

1. ELEMENTARY POINT SET TOPOLOGY

The branch of mathematics which goes under the name of *topology* is concerned with all questions directly or indirectly related to continuity. The term is traditionally used in a very wide sense and without strict limits. Topological considerations are extremely important for the foundation of the study of analytic functions, and the first systematic study of topology was motivated by this need.

The logical foundations of set theory belong to another discipline. Our approach will be quite naive, in keeping with the fact that all our applications will be to very familiar objects. In this limited framework no logical paradoxes can occur.

1.1. Sets and Elements. In our language a *set* will be a collection of identifiable objects, its *elements*. The reader is familiar with the notation $x \in X$ which expresses that x is an element of X (as a rule we denote sets by capital letters and elements by small letters). Two sets are equal if and only if they have the same elements. X is a subset of Y if every element of X is also an element of Y , and this relationship is indicated by $X \subset Y$ or $Y \supset X$ (we do not exclude the possibility that $X = Y$). The empty set is denoted by \emptyset .

A set can also be referred to as a *space*, and an element as a *point*. Subsets of a given space are usually called *point sets*. This lends a geometric flavor to the language, but should not be taken too literally. For instance, we shall have occasion to consider spaces whose elements are functions; in that case a "point" is a function.

The *intersection* of two sets X and Y , denoted by $X \cap Y$, is formed by all points which are elements of both X and Y . The *union* $X \cup Y$ consists of all points which are elements of either X or Y , including those which are elements of both. One can of course form the intersection and union of arbitrary collections of sets, whether finite or infinite in number.

The *complement* of a set X consists of all points which are not in X ; it will be denoted by $\sim X$. We note that the complement depends on the totality of points under consideration. For instance, a set of real numbers has one complement with respect to the real line and another with respect to the complex plane. More generally, if $X \subset Y$ we can consider the relative complement $Y \sim X$ which consists of all points that are in Y but not in X (we find it clearer to use this notation only when $X \subset Y$).

It is helpful to keep in mind the *distributive laws*

$$\begin{aligned} X \cup (Y \cap Z) &= (X \cup Y) \cap (X \cup Z) \\ X \cap (Y \cup Z) &= (X \cap Y) \cup (X \cap Z) \end{aligned}$$

and the *De Morgan laws*

$$\begin{aligned}\sim(X \cup Y) &= \sim X \cap \sim Y \\ \sim(X \cap Y) &= \sim X \cup \sim Y.\end{aligned}$$

These are purely logical identities, and they have obvious generalizations to arbitrary collections of sets.

1.2. Metric Spaces. For all considerations of limits and continuity it is essential to give a precise meaning to the terms "sufficiently near" and "arbitrarily near." In the spaces \mathbf{R} and \mathbf{C} of real and complex numbers, respectively, such nearness can be expressed by a quantitative condition $|x - y| < \epsilon$. For instance, to say that a set X contains all x sufficiently near to y means that there exists an $\epsilon > 0$ such that $x \in X$ whenever $|x - y| < \epsilon$. Similarly, X contains points arbitrarily near to y if to every $\epsilon > 0$ there exists an $x \in X$ such that $|x - y| < \epsilon$.

What we need to describe nearness in quantitative terms is obviously a distance $d(x, y)$ between any two points. We say that a set S is a *metric space* if there is defined, for every pair $x \in S, y \in S$, a nonnegative real number $d(x, y)$ in such a way that the following conditions are fulfilled:

1. $d(x, y) = 0$ if and only if $x = y$.
2. $d(y, x) = d(x, y)$.
3. $d(x, z) \leq d(x, y) + d(y, z)$.

The last condition is the *triangle inequality*.

For instance, \mathbf{R} and \mathbf{C} are metric spaces with $d(x, y) = |x - y|$. The n -dimensional euclidean space \mathbf{R}^n is the set of real n -tuples

$$x = (x_1, \dots, x_n)$$

with a distance defined by $d(x, y)^2 = \sum_1^n (x_i - y_i)^2$. We recall that we have defined a distance in the extended complex plane by

$$d(z, z') = \frac{2|z - z'|}{\sqrt{(1 + |z|^2)(1 + |z'|^2)}}$$

(see Chap. 1, Sec. 2.4); since this represents the euclidean distance between the stereographic images on the Riemann sphere, the triangle inequality is obviously fulfilled. An example of a function space is given by $C[a, b]$, the set of all continuous functions defined on the interval $a \leq x \leq b$. It becomes a metric space if we define distance by $d(f, g) = \max |f(x) - g(x)|$.

In terms of distance, we introduce the following terminology: For any $\delta > 0$ and any $y \in S$, the set $B(y, \delta)$ of all $x \in S$ with $d(x, y) < \delta$ is called

the ball with center y and radius δ . It is also referred to as the δ -neighborhood of y . The general definition of neighborhood is as follows:

Definition 1. A set $N \subset S$ is called a neighborhood of $y \in S$ if it contains a ball $B(y, \delta)$.

In other words, a neighborhood of y is a set which contains all points sufficiently near to y . We use the notion of neighborhood to define open set:

Definition 2. A set is open if it is a neighborhood of each of its elements.

The definition is interpreted to mean that the empty set is open (the condition is fulfilled because the set has no elements). The following is an immediate consequence of the triangle inequality:

Every ball is an open set.

Indeed, if $z \in B(y, \delta)$, then $\delta' = \delta - d(y, z) > 0$. The triangle inequality shows that $B(z, \delta') \subset B(y, \delta)$, for $d(x, z) < \delta'$ gives $d(x, y) < \delta' + d(y, z) = \delta$. Hence $B(y, \delta)$ is a neighborhood of z , and since z was any point in $B(y, \delta)$ we conclude that $B(y, \delta)$ is an open set. For greater emphasis a ball is sometimes referred to as an *open ball*, to distinguish it from the *closed ball* formed by all $x \in S$ with $d(x, y) \leq \delta$.

In the complex plane $B(z_0, \delta)$ is an *open disk* with center z_0 and radius δ ; it consists of all complex numbers z which satisfy the strict inequality $|z - z_0| < \delta$. We have just proved that it is an open set, and the reader is urged to interpret the proof in geometric terms.

The complement of an open set is said to be *closed*. In any metric space the empty set and the whole space are at the same time open and closed, and there may be other sets with the same property.

The following properties of open and closed sets are fundamental:

The intersection of a finite number of open sets is open.

The union of any collection of open sets is open.

The union of a finite number of closed sets is closed.

The intersection of any collection of closed sets is closed.

The proofs are so obvious that they can be left to the reader. It should be noted that the last two statements follow from the first two by use of the De Morgan laws.

There are many terms in common usage which are directly related to the idea of open sets. A complete list would be more confusing than helpful, and we shall limit ourselves to the following: *interior, closure,*

boundary, exterior.

(i) The interior of a set X is the largest open set contained in X . It exists, for it may be characterized as the union of all open sets $\subset X$. It can also be described as the set of all points of which X is a neighborhood. We denote it by $\text{Int } X$.

(ii) The closure of X is the smallest closed set which contains X , or the intersection of all closed sets $\supset X$. A point belongs to the closure of X if and only if all its neighborhoods intersect X . The closure is usually denoted by X^- , infrequently by $\text{Cl } X$.

(iii) The boundary of X is the closure minus the interior. A point belongs to the boundary if and only if all its neighborhoods intersect both X and $\sim X$. Notation: $\text{Bd } X$ or ∂X .

(iv) The exterior of X is the interior of $\sim X$. It is also the complement of the closure. As such it can be denoted by $\sim X^-$.

Observe that $\text{Int } X \subset X \subset X^-$ and that X is open if $\text{Int } X = X$, closed if $X^- = X$. Also, $X \subset Y$ implies $\text{Int } X \subset \text{Int } Y$, $X^- \subset Y^-$. For added convenience we shall also introduce the notions of *isolated point* and *accumulation point*. We say that $x \in X$ is an isolated point of X if x has a neighborhood whose intersection with X reduces to the point x . An accumulation point is a point of X^- which is not an isolated point. It is clear that x is an accumulation point of X if and only if every neighborhood of x contains infinitely many points from X .

EXERCISES

1. If S is a metric space with distance function $d(x,y)$, show that S with the distance function $\delta(x,y) = d(x,y)/[1 + d(x,y)]$ is also a metric space. The latter space is bounded in the sense that all distances lie under a fixed bound.

2. Suppose that there are given two distance functions $d(x,y)$ and $d_1(x,y)$ on the same space S . They are said to be equivalent if they determine the same open sets. Show that d and d_1 are equivalent if to every $\epsilon > 0$ there exists a $\delta > 0$ such that $d(x,y) < \delta$ implies $d_1(x,y) < \epsilon$, and vice versa. Verify that this condition is fulfilled in the preceding exercise.

3. Show by strict application of the definition that the closure of $|z - z_0| < \delta$ is $|z - z_0| \leq \delta$.

4. If X is the set of complex numbers whose real and imaginary parts are rational, what is $\text{Int } X$, X^- , ∂X ?

5. It is sometimes typographically simpler to write X' for $\sim X$. With this notation, how is X'^{-} related to X ? Show that $X'^{-} = X$.

6. A set is said to be discrete if all its points are isolated. Show that a discrete set in \mathbb{R} or \mathbb{C} is countable.

7. Show that the accumulation points of any set form a closed set.

1.3. Connectedness. If E is any nonempty subset of a metric space S we may consider E as a metric space in its own right with the same distance function $d(x,y)$ as on all of S . Neighborhoods and open sets on E are defined as on any metric space, but an open set on E need not be open when regarded as a subset of S . To avoid confusion neighborhoods and open sets on E are often referred to as relative neighborhoods and relatively open sets. As an example, if we regard the closed interval $0 \leq x \leq 1$ as a subspace of \mathbf{R} , then the semiclosed interval $0 \leq x < 1$ is relatively open, but not open in \mathbf{R} . Henceforth, when we say that a subset E has some specific topological property, we shall always mean that it has this property as a subspace, and its subspace topology is called the relative topology.

Intuitively speaking, a space is *connected* if it consists of a single piece. This is meaningless unless we define the statement in terms of nearness. The easiest way is to give a negative characterization: S is not connected if there exists a partition $S = A \cup B$ into open subsets A and B . It is understood that A and B are *disjoint* and *nonempty*. The connectedness of a space is often used in the following manner: Suppose that we are able to construct two complementary open subsets A and B of S ; if S is connected, we may conclude that either A or B is empty.

A subset $E \subset S$ is said to be connected if it is connected in the relative topology. At the risk of being pedantic we repeat:

Definition 3. A subset of a metric space is connected if it cannot be represented as the union of two disjoint relatively open sets none of which is empty.

If E is open, a subset of E is relatively open if and only if it is open. Similarly, if E is closed, relatively closed means the same as closed. We can therefore state: An open set is connected if it cannot be decomposed into two open sets, and a closed set is connected if it cannot be decomposed into two closed sets. Again, none of the sets is allowed to be empty.

Trivial examples of connected sets are the empty set and any set that consists of a single point.

In the case of the real line it is possible to name all connected sets. The most important result is that the whole line is connected, and this is indeed one of the fundamental properties of the real-number system.

An *interval* is defined by an inequality of one of the four types: $a < x < b$, $a \leq x < b$, $a < x \leq b$, $a \leq x \leq b$.† For $a = -\infty$ or $b = +\infty$ this includes the semi-infinite intervals and the whole line.

† We denote open intervals by (a,b) and closed intervals by $[a,b]$. Another common practice is to denote open intervals by $]a,b[$ and semiclosed intervals by $]a,b]$ or $[a,b[$. It is always understood that $a < b$.

Theorem 1. *The nonempty connected subsets of the real line are the intervals.*

We reproduce one of the classical proofs, based on the fact that any monotone sequence has a finite or infinite limit.

Suppose that the real line \mathbf{R} is represented as the union $\mathbf{R} = A \cup B$ of two disjoint closed sets. If neither is empty we can find $a_1 \in A$ and $b_1 \in B$; we may assume that $a_1 < b_1$. We bisect the interval (a_1, b_1) and note that one of the two halves has its left end point in A and its right end point in B . We denote this interval by (a_2, b_2) and continue the process indefinitely. In this way we obtain a sequence of nested intervals (a_n, b_n) with $a_n \in A$, $b_n \in B$. The sequences $\{a_n\}$ and $\{b_n\}$ have a common limit c . Since A and B are closed c would have to be a common point of A and B . This contradiction shows that either A or B is empty, and hence \mathbf{R} is connected.

With minor modifications the same proof applies to any interval.

Before proving the converse we make an important remark. Let E be an arbitrary subset of \mathbf{R} and call α a *lower bound* of E if $\alpha \leq x$ for all $x \in E$. Consider the set A of all lower bounds. It is evident that the complement of A is open. As to A itself it is easily seen that A is open whenever it does not contain any largest number. Because the line is connected, A and its complement cannot both be open unless one of them is empty. There are thus three possibilities: either A is empty, A contains a largest number, or A is the whole line. The largest number a of A , if it exists, is called the *greatest lower bound* of E ; it is commonly denoted as g.l.b. x or $\inf x$ for $x \in E$. If A is empty, we agree to set $a = -\infty$, and if A is the whole line we set $a = +\infty$. With this convention every set of real numbers has a uniquely determined greatest lower bound; it is clear that $a = +\infty$ if and only if the set E is empty. The *least upper bound*, denoted as l.u.b. x or $\sup x$ for $x \in E$, is defined in a corresponding manner.†

Returning to the proof, we assume that E is a connected set with the greatest lower bound a and the least upper bound b . All points of E lie between a and b , limits included. Suppose that a point ξ from the open interval (a, b) did not belong to E . Then the open sets defined by $x < \xi$ and $x > \xi$ cover E , and because E is connected, one of them must fail to meet E . Suppose, for instance, that no point of E lies to the left of ξ . Then ξ would be a lower bound, in contradiction with the fact that a is the greatest lower bound. The opposite assumption would lead to a similar contradiction, and we conclude that ξ must belong to E . It follows that E is an open, closed, or semiclosed interval with the end points a and b ; the cases $a = -\infty$ and $b = +\infty$ are to be included.

† The supremum of a sequence was introduced already in Chap. 2, Sec. 2.1.

In the course of the proof we have introduced the notions of greatest lower bound and least upper bound. If the set is closed and if the bounds are finite, they must belong to the set, in which case they are called the minimum and the maximum. In order to be sure that the bounds are finite we must know that the set is not empty and that there is some finite lower bound and some finite upper bound. In other words, the set must lie in a finite interval; such a set is said to be *bounded*. We have proved:

Theorem 2. *Any closed and bounded nonempty set of real numbers has a minimum and a maximum.*

The structure of connected sets in the plane is not nearly so simple as in the case of the line, but the following characterization of open connected sets contains essentially all the information we shall need.

Theorem 3. *A nonempty open set in the plane is connected if and only if any two of its points can be joined by a polygon which lies in the set.*

The notion of a joining polygon is so simple that we need not give a formal definition.

We prove first that the condition is necessary. Let A be an open connected set, and choose a point $a \in A$. We denote by A_1 the subset of A whose points can be joined to a by polygons in A , and by A_2 the subset whose points cannot be so joined. Let us prove that A_1 and A_2 are both open. First, if $a_1 \in A_1$ there exists a neighborhood $|z - a_1| < \epsilon$ contained in A . All points in this neighborhood can be joined to a_1 by a line segment, and from there to a by a polygon. Hence the whole neighborhood is contained in A_1 , and A_1 is open. Secondly, if $a_2 \in A_2$, let $|z - a_2| < \epsilon$ be a neighborhood contained in A . If a point in this neighborhood could be joined to a by a polygon, then a_2 could be joined to this point by a line segment, and from there to a . This is contrary to the definition of A_2 , and we conclude that A_2 is open. Since A was connected either A_1 or A_2 must be empty. But A_1 contains the point a ; hence A_2 is empty, and all points can be joined to a . Finally, any two points in A can be joined by way of a , and we have proved that the condition is necessary.

For future use we remark that it is even possible to join any two points by a polygon whose sides are parallel to the coordinate axes. The proof is the same.

In order to prove the sufficiency we assume that A has a representation $A = A_1 \cup A_2$ as the union of two disjoint open sets. Choose $a_1 \in A_1$, $a_2 \in A_2$ and suppose that these points can be joined by a polygon in A .

One of the sides of the polygon must then join a point in A_1 to a point in A_2 , and for this reason it is sufficient to consider the case where a_1 and a_2 are joined by a line segment. This segment has a parametric representation $z = a_1 + t(a_2 - a_1)$ where t runs through the interval $0 \leq t \leq 1$. The subsets of the interval $0 < t < 1$ which correspond to points in A_1 and A_2 , respectively, are evidently open, disjoint, and nonvoid. This contradicts the connectedness of the interval, and we have proved that the condition of the theorem is sufficient.

The theorem generalizes easily to \mathbb{R}^n and \mathbb{C}^n .

Definition 4. *A nonempty connected open set is called a region.*

By Theorem 3 the whole plane, an open disk $|z - a| < \rho$, and a half plane are regions. The same is true of any δ -neighborhood in \mathbb{R}^n . A region is the more-dimensional analogue of an open interval. The closure of a region is called a *closed region*. It should be observed that different regions may have the same closure.

It happens frequently that we have to analyze the structure of sets which are defined very implicitly, for instance in the course of a proof. In such cases the first step is to decompose the set into its maximal connected *components*. As the name indicates, a component of a set is a connected subset which is not contained in any larger connected subset.

Theorem 4. *Every set has a unique decomposition into components.*

If E is the given set, consider a point $a \in E$ and let $C(a)$ denote the union of all connected subsets of E that contain a . Then $C(a)$ is sure to contain a , for the set consisting of the single point a is connected. If we can show that $C(a)$ is connected, then it is a maximal connected set, in other words a component. It would follow, moreover, that any two components are either disjoint or identical, which is precisely what we want to prove. Indeed, if $c \in C(a) \cap C(b)$, then $C(a) \subset C(c)$ by the definition of $C(c)$ and the connectedness of $C(a)$. Hence $a \in C(c)$, and by the same reasoning $C(c) \subset C(a)$, so that in fact $C(a) = C(c)$. Similarly $C(b) = C(c)$, and consequently $C(a) = C(b)$. We call $C(a)$ the component of a .

Suppose that $C(a)$ were not connected. Then we could find relatively open sets $A, B \neq \emptyset$ such that $C(a) = A \cup B$, $A \cap B = \emptyset$. We may assume that $a \in A$ while B contains a point b . Since $b \in C(a)$ there is a connected set $E_0 \subset E$ which contains a and b . The representation $E_0 = (E_0 \cap A) \cup (E_0 \cap B)$ would be a decomposition into relatively open subsets, and since $a \in E_0 \cap A$, $b \in E_0 \cap B$ neither part would be empty. This is a contradiction, and we conclude that $C(a)$ is connected.

Theorem 5. *In \mathbb{R}^n the components of any open set are open.*

This is a consequence of the fact that the δ -neighborhoods in \mathbb{R}^n are connected. Consider $a \in C(a) \subset E$. If E is open it contains $B(a, \delta)$ and because $B(a, \delta)$ is connected $B(a, \delta) \subset C(a)$. Hence $C(a)$ is open. A little more generally the assertion is true for any space S which is *locally connected*. By this we mean that any neighborhood of a point a contains a connected neighborhood of a . The proof is left to the reader.

In the case of \mathbb{R}^n we can conclude, furthermore, that the number of components is countable. To see this we observe that every open set must contain a point with rational coordinates. The set of points with rational coordinates is countable, and may thus be expressed as a sequence $\{p_k\}$. For each component $C(a)$, determine the smallest k such that $p_k \in C(a)$. To different components correspond different k . We conclude that the components are in one-to-one correspondence with a subset of the natural numbers, and consequently the set of components is countable.

For instance, *every open subset of \mathbb{R} is a countable union of disjoint open intervals.*

Again, it is possible to analyze the proof and thereby arrive at a more general result. We shall say that a set E is *dense* in S if $E^- = S$, and we shall say that a metric space is *separable* if there exists a countable subset which is dense in S . We are led to the following result:

In a locally connected separable space every open set is a countable union of disjoint regions.

EXERCISES

1. If $X \subset S$, show that the relatively open (closed) subsets of X are precisely those sets that can be expressed as the intersection of X with an open (closed) subset of S .

2. Show that the union of two regions is a region if and only if they have a common point.

3. Prove that the closure of a connected set is connected.

4. Let A be the set of points $(x, y) \in \mathbb{R}^2$ with $x = 0$, $|y| \leq 1$, and let B be the set with $x > 0$, $y = \sin 1/x$. Is $A \cup B$ connected?

5. Let E be the set of points $(x, y) \in \mathbb{R}^2$ such that $0 \leq x \leq 1$ and either $y = 0$ or $y = 1/n$ for some positive integer n . What are the components of E ? Are they all closed? Are they relatively open? Verify that E is not locally connected.

6. Prove that the components of a closed set are closed (use Ex. 3).

7. A set is said to be *discrete* if all its points are isolated. Show that a discrete set in a separable metric space is countable.

1.4. Compactness. The notions of convergent sequences and Cauchy sequences are obviously meaningful in any metric space. Indeed, we would say that $x_n \rightarrow x$ if $d(x_n, x) \rightarrow 0$, and we would say that $\{x_n\}$ is a Cauchy sequence if $d(x_n, x_m) \rightarrow 0$ as n and m tend to ∞ . It is clear that every convergent sequence is a Cauchy sequence. For \mathbf{R} and \mathbf{C} we have proved the converse, namely that every Cauchy sequence is convergent (Chap. 2, Sec. 2.1), and it is not hard to see that this property carries over to any \mathbf{R}^n . In view of its importance the property deserves a special name.

Definition 5. *A metric space is said to be complete if every Cauchy sequence is convergent.*

A subset is complete if it is complete when regarded as a subspace. The reader will find no difficulty in proving that *a complete subset of a metric space is closed*, and that *a closed subset of a complete space is complete*.

We shall now introduce the stronger concept of *compactness*. It is stronger than completeness in the sense that every compact space or set is complete, but not conversely. As a matter of fact it will turn out that the compact subsets of \mathbf{R} and \mathbf{C} are the closed bounded sets. In view of this result it would be possible to dispense with the notion of compactness, at least for the purposes of this book, but this would be unwise, for it would mean shutting our eyes to the most striking property of bounded and closed sets of real or complex numbers. The outcome would be that we would have to repeat essentially the same proof in many different connections.

There are several equivalent characterizations of compactness, and it is a matter of taste which one to choose as definition. Whatever we do the uninitiated reader will feel somewhat bewildered, for he will not be able to discern the purpose of the definition. This is not surprising, for it took a whole generation of mathematicians to agree on the best approach. The consensus of present opinion is that it is best to focus the attention on the different ways in which a given set can be covered by open sets.

Let us say that a collection of open sets is an *open covering* of a set X if X is contained in the union of the open sets. A *subcovering* is a subcollection with the same property, and a *finite covering* is one that consists of a finite number of sets. The definition of compactness reads:

Definition 6. *A set X is compact if and only if every open covering of X contains a finite subcovering.*

In this context we are thinking of X as a subset of a metric space S ,

and the covering is by open sets of S . But if U is an open set in S , then $U \cap X$ is an open subset of X (a relatively open set), and conversely every open subset of X can be expressed in this form (Sec. 1.3, Ex. 1). For this reason it makes no difference whether we formulate the definition for a full space or for a subset.

The property in the definition is frequently referred to as the *Heine-Borel property*. Its importance lies in the fact that many proofs become particularly simple when formulated in terms of open coverings.

We prove first that every compact space is complete. Suppose that X is compact, and let $\{x_n\}$ be a Cauchy sequence in X . If y is not the limit of $\{x_n\}$ there exists an $\epsilon > 0$ such that $d(x_n, y) > 2\epsilon$ for infinitely many n . Determine n_0 such that $d(x_m, x_n) < \epsilon$ for $m, n \geq n_0$. We choose a fixed $n \geq n_0$ for which $d(x_n, y) > 2\epsilon$. Then $d(x_m, y) \geq d(x_n, y) - d(x_m, x_n) > \epsilon$ for all $m \geq n_0$. It follows that the ϵ -neighborhood $B(y, \epsilon)$ contains only finitely many x_n (better: contains x_n only for finitely many n).

Consider now the collection of all open sets U which contain only finitely many x_n . If $\{x_n\}$ is not convergent, it follows by the preceding reasoning that this collection is an open covering of X . Therefore it must contain a finite subcovering, formed by U_1, \dots, U_N . But that is clearly impossible, for since each U_i contains only finitely many x_n it would follow that the given sequence is finite.

Secondly, a compact set is necessarily *bounded* (a metric space is bounded if all distances lie under a finite bound). To see this, choose a point x_0 and consider all balls $B(x_0, r)$. They form an open covering of X , and if X is compact, it contains a finite subcovering; in other words, $X \subset B(x_0, r_1) \cup \dots \cup B(x_0, r_m)$, which means the same as $X \subset B(x_0, r)$ with $r = \max(r_1, \dots, r_m)$. For any $x, y \in X$ it follows that $d(x, y) \leq d(x, x_0) + d(y, x_0) < 2r$, and we have proved that X is bounded.

But boundedness is not all we can prove. It is convenient to define a stronger property called *total boundedness*:

Definition 7. A set X is *totally bounded* if, for every $\epsilon > 0$, X can be covered by finitely many balls of radius ϵ .

This is certainly true of any compact set. For the collection of all balls of radius ϵ is an open covering, and the compactness implies that we can select finitely many that cover X . We observe that a totally bounded set is necessarily bounded, for if $X \subset B(x_1, \epsilon) \cup \dots \cup B(x_m, \epsilon)$, then any two points of X have a distance $< 2\epsilon + \max d(x_i, x_j)$. (The preceding proof that any compact set is bounded becomes redundant.)

We have already proved one part of the following theorem:

Theorem 6. A set is compact if and only if it is complete and totally bounded.

To prove the other part, assume that the metric space S is complete and totally bounded. Suppose that there exists an open covering which does not contain any finite subcovering. Write $\varepsilon_n = 2^{-n}$. We know that S can be covered by finitely many $B(x, \varepsilon_1)$. If each had a finite subcovering, the same would be true of S ; hence there exists a $B(x_1, \varepsilon_1)$ which does not admit a finite subcovering. Because $B(x_1, \varepsilon_1)$ is itself totally bounded we can find an $x_2 \in B(x_1, \varepsilon_1)$ such that $B(x_2, \varepsilon_2)$ has no finite subcovering.† It is clear how to continue the construction: we obtain a sequence x_n with the property that $B(x_n, \varepsilon_n)$ has no finite subcovering and $x_{n+1} \in B(x_n, \varepsilon_n)$. The second property implies $d(x_n, x_{n+1}) < \varepsilon_n$ and hence $d(x_n, x_{n+p}) < \varepsilon_n + \varepsilon_{n+1} + \cdots + \varepsilon_{n+p-1} < 2^{-n+1}$. It follows that x_n is a Cauchy sequence. It converges to a limit y , and this y belongs to one of the open sets U in the given covering. Because U is open, it contains a ball $B(y, \delta)$. Choose n so large that $d(x_n, y) < \delta/2$ and $\varepsilon_n < \delta/2$. Then $B(x_n, \varepsilon_n) \subset B(y, \delta)$, for $d(x, x_n) < \varepsilon_n$ implies $d(x, y) \leq d(x, x_n) + d(x_n, y) < \delta$. Therefore $B(x_n, \varepsilon_n)$ admits a finite subcovering, namely by the single set U . This is a contradiction, and we conclude that S has the Heine-Borel property.

Corollary. *A subset of \mathbf{R} or \mathbf{C} is compact if and only if it is closed and bounded.*

We have already mentioned this particular consequence. In one direction the conclusion is immediate: We know that a compact set is bounded and complete; but \mathbf{R} and \mathbf{C} are complete, and complete subsets of a complete space are closed. For the opposite conclusion we need to show that every bounded set in \mathbf{R} or \mathbf{C} is totally bounded. Let us take the case of \mathbf{C} . If X is bounded it is contained in a disk, and hence in a square. The square can be subdivided into a finite number of squares with arbitrarily small side, and the squares can in turn be covered by disks with arbitrarily small radius. This proves that X is totally bounded, except for a small point that should not be glossed over. When Definition 7 is applied to a subset $X \subset S$ it is slightly ambiguous, for it is not clear whether the ε -neighborhoods should be with respect to X or with respect to S ; that is, it is not clear whether we require their centers to lie on X . It happens that this is of no avail. In fact, suppose that we have covered X by ε -neighborhoods whose centers do not necessarily lie on X . If such a neighborhood does not meet X it is superfluous, and can be dropped. If it does contain a point from X , then we can replace it by a 2ε -neighborhood around that point, and we obtain a finite covering by 2ε -neighborhoods with centers on X . For this reason the ambiguity is only apparent, and our proof that bounded subsets of \mathbf{C} are totally bounded is valid.

† Here we are using the fact that any subset of a totally bounded set is totally bounded. The reader should prove this.

There is a third characterization of compact sets. It deals with the notion of *limit point* (sometimes called *cluster value*): We say that y is a limit point of the sequence $\{x_n\}$ if there exists a subsequence $\{x_{n_k}\}$ that converges to y . A limit point is almost the same as an accumulation point of the set formed by the points x_n , except that a sequence permits repetitions of the same point. If y is a limit point, every neighborhood of y contains infinitely many x_n . The converse is also true. Indeed, suppose that $\epsilon_k \rightarrow 0$. If every $B(y, \epsilon_k)$ contains infinitely many x_n we can choose subscripts n_k , by induction, in such a way that $x_{n_k} \in B(y, \epsilon_k)$ and $n_{k+1} > n_k$. It is clear that $\{x_{n_k}\}$ converges to y .

Theorem 7. *A metric space is compact if and only if every infinite sequence has a limit point.*

This theorem is usually referred to as the *Bolzano-Weierstrass theorem*. The original formulation was that every bounded sequence of complex numbers has a convergent subsequence. It came to be recognized as an important theorem precisely because of the role it plays in the theory of analytic functions.

The first part of the proof is a repetition of an earlier argument. If y is not a limit point of $\{x_n\}$ it has a neighborhood which contains only finitely many x_n (abbreviated version of the correct phrase). If there were no limit points the open sets containing only finitely many x_n would form an open covering. In the compact case we could select a finite subcovering, and it would follow that the sequence is finite. The previous time we used this reasoning was to prove that a compact space is complete. We showed in essence that every sequence has a limit point, and then we observed that a Cauchy sequence with a limit point is necessarily convergent. For strict economy of thought it would thus have been better to prove Theorem 7 before Theorem 6, but we preferred to emphasize the importance of total boundedness as early as possible.

It remains to prove the converse. In the first place it is clear that the Bolzano-Weierstrass property implies completeness. Indeed, we just pointed out that a Cauchy sequence with a limit point must be convergent. Suppose now that the space is not totally bounded. Then there exists an $\epsilon > 0$ such that the space cannot be covered by finitely many ϵ -neighborhoods. We construct a sequence $\{x_n\}$ as follows: x_1 is arbitrary, and when x_1, \dots, x_n have been selected we choose x_{n+1} so that it does not lie in $B(x_1, \epsilon) \cup \dots \cup B(x_n, \epsilon)$. This is always possible because these neighborhoods do not cover the whole space. But it is clear that $\{x_n\}$ has no convergent subsequence, for $d(x_m, x_n) > \epsilon$ for all m and n . We conclude that the Bolzano-Weierstrass property implies total boundedness. In view of Theorem 6 that is what we had to prove.

The reader should reflect on the fact that we have exhibited three characterizations of compactness whose logical equivalence is not at all trivial. It should be clear that results of this kind are particularly valuable for the purpose of presenting proofs as concisely as possible.

EXERCISES

1. Give an alternate proof of the fact that every bounded sequence of complex numbers has a convergent subsequence (for instance by use of the limes inferior).

2. Show that the Heine-Borel property can also be expressed in the following manner: Every collection of closed sets with an empty intersection contains a finite subcollection with empty intersection.

3. Use compactness to prove that a closed bounded set of real numbers has a maximum.

4. If $E_1 \supset E_2 \supset E_3 \supset \dots$ is a decreasing sequence of nonempty compact sets, then the intersection $\bigcap_1^{\infty} E_n$ is not empty (Cantor's lemma). Show by example that this need not be true if the sets are merely closed.

5. Let S be the set of all sequences $x = \{x_n\}$ of real numbers such that only a finite number of the x_n are $\neq 0$. Define $d(x, y) = \max |x_n - y_n|$. Is the space complete? Show that the δ -neighborhoods are not totally bounded.

1.5. Continuous Functions. We shall consider functions f which are defined on a metric space S and have values in another metric space S' . Functions are also referred to as *mappings*: we say that f maps S into S' , and we write $f: S \rightarrow S'$. Naturally, we shall be mainly concerned with real- or complex-valued functions; occasionally the latter are allowed to take values in the extended complex plane, ordinary distance being replaced by distance on the Riemann sphere.

The space S is the *domain* of the function. We are of course free to consider functions f whose domain is only a subset of S , in which case the domain is regarded as a subspace. In most cases it is safe to slur over the distinction: a function on S and its restriction to a subset are usually denoted by the same symbol. If $X \subset S$ the set of all values $f(x)$ for $x \in X$ is called the *image* of X under f , and it is denoted by $f(X)$. The *inverse image* $f^{-1}(X')$ of $X' \subset S'$ consists of all $x \in S$ such that $f(x) \in X'$. Observe that $f(f^{-1}(X')) \subset X'$, and $f^{-1}(f(X)) \supset X$.

The definition of a continuous function needs practically no modification: f is continuous at a if to every $\epsilon > 0$ there exists $\delta > 0$ such that $d(x, a) < \delta$ implies $d(f(x), f(a)) < \epsilon$. We are mainly concerned with functions that are continuous at all points in the domain of definition.

The following characterizations are immediate consequences of the definition:

A function is continuous if and only if the inverse image of every open set is open.

A function is continuous if and only if the inverse image of every closed set is closed.

If f is not defined on all of S , the words "open" and "closed," when referring to the inverse image, should of course be interpreted relatively to the domain of f . It is very important to observe that these properties hold only for the inverse image, not for the direct image. For instance the mapping $f(x) = x^2/(1 + x^2)$ of \mathbf{R} into \mathbf{R} has the image $f(\mathbf{R}) = \{y; 0 \leq y < 1\}$ which is neither open nor closed. In this example $f(\mathbf{R})$ fails to be closed because \mathbf{R} is not compact. In fact, the following is true:

Theorem 8. *Under a continuous mapping the image of every compact set is compact, and consequently closed.*

Suppose that f is defined and continuous on the compact set X . Consider a covering of $f(X)$ by open sets U . The inverse images $f^{-1}(U)$ are open and form a covering of X . Because X is compact we can select a finite subcovering: $X \subset f^{-1}(U_1) \cup \dots \cup f^{-1}(U_m)$. It follows that $f(X) \subset U_1 \cup \dots \cup U_m$, and we have proved that $f(X)$ is compact.

Corollary. *A continuous real-valued function on a compact set has a maximum and a minimum.*

The image is a closed bounded subset of \mathbf{R} . The existence of a maximum and a minimum follows by Theorem 2.

Theorem 9. *Under a continuous mapping the image of any connected set is connected.*

We may assume that f is defined and continuous on the whole space S , and that $f(S)$ is all of S' . Suppose that $S' = A \cup B$ where A and B are open and disjoint. Then $S = f^{-1}(A) \cup f^{-1}(B)$ is a representation of S as a union of disjoint open sets. If S is connected either $f^{-1}(A) = \emptyset$ or $f^{-1}(B) = \emptyset$, and hence $A = \emptyset$ or $B = \emptyset$. We conclude that S' is connected.

A typical application is the assertion that a real-valued function which is continuous and never zero on a connected set is either always positive or always negative. In fact, the image is connected, and hence an interval. But an interval which contains positive and negative num-

bers also contains zero.

A mapping $f : S \rightarrow S'$ is said to be *one to one* if $f(x) = f(y)$ only for $x = y$; it is said to be *onto* if $f(S) = S'$.† A mapping with both these properties has an inverse f^{-1} , defined on S' ; it satisfies $f^{-1}(f(x)) = x$ and $f(f^{-1}(x')) = x'$. In this situation, if f and f^{-1} are both continuous we say that f is a *topological mapping* or a *homeomorphism*. A property of a set which is shared by all topological images is called a *topological property*. For instance, we have proved that compactness and connectedness are topological properties (Theorems 8 and 9). In this connection it is perhaps useful to point out that the property of being an open subset is not topological. If $X \subset S$ and $Y \subset S'$ and if X is homeomorphic to Y there is no reason why X and Y should be simultaneously open. It happens to be true if $S = S' = \mathbf{R}^n$ (*invariance of the region*), but this is a deep theorem that we shall not need.

The notion of *uniform continuity* will be in constant use. Quite generally, a condition is said to hold uniformly with respect to a parameter if it can be expressed by inequalities which do not involve the parameter. Accordingly, a function f is said to be *uniformly continuous* on X if, to every $\epsilon > 0$, there exists a $\delta > 0$ such that $d'(f(x_1), f(x_2)) < \epsilon$ for all pairs (x_1, x_2) with $d(x_1, x_2) < \delta$. The emphasis is on the fact that δ is not allowed to depend on x_1 .

Theorem 10. *On a compact set every continuous function is uniformly continuous.*

The proof is typical of the way the Heine-Borel property can be used. Suppose that f is continuous on a compact set X . For every $y \in X$ there is a ball $B(y, \rho)$ such that $d'(f(x), f(y)) < \epsilon/2$ for $x \in B(y, \rho)$; here ρ may depend on y . Consider the covering of X by the smaller balls $B(y, \rho/2)$. There exists a finite subcovering: $X \subset B(y_1, \rho_1/2) \cup \dots \cup B(y_m, \rho_m/2)$. Let δ be the smallest of the numbers $\rho_1/2, \dots, \rho_m/2$, and suppose that $d(x_1, x_2) < \delta$. There is a y_k with $d(x_1, y_k) < \rho_k/2$, and we obtain $d(x_2, y_k) < \rho_k/2 + \delta \leq \delta_k$. Hence $d'(f(x_1), f(y_k)) < \epsilon/2$ and $d'(f(x_2), f(y_k)) < \epsilon/2$ so that $d'(f(x_1), f(x_2)) < \epsilon$ as desired.

On sets which are not compact some continuous functions are uniformly continuous and others are not. For instance, the function z is uniformly continuous on the whole complex plane, but the function z^2 is not.

† These linguistically clumsy terms can be replaced by *injective* (for one to one) and *surjective* (for onto). A mapping with both properties is called *bijective*.

EXERCISES

1. Construct a topological mapping of the open disk $|z| < 1$ onto the whole plane.

2. Prove that a subset of the real line which is topologically equivalent to an open interval is an open interval. (Consider the effect of removing a point.)

3. Prove that every continuous one-to-one mapping of a compact space is topological. (Show that closed sets are mapped on closed sets.)

4. Let X and Y be compact sets in a complete metric space. Prove that there exist $x \in X, y \in Y$ such that $d(x, y)$ is a minimum.

5. Which of the following functions are uniformly continuous on the whole real line: $\sin x, x \sin x, x \sin(x^2), |x|^{\frac{1}{2}} \sin x$?

1.6. Topological Spaces. It is not necessary, and not always convenient, to express nearness in terms of distance. The observant reader will have noticed that most results in the preceding sections were formulated in terms of open sets. True enough, we used distances to define open sets, but there is really no strong reason to do this. If we decide to consider the open sets as the primary objects we must postulate axioms that they have to satisfy. The following axioms lead to the commonly accepted definition of a *topological space*:

Definition 8. A topological space is a set T together with a collection of its subsets, called open sets. The following conditions have to be fulfilled:

- (i) The empty set \emptyset and the whole space T are open sets.
- (ii) The intersection of any two open sets is an open set.
- (iii) The union of an arbitrary collection of open sets is an open set.

We recognize at once that this terminology is consistent with our earlier definition of an open subset of a metric space. Indeed, properties (ii) and (iii) were strongly emphasized, and (i) is trivial.

Closed sets are the complements of open sets, and it is immediately clear how to define interior, closure, boundary, and so on. Neighborhoods could be avoided, but they are rather convenient: N is a neighborhood of x if there exists an open set U such that $x \in U$ and $U \subset N$.

Connectedness was defined purely by means of open sets. Hence the definition carries over to topological spaces, and the theorems remain true. The Heine-Borel property is also one that deals only with open sets. Therefore it makes perfect sense to speak of a compact topological space. However, Theorem 6 becomes meaningless, and Theorem 7 becomes false.

As a matter of fact, the first serious difficulty we encounter is with