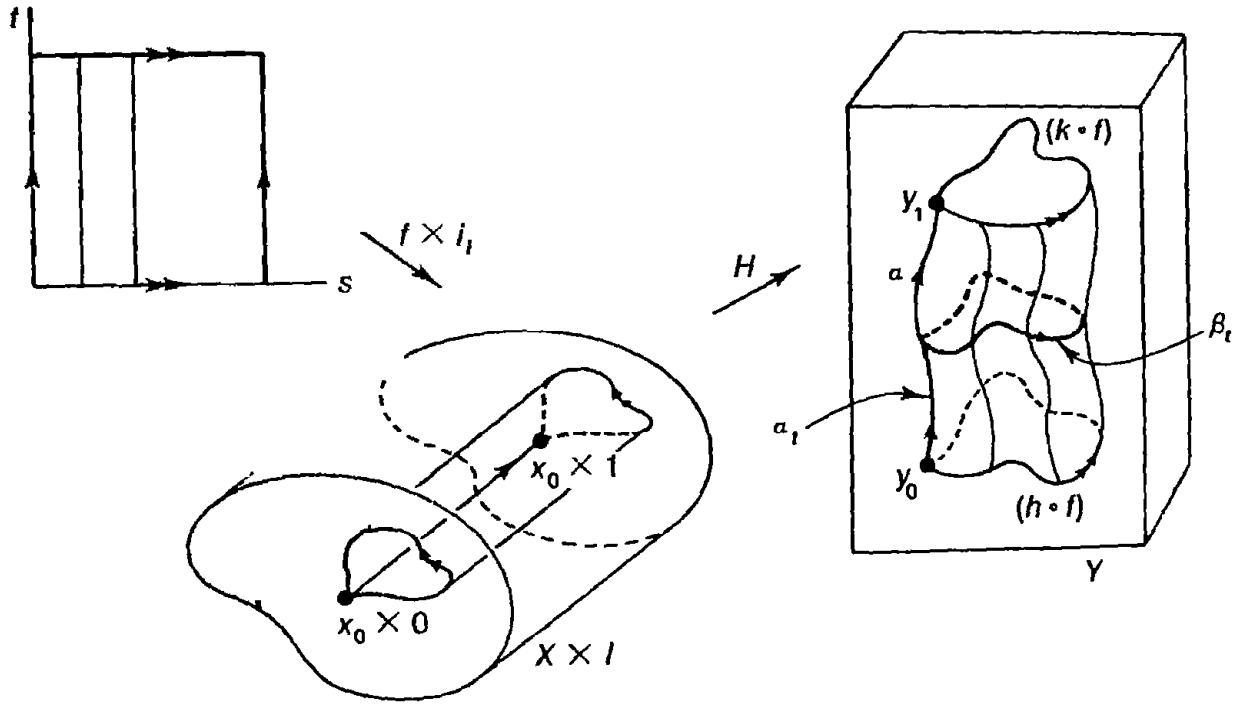


Figure 29

in Y based at $\alpha(t)$. See Figure 29. Then consider the loop

$$(\alpha_t * \beta_t) * \bar{\alpha}_t$$

based at y_0. When $t = 0$, this loop equals

$$(e_{y_0} * (h \circ f)) * e_{y_0},$$

which is path homotopic to $h \circ f$; and when $t = 1$, it equals the loop

$$(\alpha * (k \circ f)) * \bar{\alpha}.$$

Thus setting $G(s, t) = ((\alpha_t * \beta_t) * \bar{\alpha}_t)(s)$ gives us the path homotopy required to prove equation (*).

Formally, we define $G: I \times I \longrightarrow Y$ as follows:

$$G(s,t) = \begin{cases} \alpha(4st) & \text{for } s \in [0, \tfrac{1}{4}], \\ H(f(4s-1), t) & \text{for } s \in [\tfrac{1}{4}, \tfrac{1}{2}], \\ \alpha(2t(1-s)) & \text{for } s \in [\tfrac{1}{2}, 1]. \end{cases}$$

You can check that these formulas make sense, that they agree when the domains overlap, and that G is the required path homotopy.

Finally, we note that if the homotopy H leaves the base point x_0 fixed at the point y_0, then the path $\alpha(t) = H(x_0, t)$ is the constant path, so that $k_* = \hat{\alpha} \circ h_* = h_*$. □

Corollary 11.2. Let $\mathrm{h}, \mathrm{k} : \mathrm{X} \longrightarrow \mathrm{Y}$; *let* $\mathrm{h}(\mathrm{x}_0) = \mathrm{y}_0$ *and* $\mathrm{k}(\mathrm{x}_0) = \mathrm{y}_1$. *Suppose that* h *and* k *are homotopic. If* k_* *is injective (or surjective, or the zero homomorphism), then so is* h_*. *In particular, if* h *is homotopic to a constant map, then* h_* *is the zero homomorphism.*

Now we turn to the second problem we mentioned at the beginning of this section, the problem of computing fundamental groups. We obtain a general condition under which two spaces have isomorphic fundamental groups.

Definition. A continuous map $f : X \longrightarrow Y$ is called a **homotopy equivalence** if there is a continuous map $g : Y \longrightarrow X$ such that $g \circ f$ is homotopic to the identity map i_X of X and $f \circ g$ is homotopic to the identity map i_Y of Y. The map g is said to be a **homotopy inverse** for the map f.

It is easy to show that given any collection $\mathcal{C}$ of topological spaces, the relation of homotopy equivalence is an equivalence relation on $\mathcal{C}$. See Exercise 2. Two spaces that are homotopy equivalent are said to have the same **homotopy type**.

The notion of homotopy equivalence generalizes the notion of strong deformation retraction, defined in §8-5; if A is a strong deformation retract of X, then A has the same homotopy type as X. For let $j : A \longrightarrow X$ be the inclusion mapping and let $r : X \longrightarrow A$ be the retraction mapping. Then the composite $r \circ j$ *equals* the identity map of A, and the composite $j \circ r$ is by hypothesis homotopic to the identity map of X (and in fact each point of A remains fixed during the homotopy).

Under the hypothesis that two spaces are homotopy equivalent, we can prove that they have isomorphic fundamental groups:

Theorem 11.3. *Let* $\mathrm{f} : \mathrm{X} \longrightarrow \mathrm{Y}$ *be continuous; let* $\mathrm{f}(\mathrm{x}_0) = \mathrm{y}_0$. *If* f *is a homotopy equivalence, then*

$$\mathrm{f}_* : \pi_1(\mathrm{X}, \mathrm{x}_0) \longrightarrow \pi_1(\mathrm{Y}, \mathrm{y}_0)$$

is an isomorphism.

Proof. Let $g : Y \longrightarrow X$ be a homotopy inverse for f. Consider the maps

$$(X, x_0) \xrightarrow{f} (Y, y_0) \xrightarrow{g} (X, x_1) \xrightarrow{f} (Y, y_1),$$

where $x_1 = g(y_0)$ and $y_1 = f(x_1)$. We have the corresponding induced homomorphisms:

$$\begin{array}{ccc} \pi_1(X, x_0) & \xrightarrow{(f_{x_0})_*} & \pi_1(Y, y_0) \\ & \searrow^{g_*} & \\ \pi_1(X, x_1) & \xrightarrow{(f_{x_1})_*} & \pi_1(Y, y_1) \end{array}$$

[Here we have to distinguish between the homomorphisms induced by f relative to two different base points.] Now

$$g \circ f : (X, x_0) \longrightarrow (X, x_1).$$

Because $g \circ f$ is by hypothesis homotopic to the identity map, there is a path α in X such that

$$(g \circ f)_* = \hat{\alpha} \circ (i_X)_* = \hat{\alpha}.$$

In particular, $(g \circ f)_* = g_* \circ (f_{x_0})_*$ is an isomorphism.

Similarly, because $f \circ g$ is homotopic to i_Y, the homomorphism $(f \circ g)_* = (f_{x_1})_* \circ g_*$ is an isomorphism.

The first fact implies that g_* is surjective, and the second implies that g_* is injective. Therefore, g_* is an isomorphism. Applying the first equation once again, we conclude that

$$(f_{x_0})_* = (g_*)^{-1} \circ \hat{\alpha},$$

so that $(f_{x_0})_*$ is also an isomorphism.

Note that although g is a homotopy inverse for f, the homomorphism g_* is not an inverse for the homomorphism $(f_{x_0})_*$. □

The relation of homotopy equivalence is clearly more general than the notion of strong deformation retraction. The theta space and the figure eight are both strong deformation retracts of the doubly punctured plane; we noted this fact in Examples 3 and 4 of §8-5. Therefore, they are homotopy equivalent to the doubly punctured plane, and hence to each other. But neither is homeomorphic to a strong deformation retract of the other; in fact, neither of them can even be imbedded in the other. (See Exercise 4.)

It is a striking fact that the situation which occurs for these two spaces is the standard situation with regard to homotopy equivalences. Martin Fuchs has proved a theorem to the effect that two spaces X and Y have the same homotopy type if and only if they are homeomorphic to strong deformation retracts of a single space Z. The proof, while it uses only elementary tools, is difficult [F].

Exercises

1. Check the details concerning the homotopy G constructed in the proof of Theorem 11.1.
2. Show that given a collection $\mathcal{C}$ of spaces, the relation of homotopy equivalence is an equivalence relation on $\mathcal{C}$.

3. Show that X has the homotopy type of a one-point space if and only if X is contractible. (See Exercise 4 of §8-8.)

4. Show that neither of the spaces θ and 8 can be imbedded in the other.

5. Let A be a subspace of X; let $j: A \longrightarrow X$ be the inclusion mapping; let $f: X \longrightarrow A$ be a continuous map. Suppose that the map $j \circ f: X \longrightarrow X$ is homotopic to the identity map $i_X: X \longrightarrow X$ under a homotopy $H: X \times I \longrightarrow X$.
   (a) Show that if $H(a, t) \in A$ for every $a \in A$, then j_* and f_* are isomorphisms.
   (b) Show that if f is a retraction, then j_* and f_* are isomorphisms. [In this case, H is called merely a **deformation retraction** rather than a strong deformation retraction.]
   (c) Give an example to show that j_* and f_* need not be isomorphisms in general.

6. We define the *degree* of a continuous map $h: S^1 \longrightarrow S^1$ as follows:

   Let b_0 be the point $(1, 0)$ of S^1; choose a generator γ for the infinite cyclic group $\pi_1(S^1, b_0)$. (There are only two, differing in sign.) If x_0 is any point of S^1, choose a path α in S^1 from b_0 to x_0, and define $\gamma(x_0) = \hat{\alpha}(\gamma)$. Then $\gamma(x_0)$ generates $\pi_1(S^1, x_0)$. The element $\gamma(x_0)$ is independent of the choice of the path α, since the fundamental group of S^1 is abelian.

   Now given $h: S^1 \longrightarrow S^1$, choose $x_0 \in S^1$ and let $h(x_0) = x_1$. Consider the homomorphism

$$h_*: \pi_1(S^1, x_0) \longrightarrow \pi_1(S^1, x_1).$$

   Since both groups are infinite cyclic, we have

$$(*) \qquad h_*(\gamma(x_0)) = d \cdot \gamma(x_1)$$

   for some integer d. The integer d is called the **degree** of h and is denoted by $\deg h$.

   The degree of h is independent of the choice of the generator γ; choosing the other generator would merely change the sign of both sides of $(*)$.
   (a) Show that d is independent of the choice of x_0.
   (b) Show that if $h, k: S^1 \longrightarrow S^1$ are homotopic, they have the same degree.
   (c) Show that $\deg(h \circ k) = (\deg h) \cdot (\deg k)$.
   (d) Compute the degrees of the identity map, the reflection $\rho(x_1, x_2) = (x_1, -x_2)$, the constant map, and the map $h(z) = z^n$, where z is a complex number.
   *(e) Show that if $h, k: S^1 \longrightarrow S^1$ have the same degree, they are homotopic.

7. Suppose that to every map $h: S^n \longrightarrow S^n$ we have assigned an integer, denoted by $\deg h$ and called the **degree** of h, such that:
   (i) Homotopic maps have the same degree.
   (ii) $\deg(h \circ k) = (\deg h) \cdot (\deg k)$.
   (iii) The identity map has degree 1, the constant map has degree 0, and the reflection map $\rho(x_1, \ldots, x_{n+1}) = (x_1, \ldots, -x_{n+1})$ has degree -1.

   [One can construct such a function, using the tools of algebraic topology. Intuitively, $\deg h$ measures how many times h wraps S^n about itself; the sign tells you whether h preserves orientation or not.] Prove the following:
   (a) There is no retraction $r: B^{n+1} \longrightarrow S^n$.
   (b) If $h: S^n \longrightarrow S^n$ has degree different from $(-1)^{n+1}$, then h has a fixed point. [*Hint:* Show that if h has no fixed point, then h is homotopic to the antipodal map.]

(c) If $h : S^n \to S^n$ has degree different from 1, then h maps some point x to its antipode $-x$.
(d) If S^n has a nonvanishing tangent vector field v, then n is odd. [*Hint:* If v exists, show the identity map is homotopic to the antipodal map.]
(e) Every map $h : S^{2m} \to S^{2m}$ has either a fixed point or a point x such that $h(x) = -x$.
(f) Compare these results with Exercises 6, 7, and 8 of §8-10.
For the use of the notion of *degree* in differential geometry, see [G-P].

8-12 *The Jordan Separation Theorem*

We now consider one of the classical theorems of mathematics, the Jordan curve theorem. It states a fact that is geometrically quite "obvious," the fact that a simple closed curve always separates the plane into two pieces. But it is quite difficult to prove by direct geometric arguments. Here we prove it as a corollary of our study of covering spaces and the fundamental group.

There are actually three separation theorems involved. The first, which we call the *Jordan separation theorem*, states that a simple closed curve in S^2 (or in the plane) separates it into *at least* two components. The second says that an arc does not separate S^2 (or the plane). And the third, the *Jordan curve theorem* proper, says that a simple closed curve C in S^2 (or in the plane) separates it into *precisely* two components, of which C is the common boundary.

We shall prove the first of these theorems now. We need the following fact:

Lemma 12.1. *Let* a *and* b *be points of* S^2; *let* A *be a compact space; let* $f : A \to S^2 - a - b$ *be a continuous map. If* a *and* b *lie in the same component of* $S^2 - f(A)$, *then* f *is inessential.*

Intuitively, this lemma says that if the set $f(A)$ does not separate a from b in S^2, then f can be shrunk to a point without touching a or b.

Proof. Step 1. We show there is a homeomorphism h of S^2 with the one-point compactification $R^2 \cup \{\infty\}$ of R^2 such that $h(a) = \infty$ and $h(b) = 0$.

To construct h, we first take a rotation h_1 of S^2 that carries a to the north pole p. We then take stereographic projection h_2, mapping $S^2 - p$ homeomorphically onto R^2; we extend it to S^2 by letting $h_2(p) = \infty$. Finally, we take a translation h_3 of R^2 carrying $h_2(h_1(b))$ to the origin 0, and we extend it to $R^2 \cup \{\infty\}$ by letting $h_3(\infty) = \infty$. It is easy to check that h_2 and h_3 are homeomorphisms, as is h_1. Then $h_3 \circ h_2 \circ h_1$ is the desired homeomorphism.

Step 2. In view of Step 1, the lemma reduces to the following statement: Let A be compact; let $f : A \to R^2 - 0$ be a continuous map. If 0 lies in the unbounded component of $R^2 - f(A)$, then f is inessential.

This statement is easy to prove. Assume that **0** lies in the unbounded component of $R^2 - f(A)$. Choose a ball B centered at the origin, of sufficiently large radius that it contains the set $f(A)$. Choose a point p of R^2 lying outside B. Then **0** and p lie in the same component of $R^2 - f(A)$.

Because R^2 is locally path connected, so is the open set $R^2 - f(A)$. Therefore, the components and path components of $R^2 - f(A)$ are the same. Hence we can choose a path α in $R^2 - f(A)$ from **0** to p. We define a homotopy $F : A \times I \to R^2 - \mathbf{0}$ by the equation

$$F(x, t) = f(x) - \alpha(t);$$

it is pictured in Figure 30. The homotopy F is a homotopy between the map f and the map g defined by $g(x) = f(x) - p$. Note that $F(x, t) \neq \mathbf{0}$ because the path α does not intersect the set $f(A)$.

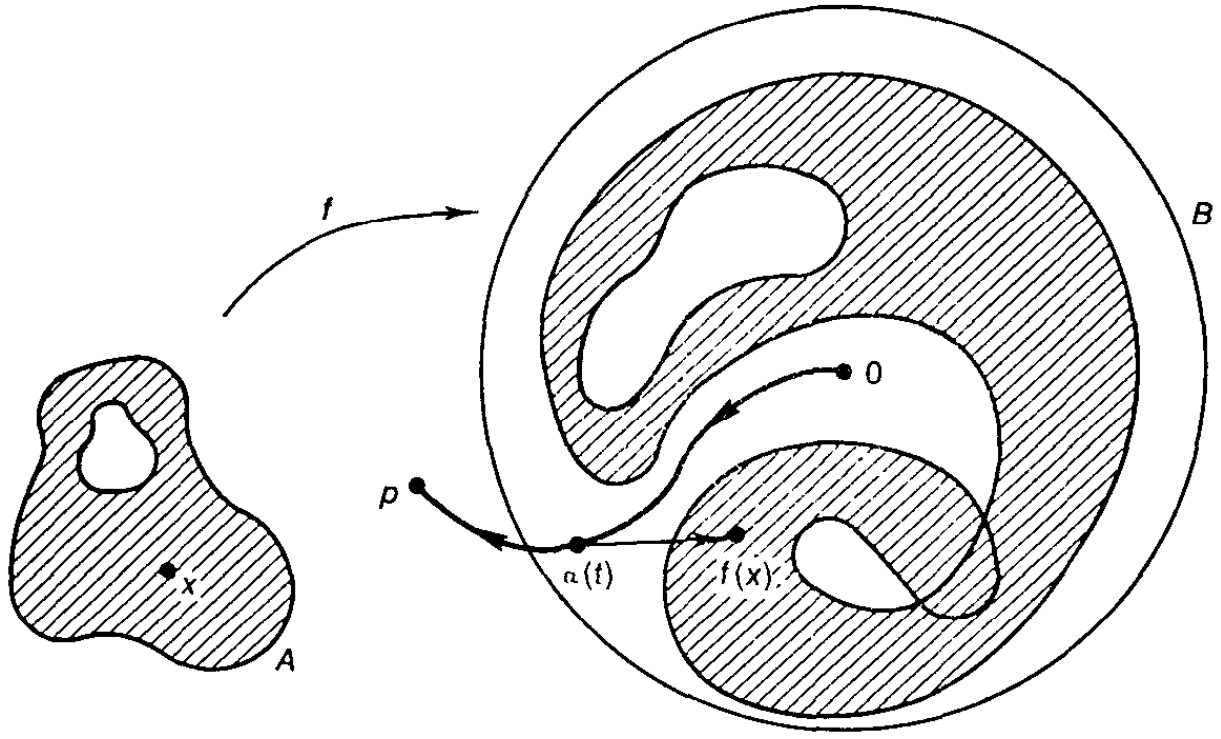


Figure 30

Now we define a homotopy $G : A \times I \to R^2 - \mathbf{0}$ by the equation

$$G(x, t) = tf(x) - p.$$

It is a homotopy between the map g and a constant map. Note that $G(x, t) \neq \mathbf{0}$ because $tf(x)$ lies inside the ball B and p does not.

Thus we have proved that f is inessential. □

Definition. An **arc** is a space homeomorphic to the unit interval [0, 1]. A **simple closed curve** is a space homeomorphic to the circle S^1.

Theorem 12.2 (The Jordan separation theorem). *Let* C *be a simple closed curve in* S^2. *Then* $S^2 - C$ *is not connected.*

Proof. Because $S^2 - C$ is locally path connected, its components and path components are the same. We assume that $S^2 - C$ is path connected and derive a contradiction.

We are going to apply Theorem 6.1, the special case we proved of the Van Kampen theorem. It expresses the fundamental group of $X = U \cup V$ in terms of the fundamental groups of U and V, provided the intersection $U \cap V$ is path connected.

Let us first write C as the union of two arcs A_1 and A_2 that intersect only at their end points a and b. Then let X denote the space $S^2 - a - b$. The space X is homeomorphic to the punctured plane $R^2 - 0$, so its fundamental group is infinite cyclic.

Let U be the open set $S^2 - A_1$ and let V be the open set $S^2 - A_2$; then $X = U \cup V$. Now

$$U \cap V = S^2 - (A_1 \cup A_2) = S^2 - C,$$

so that $U \cap V$ is path connected by assumption. Let x_0 be a point of $U \cap V$. If we can prove that the inclusions

$$i : (U, x_0) \longrightarrow (X, x_0) \quad \text{and} \quad j : (V, x_0) \longrightarrow (X, x_0)$$

induce zero homomorphisms of the fundamental group, Theorem 6.1 will apply to show that $\pi_1(X, x_0) = 0$. This contradicts the fact that $\pi_1(X, x_0)$ is infinite cyclic, so our assumption that $S^2 - C$ is path connected must have been mistaken.

Let us prove that i_* is the zero homomorphism. Given a loop $f : I \to U$ based at x_0, we show that $i_*([f])$ is trivial. For this purpose, let $\phi : I \to S^1$ be the standard loop generating $\pi_1(S^1, b_0)$. The map $f : I \to U$ induces a continuous map $h : S^1 \to U$ such that $h \circ \phi = f$. See Figure 31.

Consider the map $i \circ h : S^1 \to S^2 - a - b$. By hypothesis, the set $i(h(S^1)) = h(S^1)$ does not intersect the arc A_1 joining a and b. Therefore, a

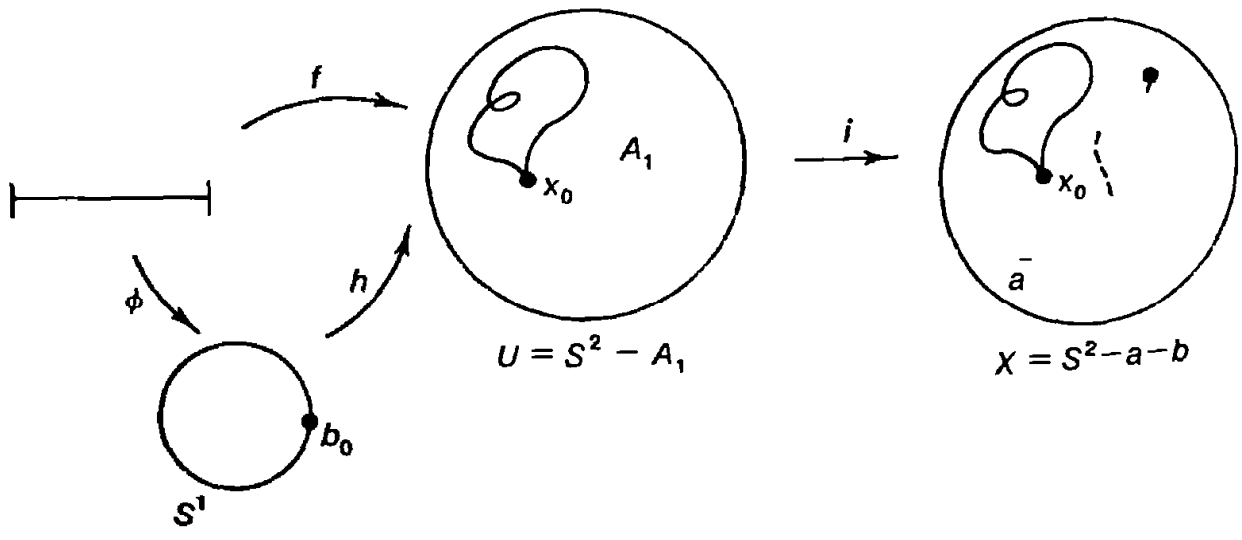


Figure 31

and b lie in the same component of $S^2 - \imath(h(S^1))$. By the preceding lemma, the map $i \circ h$ is inessential. It follows from Theorem 8.2 that $(i \circ h)_*$ is the zero homomorphism of fundamental groups. But

$$(i \circ h)_*([\phi]) = [i \circ h \circ \phi] = [i \circ f] = i_*([f]).$$

Therefore, $i_*([f])$ is trivial, as desired. □

Exercises

1. By identifying S^2 with $R^2 \cup \{\infty\}$, prove that every simple closed curve in R^2 separates R^2.
2. We assumed in the proof of Theorem 12.2 that the simple closed curve C cannot equal all of S^2. Why is this justified?
3. Give examples to show that a simple closed curve in the torus may or may not separate the torus.
4. (a) *Lemma. Let A and B be closed connected subsets of S^2 whose intersection consists of precisely two points. Then $A \cup B$ separates S^2.*
   (b) Let C be a subset of S^2 homeomorphic to the space which is the union of the topologist's sine curve and the broken-line path from (0, 1) to (0, −2) to (1, −2) to (1, sin 1). (See Figure 32.) Show that C separates S^2.

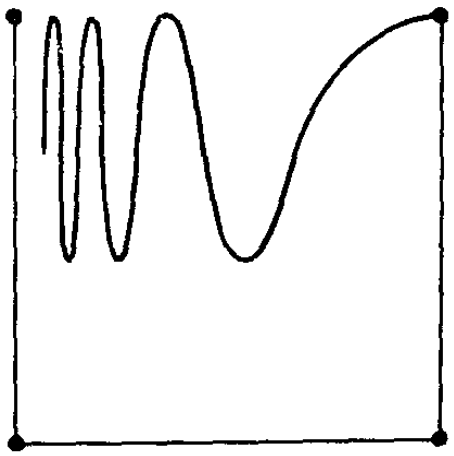


Figure 32

*5. This exercise leads to a proof of the invariance of domain theorem for R^2.
   (a) Prove the following:
   Lemma (Homotopy extension lemma). Let X and $X \times I$ be normal; let A be a closed subset of X. If $f: A \longrightarrow R^2 - \mathbf{0}$ is a continuous map and f is inessential, then f can be extended to a continuous map $g: X \longrightarrow R^2 - \mathbf{0}$.
   [*Hint:* Define a map of the subspace $Y = (A \times I) \cup (X \times 1)$ of $X \times I$ into $R^2 - \mathbf{0}$ by using the homotopy between f and a constant map. Extend to a neighborhood U of Y in $X \times I$. Then construct a map of $X \times 0$ into U that equals the identity on $A \times 0$.]
   (b) Prove:
   Theorem (Borsuk). Let X be a compact subset of R^2. If $\mathbf{0}$ lies in a bounded

component C of $R^2 - X$, then the inclusion map $j: X \to R^2 - 0$ is essential; and conversely.

[*Hint:* Let B be a large ball that contains C and X. If j is inessential, show that the inclusion map of $B - C$ into $R^2 - 0$ can be extended to a continuous map of B into $R^2 - 0$.]

(c) Prove:

Theorem (A nonseparation theorem). No arc separates R^2; no space homeomorphic to B^2 separates R^2.

(d) Prove the following. Another proof is outlined in §8-13, Exercise 9.

Theorem (Brouwer theorem on invariance of domain for R^2). If U is an open subset of R^2 and $f: U \to R^2$ is continuous and injective, then $f(U)$ is open in R^2 and f is an imbedding.

[*Hint:* If $f: B^2 \to R^2$ is an imbedding, note that $f(S^1)$ separates R^2 and $f(B^2)$ does not; conclude that $f(\text{Int } B^2)$ must equal one of the components of $R^2 - f(S^1)$.]

8-13 The Jordan Curve Theorem

The special case of the Van Kampen theorem, which we just applied in proving the Jordan separation theorem, tells us something about the fundamental group of the space $X = U \cup V$ in the case where the intersection $U \cap V$ is path connected. In the next two lemmas we examine what happens when $U \cap V$ is *not* connected. In the first lemma we suppose that $U \cap V$ has at least two components, and in the second we assume that it has at least three.

These lemmas will enable us to complete the proof of the Jordan curve theorem. The proof we give is, as far as we know, a new one. It involves a construction similar to that used in complex analysis to construct Riemann surfaces. But it makes no use of normality or the Tietze theorem.

Lemma 13.1. Let X *be the union of two open sets* U *and* V. *Suppose that* $U \cap V$ *can be written as the union of two disjoint open sets* A *and* B. *Let* $a \in A$ *and* $b \in B$; *suppose that* a *and* b *can be joined by paths in* U *and in* V. *Then* $\pi_1(X, a) \neq 0$.

Proof. The proof is in many ways an imitation of the proof in §8-4 that the fundamental group of the circle is infinite cyclic. As in that proof, the crucial step is to find an appropriate covering space E for the space X.

Step 1. (Construction of E). We construct E by pasting together copies of the subspaces U and V. Let us take countably many copies of U and countably many copies of V, all disjoint, say

$$U \times (2n) \quad \text{and} \quad V \times (2n + 1)$$

for all $n \in Z$, where Z denotes the integers. Let Y denote the union of these spaces; Y is a subspace of $X \times Z$. Now we form a new space E as a quotient

space of Y by identifying the points

$$x \times (2n) \quad \text{and} \quad x \times (2n-1) \quad \text{for } x \in A$$

and by identifying the points

$$x \times (2n) \quad \text{and} \quad x \times (2n+1) \quad \text{for } x \in B.$$

Let $\pi : Y \longrightarrow E$ be the quotient map.

Now the map $\rho : Y \longrightarrow X$ defined by $\rho(x \times m) = x$ induces a map $p : E \longrightarrow X$; the map p is continuous because E has the quotient topology. The map p is also surjective. We shall show that p is a covering map.

The space E is illustrated in Figure 33.

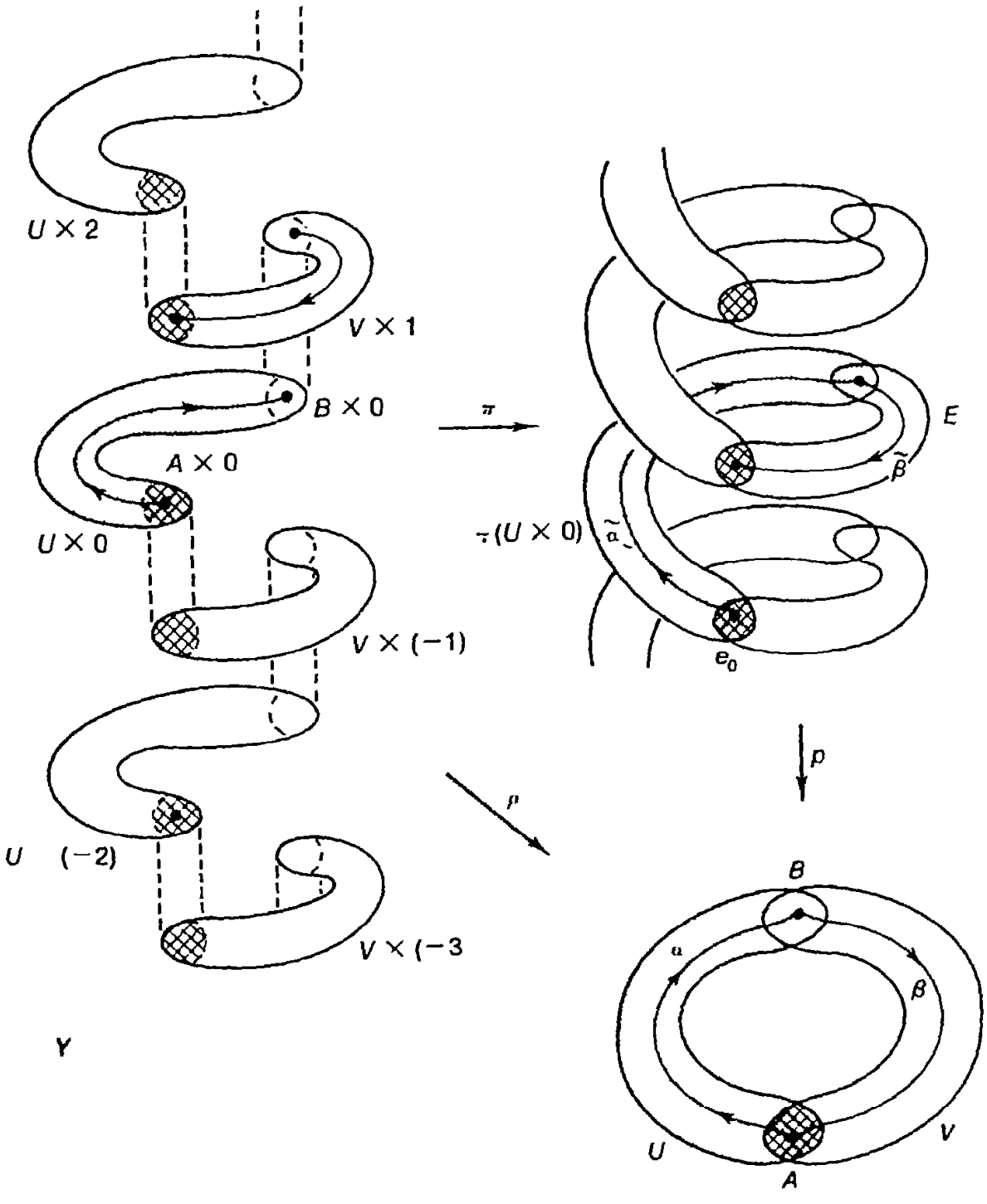


Figure 33

First let us show that the map π is an open map. Since Y is the union of the disjoint open sets $\{U \times (2n)\}$ and $\{V \times (2n+1)\}$, it will suffice to show that $\pi|(U \times 2n)$ and $\pi|(V \times (2n+1))$ are open maps. And this is easy. Take an open set in $U \times 2n$, for example; it will be of the form $W \times 2n$, where W is open in U. Then

$$\pi^{-1}(\pi(W \times 2n)) = [W \times 2n] \cup [(W \cap B) \times (2n+1)] \cup [(W \cap A) \times (2n-1)],$$

which is the union of three open sets of Y and hence open in Y. By definition of the quotient topology, $\pi(W \times 2n)$ is open in E, as desired.

Now we prove that p is a covering map; we show that the open sets U and V are evenly covered by p. Consider U, for example. The set $p^{-1}(U)$ is the union of the disjoint sets $\pi(U \times 2n)$ for $n \in Z$. Each of these sets is open in E because π is an open map. Let π_{2n} denote the restriction of π to the open set $U \times 2n$, mapping it onto $\pi(U \times 2n)$. It is a homeomorphism, because it is bijective, continuous, and open. Then when restricted to $\pi(U \times 2n)$, the map p is just the composite of the two homeomorphisms

$$\pi(U \times 2n) \xrightarrow{\pi_{2n}^{-1}} U \times 2n \xrightarrow{\rho} U$$

and is thus a homeomorphism. Therefore, $p|\pi(U \times 2n)$ maps this set homeomorphically onto U, as desired.

Step 2. Now we prove that $\pi_1(X, a) \neq 0$. Take a path α in U from a to b, and a path β in V from b to a. We assert that the loop $\alpha * \beta$ based at a is not path homotopic to a constant. Let us lift $\alpha * \beta$ to a path on E beginning at the base point $e_0 = \pi(a \times 0)$. Define

$$\tilde{\alpha}(t) = \pi(\alpha(t) \times 0),$$
$$\tilde{\beta}(t) = \pi(\beta(t) \times 1).$$

Since $\tilde{\alpha}(1) = \pi(b \times 0) = \pi(b \times 1) = \tilde{\beta}(0)$, the path $\tilde{\alpha} * \tilde{\beta}$ is defined. It is easily seen to be a lifting of $\alpha * \beta$, because $p \circ \tilde{\alpha} = \alpha$ and $p \circ \tilde{\beta} = \beta$. Since the lifted path $\tilde{\alpha} * \tilde{\beta}$ begins at the base point e_0 and ends at a point different from the base point, the loop $\alpha * \beta$ cannot be path homotopic to the constant loop. □

Lemma 13.2. Let X *be the union of two open sets* U *and* V. *Suppose that* $U \cap V$ *can be written as the union of three disjoint open sets* A, A′, *and* B. *Let* a ∈ A, a′ ∈ A′, *and* b ∈ B; *suppose that all three points can be joined by paths in* U *and in* V. *Then* $\pi_1(X, a)$ *is not infinite cyclic.*

Proof. Let α and β be paths in U and V, respectively, from a to b and from b to a, respectively. Let δ and γ be paths in U and V, respectively, from a to a' and from a' to a, respectively. Writing $U \cap V$ as the union of the disjoint open sets

$$A \quad \text{and} \quad A' \cup B,$$

we see from the proof of the preceding lemma that both the loop $\alpha * \beta$ and the loop $\delta * \gamma$ represent nontrivial elements of $\pi_1(X, a)$.

Now if $\pi_1(X, a)$ were infinite cyclic, each of these loops would represent a multiple of the generator of the group. Thus there would exist nonzero integers n and m such that

$$n[\alpha * \beta] = m[\delta * \gamma].$$

We shall prove that such integers do not exist.

Apply the construction of the preceding lemma to this situation, writing $U \cap V$ now as the union of the two disjoint open sets

$$A \cup A' \quad \text{and} \quad B.$$

One obtains the covering space E of X as before. See Figure 34. (Remember that we never assumed A and B were *connected* when we constructed E.)

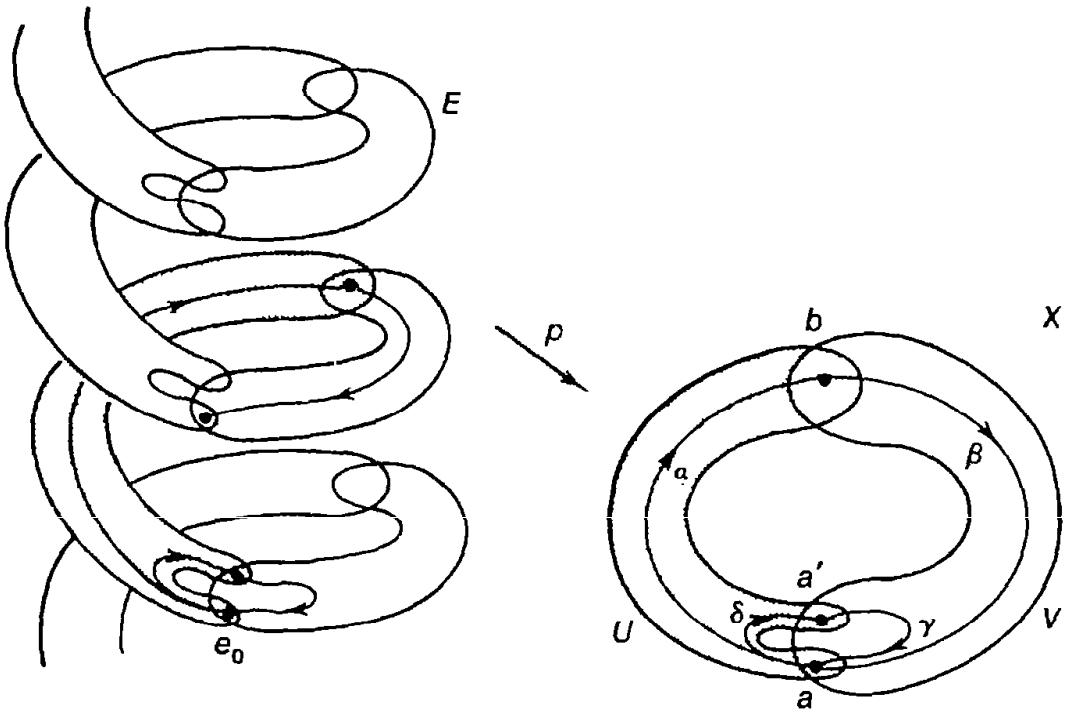


Figure 34

Now consider the loops $\alpha * \beta$ and $\delta * \gamma$ and lift them to paths on E beginning at e_0. It is easy to see that $\delta * \gamma$ lifts to a *loop* in E, while $\alpha * \beta$ does not. More generally, every *multiple* of $\delta * \gamma$ will lift to a loop in E, while every nonzero multiple of $\alpha * \beta$ lifts to a path that begins at e_0 and ends at some other point of E.

Therefore, no multiple of $\delta * \gamma$ can be path homotopic to a multiple of $\alpha * \beta$. □

Now we prove that no arc in S^2 separates S^2. One proof was outlined in Exercise 5(c) of the preceding section. Here is another, which you can skip if you have worked the exercise cited.

Theorem 13.3 (A nonseparation theorem). Let A *be an arc in* S^2. *Then* $S^2 - A$ *is connected.*

Proof. Step 1. Let D be an arc in S^2 which is written as the union of two arcs D_1 and D_2 having precisely one point d in common. Let a and b be points not in D. We assert that if a and b can be joined by paths in $S^2 - D_1$ and $S^2 - D_2$, then they can be joined by a path in $S^2 - D$. Figure 35 illustrates the fact that this assertion is not entirely trivial.

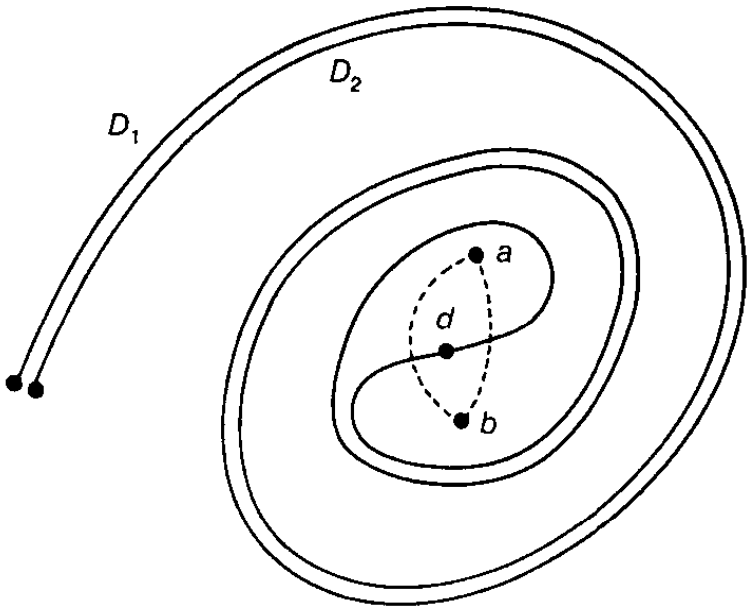


Figure 35

We suppose that a and b cannot be joined by a path in $S^2 - D$ and derive a contradiction. We apply Lemma 13.1. Let X be the space $S^2 - d$. Let U and V be the open sets

$$U = S^2 - D_1 \quad \text{and} \quad V = S^2 - D_2.$$

Then $X = U \cup V$, and $U \cap V = S^2 - D$. By hypothesis, a and b are points of $S^2 - D$ that cannot be joined by a path in $S^2 - D$. Therefore, $U \cap V$ is not path connected. Let A be the path component of $U \cap V$ containing a; let B be the union of the other path components of $U \cap V$. Since $U \cap V$ is locally path connected (being open in S^2), the path components of $U \cap V$ are open; hence A and B are open in X. We are given that a and b can be joined by paths in $U = S^2 - D_1$ and $V = S^2 - D_2$. We conclude from Lemma 13.1 that $\pi_1(X, a) \neq 0$.

But X is the space $S^2 - d$, which is homeomorphic to R^2 and is thus simply connected. Hence it must be possible to join a and b by a path in $S^2 - D$.

Step 2. Now, given the arc A and the points a and b in $S^2 - A$, we suppose that a and b cannot be joined by a path in $S^2 - A$ and derive a contradiction. This proves our theorem.

Choose a homeomorphism $h : [0, 1] \to A$; let $A_1 = h([0, \frac{1}{2}])$ and $A_2 =$

$h([\frac{1}{2}, 1])$. The result of Step 1 shows that since a and b cannot be joined by a path in $S^2 - A$, they cannot be joined by paths in both $S^2 - A_1$ and $S^2 - A_2$. To be definite, suppose that a and b cannot be joined by a path in $S^2 - A_2$.

Now repeat the argument, breaking A_2 up into two arcs $B_1 = h([\frac{1}{2}, \frac{3}{4}])$ and $B_2 = h([\frac{3}{4}, 1])$. We conclude from Step 1 that a and b cannot be joined by paths in both $S^2 - B_1$ and $S^2 - B_2$.

Continue similarly. In this way we define a sequence

$$I \supset I_1 \supset I_2 \supset \cdots$$

of closed intervals such that I_n has length $(\frac{1}{2})^n$ and such that for each n, the points a and b cannot be joined by a path in $S^2 - h(I_n)$. Compactness of the unit interval guarantees there is a point x in $\bigcap I_n$; since the lengths of the intervals converge to zero, there is only one such point.

Then consider the space $S^2 - h(x)$. Since this space is homeomorphic to R^2, the points a and b can be joined by a path α in $S^2 - h(x)$. Since $h(x) = \bigcap h(I_n)$, the space $S^2 - h(x)$ equals the union of the open sets

$$(S^2 - h(I_1)) \subset (S^2 - h(I_2)) \subset \cdots.$$

Because $\alpha(I)$ is compact, it must lie in the union of finitely many of these sets, and hence in some one of them, say $S^2 - h(I_m)$. But then α is a path in the set $S^2 - h(I_m)$ joining a and b, contrary to hypothesis. □

Theorem 13.4 (The Jordan curve theorem). Let C *be a simple closed curve in* S^2. *Then* $S^2 - C$ *has precisely two components* W_1 *and* W_2, *of which* C *is the common boundary.* [*That is,* $C = \bar{W}_1 - W_1 = \bar{W}_2 - W_2$.]

Proof. Step 1. We first prove that $S^2 - C$ has exactly two components. We apply Lemma 13.2. Write C as the union of two arcs C_1 and C_2 having precisely two points p and q in common. Let X be the space $S^2 - p - q$. Let U and V be the open sets

$$U = S^2 - C_1 \quad \text{and} \quad V = S^2 - C_2.$$

Then $X = U \cup V$. And $U \cap V$ equals $S^2 - C$, which has at least two components, by Theorem 12.2.

Suppose that $S^2 - C$ has more than two components. Let A and A' be two of the components and let B be the union of the remaining ones. Because $S^2 - C$ is locally connected, each of the sets A, A', and B is open. Let a, a', and b be points of A, A', and B, respectively. By the preceding theorem, the sets $U = S^2 - C_1$ and $V = S^2 - C_2$ are path connected (since no arc separates S^2). Therefore, a, a', and b may be joined by paths in U and in V. Lemma 13.2 now implies that $\pi_1(X, a)$ is not infinite cyclic.

But X equals $S^2 - p - q$, which is homeomorphic to the punctured plane $R^2 - 0$ and thus has an infinite cyclic fundamental group. Therefore, our assumption that $S^2 - C$ has more than two components must have been mistaken.

Step 2. Now we show that C is the common boundary of W_1 and W_2.

Because S^2 is locally connected, each of the components W_1 and W_2 of $S^2 - C$ is open in S^2. In particular, neither contains a limit point of the other, so that $\bar{W}_1 \cap W_2 = W_1 \cap \bar{W}_2 = \varnothing$. Since S^2 is the disjoint union of W_1, W_2, and C, both the sets $\bar{W}_1 - W_1$ and $\bar{W}_2 - W_2$ must be contained in C.

To prove the converse, we show that if x is a point of C, every neighborhood U of x intersects the closed set $\bar{W}_1 - W_1$. It follows that x is in the set $\bar{W}_1 - W_1$.

So let U be a neighborhood of x. Because C is homeomorphic to the circle S^1, we can break C up into two arcs C_1 and C_2 that intersect only in their end points, such that C_1 is small enough that it lies inside U. See Figure 36.

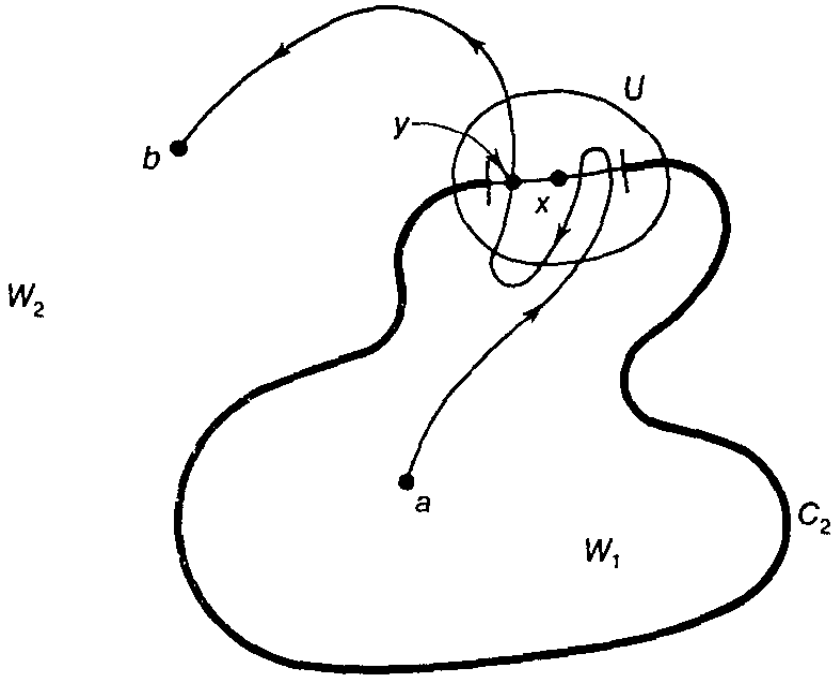


Figure 36

Let a and b be points of W_1 and W_2, respectively. Because C_2 does not separate S^2 (by Theorem 13.3), we can find a path α in $S^2 - C_2$ joining a and b. The set $\alpha(I)$ must contain a point y of the set $\bar{W}_1 - W_1$, because otherwise $\alpha(I)$ would be a connected set lying in the union of the disjoint open sets W_1 and $S^2 - \bar{W}_1$ and intersecting each of them. The point y necessarily belongs to the closed curve C, since $(\bar{W}_1 - W_1) \subset C$. Because the path α does not intersect the arc C_2, the point y must therefore lie in the arc C_1, which in turn lies in the open set U. Thus U intersects $\bar{W}_1 - W_1$ in the point y, as desired. □

EXAMPLE 1. The second half of the Jordan curve theorem, to the effect that C is the common boundary of W_1 and W_2, may seem so obvious as hardly to require comment. But, in fact, it depends crucially on the fact that C is homeomorphic to S^1.

For example, the subset C of the plane consisting of the union of the unit circle S^1 and the line segment $[1, 2] \times 0$ separates S^2 into two components W_1 and W_2 just as the circle does. But C does not equal the common boundary of W_1 and W_2 in this case. See Figure 37.

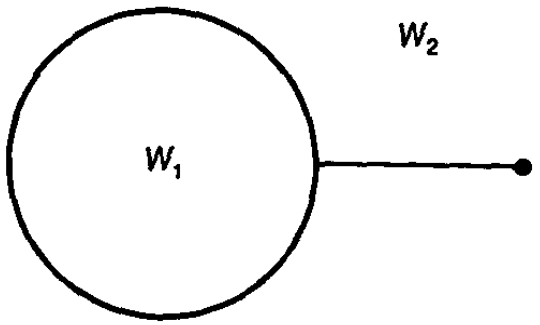


Figure 37

There is a fourth theorem that is often considered along with these three separation theorems. It is called the *Schoenflies theorem*, and it states that if C is a simple closed curve in S^2 and U and V are the components of $S^2 - C$, then $\bar{U}$ and $\bar{V}$ are each homeomorphic to the closed unit ball B^2. We are not going to prove this theorem. It is, however, a corollary of the Riemann mapping theorem of complex variable theory.

The separation theorems can be generalized to higher dimensions as follows:

(1) Any subset C of S^n homeomorphic to S^{n-1} separates S^n.

(2) No subspace A of S^n homeomorphic to $[0, 1]$ or to some ball B^m separates S^n.

(3) Any subset C of S^n homeomorphic to S^{n-1} separates S^n into two components, of which C is the common boundary.

These theorems can be proved quite readily once one has studied singular homology groups in algebraic topology. (See [S], p. 198.) The Brouwer theorem on invariance of domain for R^n follows as a corollary.

The Schoenflies theorem, however, does not generalize to higher dimensions without some restrictions on the way the space C is imbedded in S^n. This is shown by the famous example of the "Alexander horned sphere," a homeomorphic image of S^2 in S^3, one of whose complementary domains is not simply connected! (See [H-Y], p. 176.)

The separation theorems can be generalized even further than this. The definitive theorem along these lines is the famous *Alexander–Pontryagin duality theorem*, a rather deep theorem of algebraic topology, which we shall not attempt to state here. (See [S].) It implies that if C separates S^n into k components, so does any subset of S^n that is homeomorphic to C (or even homotopy equivalent to C). The separation theorems (1)–(3) are immediate corollaries.

Exercises

1. (a) Show that no arc in R^2 separates R^2.
   (b) Show that a simple closed curve C in R^2 separates R^2 into two components, of which C is the common boundary.

2. Show that under the hypotheses of Lemma 13.1, $\pi_1(X, a)$ contains an infinite cyclic subgroup.

3. *Lemma. Let A and B be closed connected subsets of S^2 whose intersection consists of precisely two points. If neither A nor B separates S^2, then $S^2 - (A \cup B)$ has precisely two components.*

4. Recall that a *theta space* is a space which is the union of three arcs a, b, and c, each two of which intersect precisely in their end points.
   (a) Show that a theta space in S^2 separates S^2 into precisely three components. [*Hint:* If C is the component of $S^2 - (a \cup b)$ not intersecting c, then $\bar{C} \cup c$ separates S^2 into precisely two components, by the preceding exercise.]
   (b) Show that these three components have boundaries $a \cup b$, $a \cup c$, and $b \cup c$, respectively. [*Hint:* Show that C is one of these components; then apply symmetry.]
   (c) *Theorem. The gas-water-electricity graph cannot be imbedded in S^2.* (See Figure 13 of §7-9.)
   [*Hint:* Consider the theta space obtained by joining H_1 and H_2 to G, W, and E.]

5. Suppose that a_1, a_2, a_3, and a_4 are four distinct points of S^2, and that for each pair a_i, a_j of these points ($i \neq j$), we have an arc a_ia_j in S^2 joining them. Suppose, moreover, that each pair of these arcs intersect in at most a common end point. The union G_4 of these arcs is called the complete graph on four vertices.
   (a) Let C be the curve $(a_1a_2) \cup (a_2a_3) \cup (a_3a_4) \cup (a_4a_1)$, which we denote by $a_1a_2a_3a_4a_1$ for short. Let A and B be the components of $S^2 - C$. Show that if A is disjoint from a_1a_3, then B is disjoint from a_2a_4. [*Hint:* Assume Exercise 4. Note that A is one of the three components of $S^2 - (C \cup a_1a_3)$, and these three components have boundaries C, $a_1a_2a_3a_1$, and $a_1a_3a_4a_1$, respectively.]
   (b) Show that $S^2 - G_4$ has four components, whose boundaries are $a_1a_2a_3a_1$, $a_1a_3a_4a_1$, $a_1a_2a_4a_1$, and $a_2a_3a_4a_2$, respectively.
   (c) *Theorem. The complete graph on five vertices cannot be imbedded in S^2.* (See Figure 13 of §7-9.)

6. (a) Assume the hypotheses and notation of Lemma 13.1. Suppose U and V are path connected and $\pi_1(X, a)$ is infinite cyclic. Show that E is simply connected and $[\alpha * \beta]$ *generates* $\pi_1(X, a)$. [*Hint:* Use Exercise 10 of §8-4.]
   (b) Consider the complete graph G_4 on four vertices a, b, c, d; suppose G_4 is a subset of S^2. Show that if p is an interior point of the arc ac and q is an interior point of the arc bd, then the loop *abcda generates* the fundamental group of $S^2 - p - q$. [*Hint:* Let $U = S^2 - pabq$ and $V = S^2 - pcdq$. Let x be an interior point of ad and y be an interior point of bc; by Exercise

5(a), they lie in different components of $U \cap V$. Let $\alpha = xaby$ and $\beta = ycdx$.]

7. Prove the following:

Theorem. Let A be a simple closed curve in R^2; let $\mathbf{0}$ lie in the bounded component of $R^2 - A$. Then A generates *the fundamental group of $R^2 - \mathbf{0}$.*

Proof. Let $S = R^2 \cup \{\infty\}$ be the one-point compactification of R^2. Let B be the bounded component of $R^2 - A$.

(a) Construct an arc D in S containing ∞ that intersects A in precisely its end points a and c.

(b) Construct an arc E in S containing $\mathbf{0}$ that intersects A in precisely its end points b and d, where b and d lie in different components of $A - a - c$. [*Hint*: Show that if x and y are two points of the open, connected subset U of S^2, then (i) there is an arc in U joining x and y, and (ii) there is an homeomorphism $h: S^2 \to S^2$ which equals the identity outside U and maps x to y. Let A_1 and A_2 be the components of $A - a - c$. Choose a point x of B; join x to ∞ by arcs in $S - \bar{A}_1$ and $S - \bar{A}_2$; obtain arcs joining x to b and d; then an arc from b to d; then apply (ii) to obtain an arc containing $\mathbf{0}$.]

(c) Apply Exercise 6.

8. Prove the following basic theorem of complex analysis:

Theorem. Let A be a piecewise-differentiable simple closed curve in the complex plane C; *let z_0 be a point of* C *not on A. Then the integral*

$$\frac{1}{2\pi i}\int_A \frac{dz}{z - z_0}$$

has value ± 1 if z_0 lies in the bounded component of C $- A$, *and it has value 0 if z_0 lies in the unbounded component.*

[*Hint:* Use the preceding exercise and Exercise 6 of §8-5.]

9. A proof of the invariance of domain theorem for R^2 was outlined in the exercises of §8-12. Give an independent proof based on Exercise 7 of this section.

8-14 The Classification of Covering Spaces

Up to now we have used covering spaces primarily as a tool for computing fundamental groups. Now we are going to turn things around and use the fundamental group as a tool for studying covering spaces. It turns out that one can determine all possible covering spaces of a given space B, merely by examining the fundamental group of B. We explain this process now.

If H_1 and H_2 are subgroups of a group G, you may recall from algebra that they are said to be **conjugate** subgroups if $H_2 = \alpha \cdot H_1 \cdot \alpha^{-1}$ for some element α of G. Said differently, they are conjugate if the isomorphism of G with itself which maps x to $\alpha \cdot x \cdot \alpha^{-1}$ carries the group H_1 onto the group H_2. It is easy to check that conjugacy is an equivalence relation on the collection

of subgroups of G. The equivalence class of the subgroup H is called the **conjugacy class** of H.

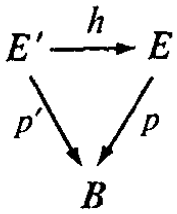
Now suppose we have a fixed path-connected space B and a fixed base point $b_0 \in B$. Let $p : E \longrightarrow B$ be a covering map, where E is path connected. If we choose a point e_0 in $p^{-1}(b_0)$ as a base point for E, then we have an induced homomorphism

$$p_* : \pi_1(E, e_0) \longrightarrow \pi_1(B, b_0)$$

(which is, in fact, injective). The image group $H_0 = p_*(\pi_1(E, e_0))$ depends of course on the choice of the base point e_0 in E, but in a very simple way. We shall show that as e_0 ranges over all the various points of $p^{-1}(b_0)$, the image subgroup ranges over those subgroups of G that are conjugate to H_0. Thus to each path-connected covering space $p : E \longrightarrow B$ of B we can assign a certain conjugacy class of subgroups of $\pi_1(B, b_0)$, the collection of the images of the groups $\pi_1(E, e)$ under the homomorphisms induced by p, for $e \in p^{-1}(b_0)$.

It turns out that every conjugacy class corresponds to a covering space of B under this assignment, and that this conjugacy class determines the covering space *uniquely*, up to homeomorphism. This is the basic classification theorem for covering spaces.

More precisely, we define two covering spaces $p : E \longrightarrow B$ and $p' : E' \longrightarrow B$ to be **equivalent** if there is a homeomorphism $h : E' \longrightarrow E$ such that $p \circ h = p'$.



Then the two main theorems concerning classification of covering spaces can be stated as follows:

(1) For every conjugacy class of subgroups of $\pi_1(B, b_0)$, there exists a path-connected covering space $p : E \longrightarrow B$ of B such that the groups $p_*(\pi_1(E, e))$, for e in $p^{-1}(b_0)$, form this conjugacy class.

(2) Two path-connected covering spaces of B are equivalent if and only if they correspond to the same conjugacy class of subgroups of $\pi_1(B, b_0)$ under this correspondence.

By means of these theorems, the topological problem of determining all possible path-connected covering spaces of a given space B is reduced completely to the algebraic problem of determining all conjugacy classes of subgroups of $\pi_1(B, b_0)$.

We should note, however, that these theorems do not hold for a completely arbitrary path-connected space B. We shall need to assume certain "local niceness" conditions on B in order to carry out the proofs. (See Theorems 14.3 and 14.4.)

Lemma 14.1. *Let* $\mathrm{p}: \mathrm{E} \longrightarrow \mathrm{B}$ *be a covering map, where* E *and* B *are path connected. Let* $\mathrm{b}_0 \in \mathrm{B}$. *As* e *ranges over the points of* $\mathrm{p}^{-1}(\mathrm{b}_0)$, *the group* $\mathrm{p}_*(\pi_1(\mathrm{E}, \mathrm{e}))$ *ranges precisely over a conjugacy class of subgroups of* $\pi_1(\mathrm{B}, \mathrm{b}_0)$.

Proof. Given $e_0, e_1 \in p^{-1}(b_0)$, let

$$H_0 = p_*(\pi_1(E, e_0)) \quad \text{and} \quad H_1 = p_*(\pi_1(E, e_1)).$$

Step 1. Let γ be a path in E from e_0 to e_1; let α be the loop $p \circ \gamma$ in B based at b_0. We show that

$$[\alpha] * H_1 * [\alpha]^{-1} \subset H_0.$$

Let $[h]$ be an element of H_1; then $[h] = p_*[\tilde{h}]$ for some loop $\tilde{h}$ in E based at e_1. Let $\tilde{k} = (\gamma * \tilde{h}) * \bar{\gamma}$. Then $\tilde{k}$ is a loop based at e_0, and

$$p_*([\tilde{k}]) = [(\alpha * h) * \bar{\alpha}].$$

Thus $[\alpha] * [h] * [\alpha]^{-1} \in H_0$, as desired. See Figure 38.

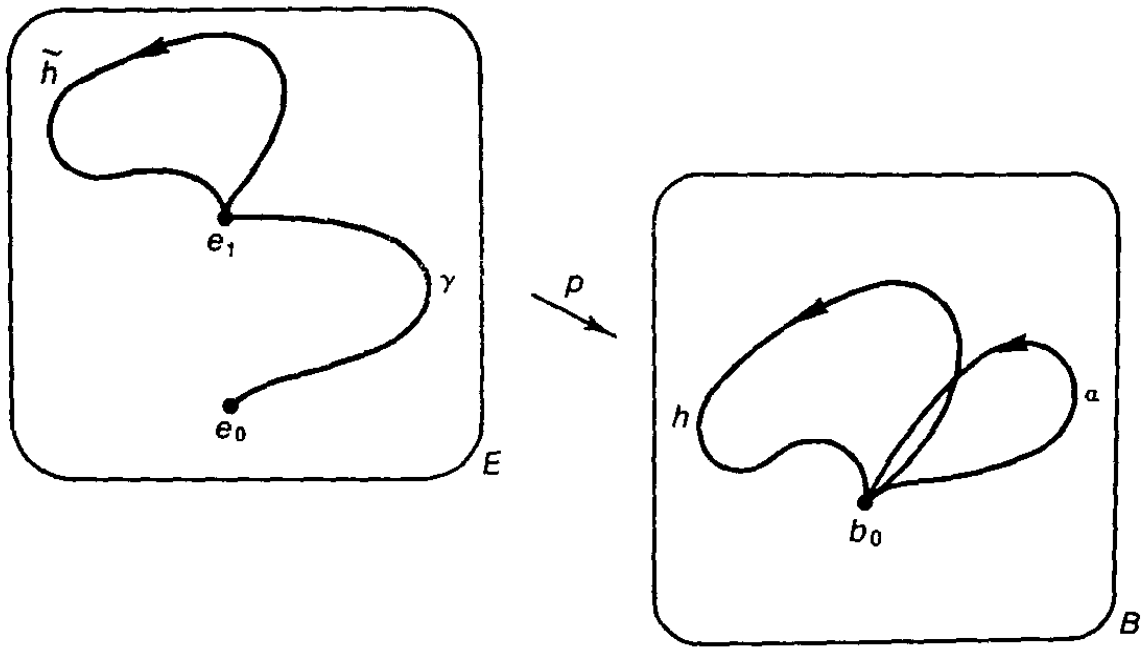


Figure 38

Step 2. It follows that H_0 and H_1 are conjugate subgroups. For let γ be a path in E from e_0 to e_1, and let $\alpha = p \circ \gamma$. Then $\bar{\gamma}$ is a path from e_1 to e_0, and $\bar{\alpha} = p \circ \bar{\gamma}$. Applying Step 1 twice, we have

$$[\alpha] * H_1 * [\alpha]^{-1} \subset H_0,$$
$$[\bar{\alpha}] * H_0 * [\bar{\alpha}]^{-1} \subset H_1.$$

Thus $H_0 = [\alpha] * H_1 * [\alpha]^{-1}$, as desired.

Step 3. Now given $e_0 \in p^{-1}(b_0)$, and given a subgroup H of $\pi_1(B, b_0)$ conjugate to H_0, we wish to find a point e_1 of $p^{-1}(b_0)$ such that the corresponding subgroup H_1 equals H. By hypothesis, $H_0 = [\alpha] * H * [\alpha]^{-1}$ for

some loop α in B based at b_0. Let γ be the lifting of α to a path in E beginning at e_0, and let $e_1 = \gamma(1)$. Let $H_1 = p_*(\pi_1(E, e_1))$. It follows from Step 2 that $H_0 = [\alpha] * H_1 * [\alpha]^{-1}$. Then $H = H_1$, as desired. □

Equivalence of coverings

Now we prove that two coverings are equivalent if and only if they correspond to the same conjugacy class. We need the following generalized version of the "lifting lemmas" of §8-4:

Lemma 14.2 (The general lifting lemma). *Let* $p : E \longrightarrow B$ *be a covering map; let* $p(e_0) = b_0$. *Let* $f : Y \longrightarrow B$ *be a continuous map, with* $f(y_0) = b_0$. *Suppose* Y *is path connected and locally path connected. The map* f *can be lifted to a map* $\tilde{f} : Y \longrightarrow E$ *such that* $\tilde{f}(y_0) = e_0$ *if and only if*

$$f_*(\pi_1(Y, y_0)) \subset p_*(\pi_1(E, e_0)).$$

Furthermore, if such a lifting exists, it is unique.

Proof. If the lifting $\tilde{f}$ exists, then

$$f_*(\pi_1(Y, y_0)) = p_*(\tilde{f}_*(\pi_1(Y, y_0))) \subset p_*(\pi_1(E, e_0)).$$

This proves the "only if" part of the theorem.

Now we prove that if $\tilde{f}$ exists, it is unique. Given $y_1 \in Y$, choose a path α in Y from y_0 to y_1. Take the path $f \circ \alpha$ in B and lift it to a path γ in E beginning at e_0. If a lifting $\tilde{f}$ of f exists, then $\tilde{f}(y_1)$ must equal the end point $\gamma(1)$ of γ, for $\tilde{f} \circ \alpha$ is a lifting of $f \circ \alpha$ that begins at e_0, and path liftings are unique.

Finally, we prove the "if" part of the theorem. The uniqueness part of the proof gives us a clue how to proceed. Given $y_1 \in Y$, choose a path α in Y from y_0 to y_1. Lift the path $f \circ \alpha$ to a path γ in E beginning at e_0, and define $\tilde{f}(y_1) = \gamma(1)$. See Figure 39. It is a certain amount of work to show that $\tilde{f}$ is well-defined, independent of the choice of α. Once we prove that, continuity of $\tilde{f}$ is proved easily, as follows:

Given a neighborhood N of $\tilde{f}(y_1)$, we shall find neighborhood W of y_1 such that $\tilde{f}(W) \subset N$. First, choose a neighborhood U of $f(y_1)$ that is evenly covered by p. Let V_0 be the slice of $p^{-1}(U)$ that contains $\tilde{f}(y_1)$. Let p_0 be the restriction of p to V_0. By passing to smaller neighborhoods if necessary, we can assume that $V_0 \subset N$. Now choose a neighborhood W of y_1 that is path connected and lies in $f^{-1}(U)$. We claim that $\tilde{f}(W) \subset V_0$. For, given $y \in W$, we can choose a path β in W from y_1 to y. Consider the path $f \circ \beta$; lift it to the path $\delta = p_0^{-1} \circ f \circ \beta$ in E, beginning at $\tilde{f}(y_1)$. Then $\gamma * \delta$ is defined, it is a lifting of $f \circ (\alpha * \beta)$, and it begins at e_0. By definition, $\tilde{f}(y)$ equals the end point of $\gamma * \delta$, which lies in V_0.

Now we show $\tilde{f}$ is well defined. Given two paths α and β in Y from y_0 to y_1, their images $f \circ \alpha$ and $f \circ \beta$ are paths in B. Let γ and δ denote their

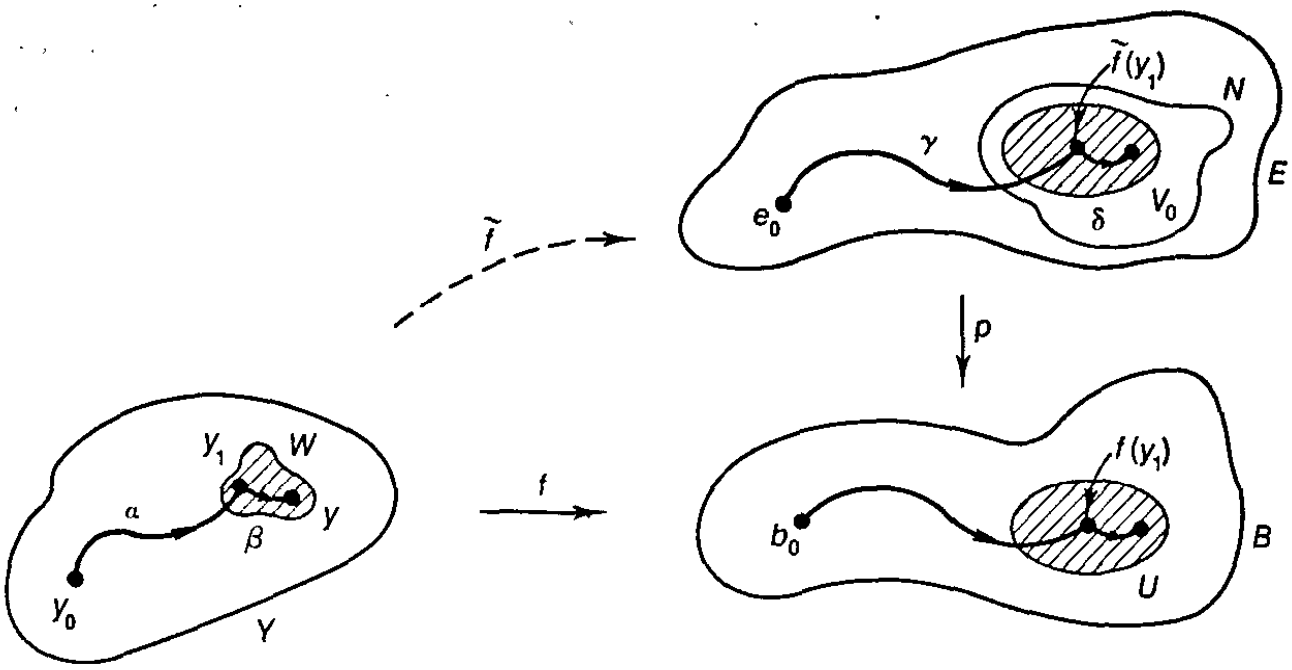


Figure 39

respective liftings to paths on E beginning at e_0. We wish to show that $\gamma(1) = \delta(1)$.

Let ϵ be a lifting of $f \circ \tilde{\beta}$ to a path on E beginning at $\gamma(1)$. See Figure 40. Then $\gamma * \epsilon$ is defined and is a lifting to E of the loop $(f \circ \alpha) * (f \circ \bar{\beta})$ in B. The homotopy class of this loop is $f_*([\alpha * \bar{\beta}])$, which by hypothesis belongs to $p_*(\pi_1(E, e_0))$. Then there is a loop ϕ in E based at e_0 such that

$$[(f \circ \alpha) * (f \circ \bar{\beta})] = [p \circ \phi].$$

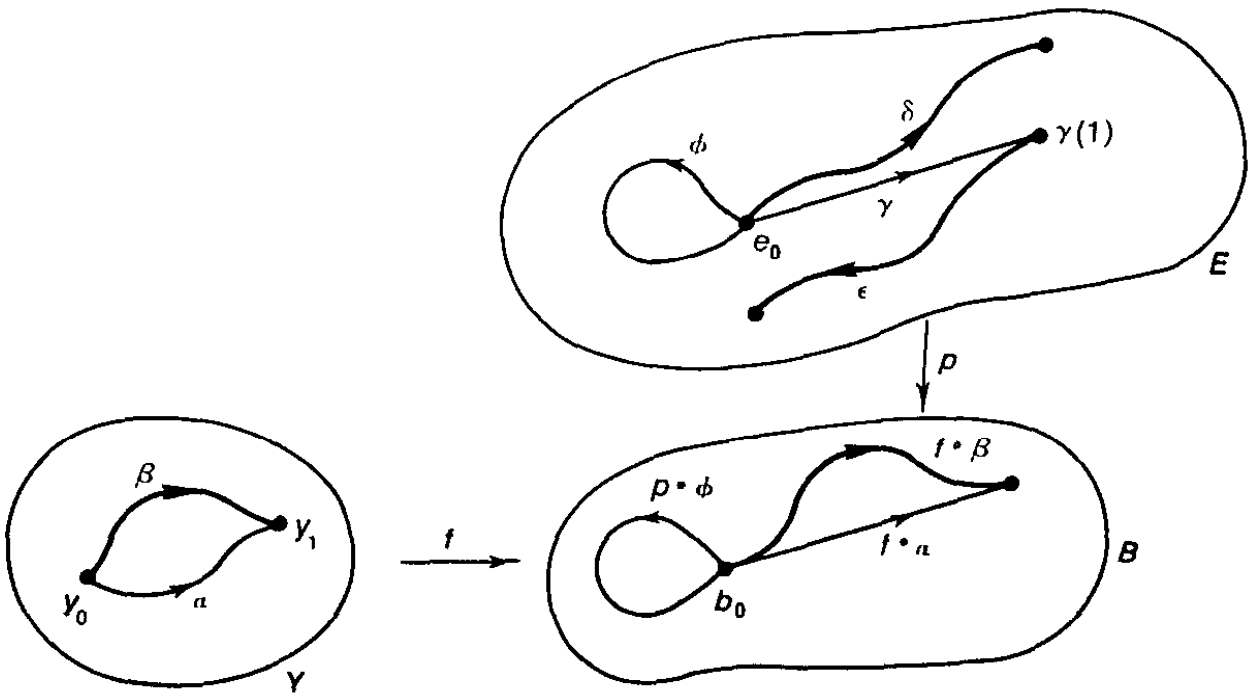


Figure 40

It follows that ϵ must end at e_0. For by Theorem 4.3, if two paths in B which are path homotopic are lifted to paths on E beginning at the same point, they must end at the same point. Since $p \circ \phi$ and $(f \circ \alpha) * (f \circ \tilde{\beta})$ are path homotopic, their liftings ϕ and $\gamma * \epsilon$ must end at the same point. Then $\gamma * \epsilon$ ends at e_0, so ϵ ends at e_0.

Now ϵ is a lifting of $f \circ \tilde{\beta}$ that begins at $\gamma(1)$ and ends at e_0. Then $\tilde{\epsilon}$ is a lifting of $f \circ \beta$ that begins at e_0 and ends at $\gamma(1)$. The path δ is another such lifting. By uniqueness of path liftings, $\delta = \tilde{\epsilon}$. In particular, $\delta(1) = \gamma(1)$, as desired. □

Theorem 14.3. *Let* B *be path connected and locally path connected. Let* $p : E \to B$ *and* $p' : E' \to B$ *be path-connected covering spaces of B; let* $p(e_0) = p'(e_0') = b_0$. *Then* p *and* p′ *are equivalent covering maps if and only if* $p_*(\pi_1(E, e_0))$ *and* $p'_*(\pi_1(E', e_0'))$ *are conjugate subgroups of* $\pi_1(B, b_0)$.

Proof. Suppose that the coverings are equivalent. Let $h : E' \to E$ be a homeomorphism such that $p \circ h = p'$. Let $h(e_0') = e_1$. Then because h is a homeomorphism, $h_*(\pi_1(E', e_0')) = \pi_1(E, e_1)$. Applying p_* to both sides, we have

$$p'_*(\pi_1(E', e_0')) = p_*(\pi_1(E, e_1)).$$

By Lemma 14.1, the latter subgroup is conjugate to $p_*(\pi_1(E, e_0))$.

Now we prove the converse. Suppose that the two subgroups are conjugate. In view of Lemma 14.1, we may by choosing a different base point in E' obtain the situation where the two subgroups are equal. So let us suppose this is done, and let e_0' now denote the new base point.

We apply the preceding lemma. Consider the diagram

$$\begin{array}{ccc} & & E' \\ & & \downarrow p' \\ E & \xrightarrow{p} & B \end{array}$$

where p' is a covering map. The space E is path connected; it is also locally path connected, being locally homeomorphic to B. Furthermore,

$$p_*(\pi_1(E, e_0)) \subset p'_*(\pi_1(E', e_0'));$$

in fact, these two groups are equal. By the preceding lemma, we can lift the map $p : E \to B$ to a continuous map $h : E \to E'$ such that $h(e_0) = e_0'$. Then $p' \circ h = p$.

Reversing the roles of E and E' in this argument, we see that the map $p' : E' \to B$ can be lifted to a continuous map $k : E' \to E$ with $k(e_0') = e_0$.

To show that h and k are inverses of each other, consider the diagram

$$\begin{array}{ccc} & & E \\ & & \downarrow p \\ E & \xrightarrow{p} & B \end{array}$$

Note that $k \circ h : E \longrightarrow E$ is a lifting to E of the map $p : E \longrightarrow B$, satisfying the condition $(k \circ h)(e_0) = e_0$. The identity map $i_E : E \longrightarrow E$ is another such lifting of p. By uniqueness of liftings, $k \circ h$ must equal i_E. Similarly, $h \circ k = i_{E'}$. □

EXAMPLE 1. Consider covering spaces of the circle $B = S^1$. Because $\pi_1(B, b_0)$ is abelian, two subgroups of $\pi_1(B, b_0)$ are conjugate if and only if they are equal. Therefore two coverings of B are equivalent if and only if they correspond to the same subgroup of $\pi_1(B, b_0)$.

Now $\pi_1(B, b_0)$ is isomorphic to the integers Z. What are the subgroups of Z? It is standard theorem of modern algebra that, given a nontrivial subgroup of Z, it must be the group G_n consisting of all multiples of n, for some $n \in Z_+$.

We have studied one covering space of the circle, the covering $p : R \longrightarrow S^1$. It must correspond to the trivial subgroup of $\pi_1(S^1, b_0)$, because R is simply connected. We have also considered the covering $p : S^1 \longrightarrow S^1$ defined by $p(z) = z^n$, where z is a complex number. In this case, the map p_* carries a generator of $\pi_1(S^1, b_0)$ into n times itself. Therefore, the group $p_*(\pi_1(S^1, b_0))$ corresponds to the subgroup G_n of Z under the standard isomorphism of $\pi_1(S^1, b_0)$ with Z.

We conclude from the preceding theorem that every path-connected covering space of S^1 is equivalent to one of these coverings.

Existence of coverings

Now we show that given a subgroup H of $\pi_1(B, b_0)$, there exists a covering space $p : E \longrightarrow B$ and a point e_0 of $p^{-1}(b_0)$ such that $p_*(\pi_1(E, e_0)) = H$. We need to assume an additional condition on the space B, called "semilocal simple connectivity."

Definition. A space B is said to be **semilocally simply connected** if every point b of B has a neighborhood V such that the homomorphism of $\pi_1(V, b)$ into $\pi_1(B, b)$ induced by inclusion is trivial.

If V is such a neighborhood of b and U is any neighborhood of b, then the neighborhood $U \cap V$ of b, which lies in U, also satisfies this condition. Semilocal simple connectivity is weaker than true local simple connectivity, which would require that within each neighborhood U of b there should exist a neighborhood V of b that is itself simply connected.

Theorem 14.4. *Let* B *be path connected, locally path connected, and semilocally simply connected. Let* $b_0 \in B$. *Given a subgroup* H *of* $\pi_1(B, b_0)$, *there exists a path connected covering space* $p : E \longrightarrow B$ *and a point* $e_0 \in p^{-1}(b_0)$ *such that*

$$p_*(\pi_1(E, e_0)) = H.$$

Proof. Step 1. (*Construction of E*). The procedure for constructing E is reminiscent of the procedure used in complex analysis for constructing Riemann surfaces. Let $\mathcal{P}$ denote the set of all paths in B beginning at b_0. Define an equivalence relation on $\mathcal{P}$ by setting $\alpha \sim \beta$ if α and β end at the same point of B and the class $[\alpha * \bar{\beta}]$ belongs to the subgroup H. This is easily seen to be an equivalence relation.

Let E denote the collection of equivalence classes; denote the equivalence class of the path α by $\alpha^{\#}$. Define $p : E \to B$ by the equation $p(\alpha^{\#}) = \alpha(1)$. Because B is path connected, p is surjective. We shall topologize E in such a way that p is a covering map.

We first note two facts:

(1) If $\alpha \simeq_p \beta$, then $\alpha^{\#} = \beta^{\#}$.

(2) If $\alpha^{\#} = \beta^{\#}$, then $(\alpha * \delta)^{\#} = (\beta * \delta)^{\#}$ for any path δ in B beginning at $\alpha(1)$.

The first follows by noting that if $\alpha \simeq_p \beta$, then $[\alpha * \bar{\beta}]$ is the identity element, which belongs to H. The second follows by noting that $\alpha * \delta$ and $\beta * \delta$ end at the same point, and

$$[(\alpha * \delta) * \overline{(\beta * \delta)}] = [(\alpha * \delta) * (\bar{\delta} * \bar{\beta})] = [\alpha * \bar{\beta}],$$

which belongs to H by hypothesis.

Step 2. (*Topologizing E*). One way to topologize E is to give $\mathcal{P}$ the compact-open topology (see Chapter 7) and E the corresponding quotient topology. But we can topologize E directly as follows:

Let α be any element of $\mathcal{P}$, and let U be any path-connected neighborhood of $\alpha(1)$. Define

$$B(U, \alpha) = \{(\alpha * \delta)^{\#} \mid \delta \text{ is a path in } U \text{ beginning at } \alpha(1)\}.$$

We assert that the sets $B(U, \alpha)$ form a basis for a topology on E. See Figure 41. Note that $\alpha^{\#}$ belongs to $B(U, \alpha)$.

First we show that if $\beta^{\#} \in B(U, \alpha)$, then $B(U, \beta) = B(U, \alpha)$. We are given that $\beta^{\#} = (\alpha * \delta)^{\#}$ for some path δ in U. The general element of $B(U, \beta)$ is of the form $(\beta * \gamma)^{\#}$, for some path γ in U. By (1) and (2) above, we have

$$(\beta * \gamma)^{\#} = ((\alpha * \delta) * \gamma)^{\#} = (\alpha * (\delta * \gamma))^{\#},$$

which lies in $B(\alpha, U)$ by definition. Hence $B(U, \beta) \subset B(U, \alpha)$. Conversely, the general element of $B(U, \alpha)$ is of the form $(\alpha * \epsilon)^{\#}$, where ϵ is a path in U. Again by (1) and (2),

$$(\alpha * \epsilon)^{\#} = ((\alpha * \delta) * (\bar{\delta} * \epsilon))^{\#} = (\beta * (\bar{\delta} * \epsilon))^{\#},$$

which lies in $B(U, \beta)$. Thus $B(U, \alpha) \subset B(U, \beta)$.

Now we show the sets $B(U, \alpha)$ form a basis. If $\beta^{\#}$ belongs to the intersection $B(U_1, \alpha_1) \cap B(U_2, \alpha_2)$, we need merely choose a path-connected

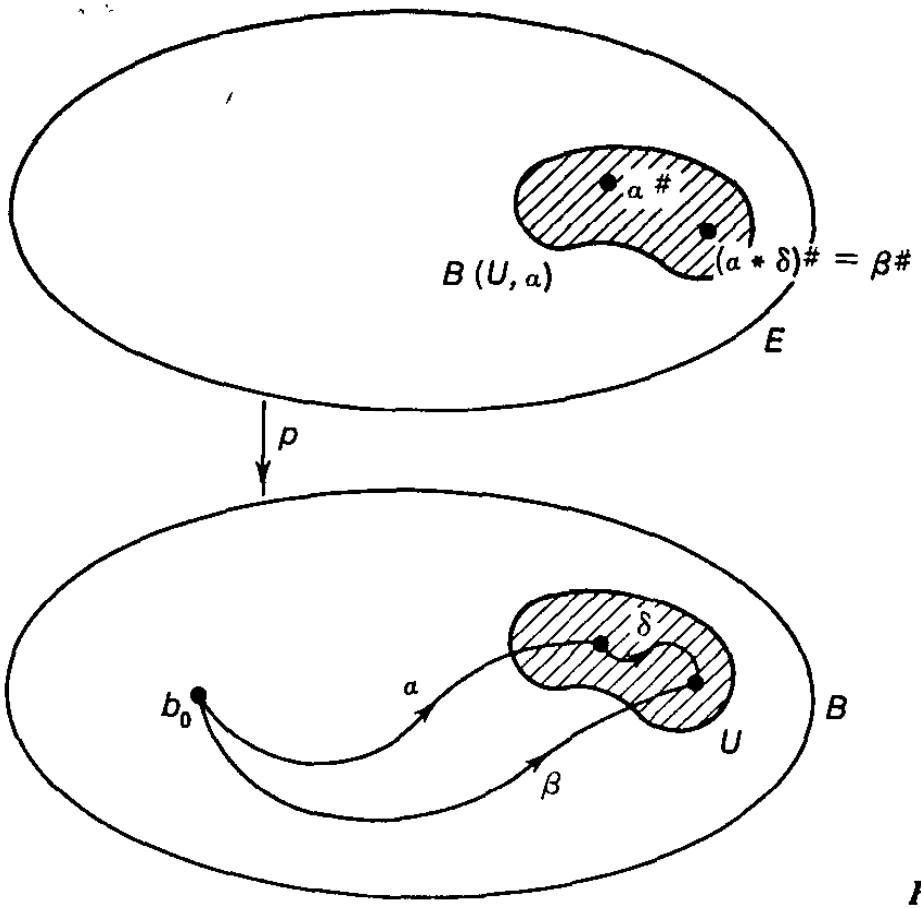


Figure 41

neighborhood V of $\beta(1)$ contained in $U_1 \cap U_2$. The inclusion

$$B(V, \beta) \subset B(U_1, \beta) \cap B(U_2, \beta)$$

follows from the definition of these sets, and the right side of the equation equals $B(U_1, \alpha_1) \cap B(U_2, \alpha_2)$ by the result just proved.

Step 3. The map p is continuous and open. It is easy to see that p is open, for the image of the basis element $B(U, \alpha)$ is the open subset U of B: Given $x \in U$, we choose a path δ in U from $\alpha(1)$ to x; then $(\alpha * \delta)\#$ is in $B(U, \alpha)$ and $p((\alpha * \delta)\#) = x$.

To show that p is continuous, let us take an element $\alpha\#$ of E and a neighborhood W of $p(\alpha\#)$. Choose a path-connected neighborhood U of the point $p(\alpha\#) = \alpha(1)$ lying in W. Then $B(U, \alpha)$ is a neighborhood of $\alpha\#$ that p maps into W. Thus p is continuous at $\alpha\#$.

Step 4. The map p is a covering map. Given $b_1 \in B$, choose U to be a path-connected neighborhood of b_1 that satisfies the further condition that the homomorphism $\pi_1(U, b_1) \longrightarrow \pi_1(B, b_1)$ induced by inclusion is trivial. We assert that U is evenly covered by p.

First, note that $p^{-1}(U)$ equals the union of the sets $B(U, \alpha)$, as α ranges over all paths in B from b_0 to b_1: Since p maps each set $B(U, \alpha)$ onto U, it is clear that $p^{-1}(U)$ contains this union. On the other hand, if $\beta\#$ is in $p^{-1}(U)$, then $\beta(1)$ is in U and we can choose a path δ in U from $\beta(1)$ to b_1. Let α be the

path $\beta * \delta$ from b_0 to b_1: Then $\beta \simeq_p \alpha * \bar{\delta}$, so that $\beta^\# = (\alpha * \bar{\delta})^\#$, which belongs to $B(U, \alpha)$. Thus $p^{-1}(U)$ is contained in the union of the sets $B(U, \alpha)$.

Second, note that distinct sets $B(U, \alpha)$ are disjoint. For if $\beta^\#$ belongs to $B(U, \alpha_1) \cap B(U, \alpha_2)$, then $B(U, \alpha_1) = B(U, \beta) = B(U, \alpha_2)$, by Step 2.

Third, we show that p defines a bijective map of $B(U, \alpha)$ with U. It follows that $p \mid B(U, \alpha)$ is a homeomorphism, being bijective and continuous and open. We already know that p maps $B(U, \alpha)$ *onto* U. To prove injectivity, suppose that

$$p((\alpha * \delta_1)^\#) = p((\alpha * \delta_2)^\#),$$

where δ_1 and δ_2 are paths in U. Then $\delta_1(1) = \delta_2(1)$. Because the homomorphism $\pi_1(U, b_1) \longrightarrow \pi_1(B, b_1)$ induced by inclusion is trivial, $\delta_1 * \bar{\delta}_2$ is path homotopic in B to the constant loop. Then $\alpha * \delta_1 \simeq_p \alpha * \delta_2$, so that we have $(\alpha * \delta_1)^\# = (\alpha * \delta_2)^\#$, as desired.

Step 5. The space E is path connected and the subgroup H equals $p_*(\pi_1(E, e_0))$ for some e_0 in $p^{-1}(b_0)$. This completes the proof of the theorem.

First, we calculate specifically what the lifting of a path α in B looks like. Let α be a path in B beginning at b_0. Given $t \in [0, 1]$, let $\alpha_t : I \longrightarrow B$ denote that "portion" of α between the points b_0 and $\alpha(t)$, defined by the equation

$$\alpha_t(s) = \alpha(ts) \quad \text{for } s \in [0, 1].$$

Then define $\tilde{\alpha} : I \longrightarrow E$ by the equation

$$\tilde{\alpha}(t) = (\alpha_t)^\#.$$

We assert that $\tilde{\alpha}$ is a lifting of α; note that $\tilde{\alpha}$ begins at the point $e_0 = (e_{b_0})^\#$, the equivalence class of the constant path, and ends at the point $\alpha^\#$. It is easy to see that $p \circ \tilde{\alpha} = \alpha$, since $p(\tilde{\alpha}(t)) = p((\alpha_t)^\#) = \alpha_t(1) = \alpha(t \cdot 1)$. The hard part is to show that $\tilde{\alpha}$ is continuous.

Once we prove that fact, our theorem follows readily. To show E is path connected, we note that if $\alpha^\#$ is any point of E, then α is a path in B beginning at b_0, and its lifting $\tilde{\alpha}$ is a path in E beginning at e_0 and ending at $\alpha^\#$.

To show that $p_*(\pi_1(E, e_0)) \supset H$, take an element $[\alpha] \in H$. Let $\tilde{\alpha}$ be the lifting of α to a path on E beginning at e_0; then $\tilde{\alpha}$ ends at $\alpha^\#$. Now α is equivalent to the constant path e_{b_0}, because $[\alpha * \bar{e}_{b_0}] = [\alpha] \in H$. Therefore, $\alpha^\# = (e_{b_0})^\#$, so that $\tilde{\alpha}$ is a *loop* in E based at e_0. Since $p_*([\tilde{\alpha}]) = [\alpha]$, it follows that $[\alpha] \in p_*(\pi_1(E, e_0))$. Thus $H \subset p_*(\pi_1(E, e_0))$.

To prove the reverse inclusion, take an element $[\tilde{\alpha}] \in \pi_1(E, e_0)$. Let $\alpha = p \circ \tilde{\alpha}$ and note that $\tilde{\alpha}$ is the unique lifting of α to E beginning at e_0. Therefore, $\tilde{\alpha}$ ends at $\alpha^\#$; since $\tilde{\alpha}$ is a loop in E, we have $\tilde{\alpha}(1) = e_0$. Therefore, $\alpha^\# = (e_{b_0})^\#$, so α is equivalent to e_{b_0}, and hence $[\alpha]$ belongs to H. Thus $p_*(\pi_1(E, e_0)) \subset H$.

We complete the proof by showing that $\tilde{\alpha}$ is continuous. Take a parameter value t_0 and a neighborhood $B(U, \alpha_{t_0})$ in E of the point $\tilde{\alpha}(t_0) = (\alpha_{t_0})^\#$. We shall find a number $\epsilon > 0$ such that whenever $|t_1 - t_0| < \epsilon$, we

have $(\alpha_{t_1})^* \in B(U, \alpha_{t_0})$. This will prove continuity. Choose ϵ small enough that $|t_1 - t_0| < \epsilon \Rightarrow \alpha(t_1) \in U$. We assert that if $|t_1 - t_0| < \epsilon$, then α_{t_1} is path homotopic to $\alpha_{t_0} * \delta$ for some path δ lying in U, whence $\tilde{\alpha}(t_1) = (\alpha_{t_1})^*$ lies in $B(U, \alpha_{t_0})$, as desired.

Assume for convenience in notation that $t_0 < t_1$. Let δ be that part of the path α lying between $\alpha(t_0)$ and $\alpha(t_1)$, reparametrized appropriately. Figure 42

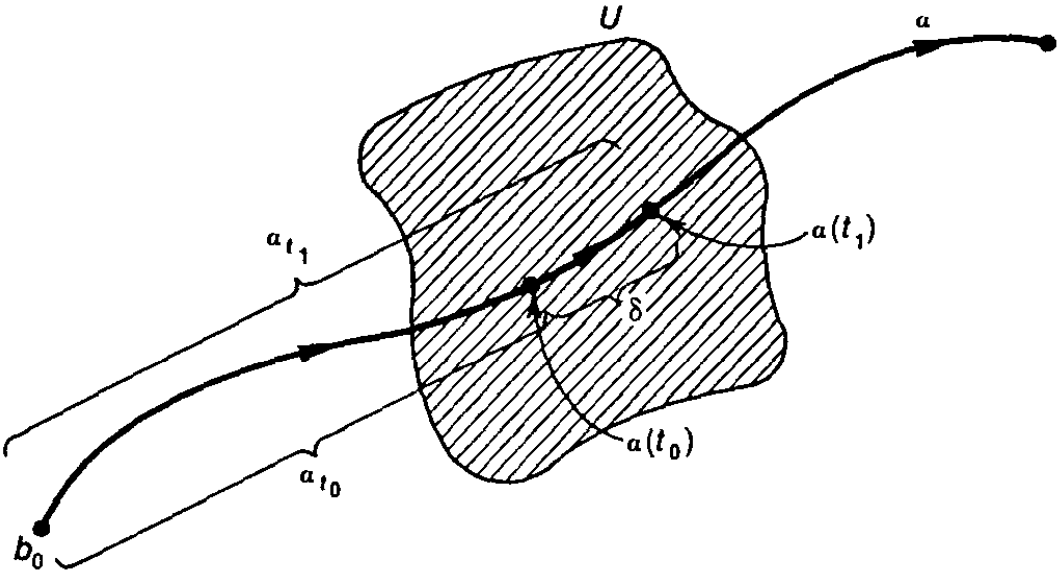


Figure 42

makes it clear that α_{t_1} is path homotopic to $\alpha_{t_0} * \delta$. Formally, let

$$\delta(s) = \alpha((1 - s)t_0 + st_1);$$

then δ lies in U. Define $F : I \times I \longrightarrow B$ by the equation

$$F(s, t) = \begin{cases} \alpha([t_0(1 - t) + t_1 t]2s/[1 + t]) & \text{for } s \in [0, (1 + t)/2], \\ \alpha(t_0(1 - (2s - 1)) + t_1(2s - 1)) & \text{for } s \in [(1 + t)/2, 1]. \end{cases}$$

It is the required path homotopy. □

Corollary 14.5. Let B *be path connected and locally path connected.*

(a) *If* B *has a universal covering space, that covering space is uniquely determined up to equivalence.*

(b) *If* B *is semilocally simply connected, then* B *has a universal covering space.*

Proof. Part (a) is an immediate consequence of Theorem 14.3; if $p : E \longrightarrow B$ is a universal covering space of B, then the group $p_*(\pi_1(E, e_0))$ is the trivial subgroup of $\pi_1(B, b_0)$. To prove (b), we apply the preceding theorem to construct a path-connected covering space $p : E \longrightarrow B$ such that $p_*(\pi_1(E, e_0))$ is the trivial subgroup of $\pi_1(B, b_0)$. Since p_* is injective (a fact we leave to you to prove), this means that E is simply connected and is thus a universal covering space of B. □

Exercises

1. We know that $\pi_1(S^1 \times S^1, b_0 \times b_0)$ is isomorphic to $Z \times Z$, the isomorphism being induced by the projections of $S^1 \times S^1$ onto its two factors.
   (a) Consider the infinite cyclic subgroup of $Z \times Z$ generated by the element 1×0; find the corresponding covering space of $S^1 \times S^1$.
   (b) Find the covering space of $S^1 \times S^1$ corresponding to the infinite cyclic subgroup generated by 1×1.
   (c) Find the covering space corresponding to the subgroup
   $$H = \{(2n, 2m) \mid n, m \in Z\}.$$

2. Let B be a topological space. Show that if B has a universal covering space, then B is semilocally simply connected.

3. Let A_n be a circle of radius $1/n$ in the xy-plane tangent to the y-axis at the origin. Let $X = \bigcup A_n$.
   (a) Show that X is not semilocally simply connected.
   (b) Let $C(X)$ be the subspace of R^3 which is the union of all line segments joining points of X to the point $(0, 0, 1)$. (It is called the *cone* on X.) Show that $C(X)$ is semilocally simply connected at the origin but not locally simply connected there.

4. Let E and B be path connected and locally path connected. Let $p : E \to B$ be a covering map. A homeomorphism $h : E \to E$ is called a **deck transformation** (or **covering transformation**) if $p \circ h = p$.
   (a) Let e_0 and e_1 be points of $p^{-1}(b_0)$. Show that there exists a deck transformation $h : E \to E$ such that $h(e_0) = e_1$ if and only if $p_*(\pi_1(E, e_0)) = p_*(\pi_1(E, e_1))$. Show that if h exists, it is unique.
   (b) If $H_0 = p_*(\pi_1(E, e_0))$ is normal in $\pi_1(B, b_0)$, then $p : E \to B$ is said to be a **regular covering map**. Show that in this case, the group of deck transformations of E is isomorphic to the quotient group $\pi_1(B, b_0)/H_0$. (Compare Exercise 10 of §8-4.)
   (c) If $p : E \to B$ is a universal covering space of B, what can you say about the group of deck transformations of E?

5. Consider the usual covering map $p : R \times R \to S^1 \times S^1$. Describe the group of deck transformations of $R \times R$.

6. Determine all deck transformations of the covering space of the figure eight constructed in Lemma 7.5.

*7. Generalize Exercise 13 of §8-4, concerning covering spaces of topological groups, to the case where the covering space is path connected but not simply connected.

Index

C

D

E

F

G

H

N

O

P

Q

R

S

U

V

W

X

Z