COMPLEX ANALYSIS

An Introduction to the Theory of Analytic Functions of One Complex Variable

Third Edition

Lars V. Ahlfors
Professor of Mathematics, Emeritus
Harvard University

McGraw-Hill, Inc.
New York St. Louis San Francisco Auckland Bogotá
Caracas Lisbon London Madrid Mexico City Milan
Montreal New Delhi San Juan Singapore
Sydney Tokyo Toronto

COMPLEX ANALYSIS

22 23 BRBBRB 9 8 7 6 5 4 3

This book was set in Modern 8A by Monotype Composition Company, Inc.
The editors were Carol Napier and Stephen Wagley;
the production supervisor was Joe Campanella.

Library of Congress Cataloging in Publication Data

Ahlfors, Lars Valerian, date
Complex analysis.

(International series in pure and applied mathematics)
Includes index.
1. Analytic functions. I. Title.
QA331.A45 1979 515'.93 78-17078
ISBN 0-07-000657-1

To Erna

Contents

3 ANALYTIC FUNCTIONS AS MAPPINGS

A function $w = f(z)$ may be viewed as a mapping which represents a point z by its image w. The purpose of this chapter is to study, in a preliminary way, the special properties of mappings defined by analytic functions.

In order to carry out this program it is desirable to develop the underlying concepts with sufficient generality, for otherwise we would soon be forced to introduce a great number of ad hoc definitions whose mutual relationship would be far from clear. Since present-day students are exposed to abstraction and generality at quite an early stage, no apologies are needed. It is perhaps more appropriate to sound a warning that greatest possible generality should not become a purpose.

In the first section we develop the fundamentals of point set topology and metric spaces. There is no need to go very far, for our main concern is with the properties that are essential for the study of analytic functions. If the student feels that he is already thoroughly familiar with this material, he should read it only for terminology.

The author believes that proficiency in the study of analytic functions requires a mixture of geometric feeling and computational skill. The second and third sections, only loosely connected with the first, are expressly designed to develop geometric feeling by way of detailed study of elementary mappings. At the same time we try to stress rigor in geometric thinking, to the point where the geometric image becomes the guide but not the foundation of reasoning.

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 $\varnothing$.

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*

$$X \cup (Y \cap Z) = (X \cup Y) \cap (X \cup Z)$$
$$X \cap (Y \cup Z) = (X \cap Y) \cup (X \cap Z)$$

and the *De Morgan laws*

$$\sim(X \cup Y) = \sim X \cap \sim Y$$
$$\sim(X \cap Y) = \sim X \cup \sim Y.$$

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 **R** and **C** of real and complex numbers, respectively, such nearness can be expressed by a quantitative condition $|x - y| < \varepsilon$. For instance, to say that a set X *contains all* x *sufficiently near to* y means that there exists an $\varepsilon > 0$ such that $x \in X$ whenever $|x - y| < \varepsilon$. Similarly, X *contains points arbitrarily near to* y if to every $\varepsilon > 0$ there exists an $x \in X$ such that $|x - y| < \varepsilon$.

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. $d(x,z) \leqq d(x,y) + d(y,z)$.

The last condition is the *triangle inequality*.

For instance, **R** and **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

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

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 $a \leqq x \leqq b$. 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 $d(x,y) \leqq \delta$.

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 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 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: 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 Int $X \subset X \subset X^-$ and that X is open if Int $X = X$, closed if $X^- = X$. Also, $X \subset Y$ implies Int $X \subset$ 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 $\varepsilon > 0$ there exists a $\delta > 0$ such that $d(x,y) < \delta$ implies $d_1(x,y) < \varepsilon$, 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 $|z - z_0| \leqq \delta$.

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

6. A set is said to be discrete if all its points are isolated. Show that a discrete set in $\mathbf{R}$ or $\mathbf{C}$ 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 $0 \leqq x \leqq 1$ as a subspace of $\mathbf{R}$, then the semiclosed interval $0 \leqq x < 1$ 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 meaningness 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: $a < x < b$, $a \leqq x < b$, $a < x \leqq b$, $a \leqq x \leqq b$.† 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 $\alpha \leqq x$ 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 $|z - a_1| < \varepsilon$ 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 $|z - a_2| < \varepsilon$ 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 $0 \leqq t \leqq 1$. 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 $\mathbf{R}^n$ and $\mathbf{C}^n$.

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 $\mathbf{R}^n$. 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* $\mathbf{R}^n$ *the components of any open set are open.*

This is a consequence of the fact that the δ-neighborhoods in $\mathbf{R}^n$ 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 $\mathbf{R}^n$ 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* $\mathbf{R}$ *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 $(x,y) \in \mathbf{R}^2$ with $x = 0$, $|y| \leqq 1$, 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 $(x,y) \in \mathbf{R}^2$ such that $0 \leqq x \leqq 1$ 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 $x_n \to x$ if $d(x_n,x) \to 0$, and we would say that $\{x_n\}$ is a Cauchy sequence if $d(x_n,x_m) \to 0$ 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 $\varepsilon > 0$ such that $d(x_n,y) > 2\varepsilon$ for infinitely many n. Determine n_0 such that $d(x_m,x_n) < \varepsilon$ for $m, n \geqq n_0$. We choose a fixed $n \geqq n_0$ for which $d(x_n,y) > 2\varepsilon$. Then $d(x_m,y) \geqq d(x_n,y) - d(x_m,x_n) > \varepsilon$ for all $m \geqq n_0$. It follows that the ε-neighborhood $B(y,\varepsilon)$ 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 $U_1, \ldots, U_N$. 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, $X \subset B(x_0,r_1) \cup \cdots \cup B(x_0,r_m)$, which means the same as $X \subset B(x_0,r)$ with $r = \max(r_1, \ldots, r_m)$. For any $x,y \in X$ it follows that $d(x,y) \leqq d(x,x_0) + d(y,x_0) < 2r$, 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 $\varepsilon > 0$, 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 $X \subset B(x_1,\varepsilon) \cup \cdots \cup B(x_m,\varepsilon)$, then any two points of X have a distance $< 2\varepsilon + \max d(x_i,x_j)$. (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 $d(x, y) \leqq d(x, x_n) + d(x_n, y) < \delta$. 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 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 $\varepsilon_k \to 0$. If every $B(y, \varepsilon_k)$ contains infinitely many x_n we can choose subscripts n_k, by induction, in such a way that $x_{n_k} \in B(y, \varepsilon_k)$ 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 $\varepsilon > 0$ 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 $x_1, \ldots, x_n$ have been selected we choose x_{n+1} so that it does not lie in $B(x_1, \varepsilon) \cup \cdots \cup B(x_n, \varepsilon)$. 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 $d(x_m, x_n) > \varepsilon$ 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 $E_1 \supset E_2 \supset E_3 \supset \cdots$ 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 $f : S \to S'$. 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 $x \in S$ 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 $\varepsilon > 0$ there exists $\delta > 0$ such that $d(x,a) < \delta$ implies $d'(f(x),f(a)) < \varepsilon$. 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 $f(\mathbf{R}) = \{y;\ 0 \leqq y < 1\}$ 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: $X \subset f^{-1}(U_1) \cup \cdots \cup f^{-1}(U_m)$. It follows that $f(X) \subset U_1 \cup \cdots \cup U_m$, 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 $f^{-1}(A) = 0$ or $f^{-1}(B) = 0$, and hence $A = 0$ or $B = 0$. 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 $f: S \to S'$ 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 $\varepsilon > 0$, there exists a $\delta > 0$ such that $d'(f(x_1),f(x_2)) < \varepsilon$ 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: $X \subset B(y_1,\rho_1/2) \cup \cdots \cup B(y_m,\rho_m/2)$. Let δ be the smallest of the numbers $\rho_1/2, \ldots, \rho_m/2$, 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 $d(x_2,y_k) < \rho_k/2 + \delta \leqq \delta_k$. 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 $\varnothing$ 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

convergent sequences. The definition is clear: we say that $x_n \to x$ if every neighborhood of x contains all but a finite number of the x_n. But if $x_n \to x$ and $x_n \to y$ we are not able to prove that $x = y$. This awkward situation is remedied by introducing a new axiom which characterizes the topological space as a *Hausdorff space:*

Definition 9. *A topological space is called a Hausdorff space if any two distinct points are contained in disjoint open sets.*

In other words, if $x \neq y$ we require the existence of open sets U, V such that $x \in U$, $y \in V$ and $U \cap V = \varnothing$. In the presence of this condition it is obvious that the limit of a convergent sequence is unique. We shall never in this book have occasion to consider a space that is not a Hausdorff space.

This is not the place to give examples of topologies that cannot be derived from a distance function. Such examples would necessarily be very complicated and would not further the purposes of this book. The point is that it may be unnatural to introduce a distance in situations when one is not really needed. The reason for including this section has been to alert the reader that distances are dispensable.

2. CONFORMALITY

We now return to our original setting where all functions and variables are restricted to real or complex numbers. The role of metric spaces will seem disproportionately small: all we actually need are some simple applications of connectedness and compactness.

The whole section is mainly descriptive. It centers on the geometric consequences of the existence of a derivative.

2.1. Arcs and Closed Curves. The equation of an *arc* γ in the plane is most conveniently given in parametric form $x = x(t)$, $y = y(t)$ where t runs through an interval $\alpha \leqq t \leqq \beta$ and $x(t)$, $y(t)$ are continuous functions. We can also use the complex notation $z = z(t) = x(t) + iy(t)$ which has several advantages. It is also customary to identify the arc γ with the continuous mapping of $[\alpha,\beta]$. When following this custom it is preferable to denote the mapping by $z = \gamma(t)$.

Considered as a point set an arc is the image of a closed finite interval under a continuous mapping. As such it is compact and connected. However, an arc is not merely a set of points, but very essentially also a succession of points, ordered by increasing values of the parameter. If a nondecreasing function $t = \varphi(\tau)$ maps an interval $\alpha' \leqq \tau \leqq \beta'$ onto $\alpha \leqq t \leqq \beta$, then $z = z(\varphi(\tau))$ defines the same succession of points as $z = z(t)$.

We say that the first equation arises from the second by a *change of parameter*. The change is *reversible* if and only if $\varphi(\tau)$ is strictly increasing. For instance, the equation $z = t^2 + it^4$, $0 \leqq t \leqq 1$ arises by a reversible change of parameter from the equation $z = t + it^2$, $0 \leqq t \leqq 1$. A change of the parametric interval (α,β) can always be brought about by a *linear* change of parameter, which is one of the form $t = a\tau + b$, $a > 0$.

Logically, the simplest course is to consider two arcs as different as soon as they are given by different equations, regardless of whether one equation may arise from the other by a change of parameter. In following this course, as we will, it is important to show that certain properties of arcs are invariant under a change of parameter. For instance, the *initial* and *terminal point* of an arc remain the same after a change of parameter.

If the derivative $z'(t) = x'(t) + iy'(t)$ exists and is $\neq 0$, the arc γ has a *tangent* whose direction is determined by $\arg z'(t)$. We shall say that the arc is *differentiable* if $z'(t)$ exists and is continuous (the term continuously differentiable is too unwieldy); if, in addition, $z'(t) \neq 0$ the arc is said to be *regular*. An arc is *piecewise differentiable* or *piecewise regular* if the same conditions hold except for a finite number of values t; at these points $z(t)$ shall still be continuous with left and right derivatives which are equal to the left and right limits of $z'(t)$ and, in the case of a piecewise regular arc, $\neq 0$.

The differentiable or regular character of an arc is invariant under the change of parameter $t = \varphi(\tau)$ provided that $\varphi'(\tau)$ is continuous and, for regularity, $\neq 0$. When this is the case, we speak of a differentiable or regular change of parameter.

An arc is *simple*, or a *Jordan arc*, if $z(t_1) = z(t_2)$ only for $t_1 = t_2$. An arc is a *closed curve* if the end points coincide: $z(\alpha) = z(\beta)$. For closed curves a *shift* of the parameter is defined as follows: If the original equation is $z = z(t)$, $\alpha \leqq t \leqq \beta$, we choose a point t_0 from the interval (α,β) and define a new closed curve whose equation is $z = z(t)$ for $t_0 \leqq t \leqq \beta$ and $z = z(t - \beta + \alpha)$ for $\beta \leqq t \leqq t_0 + \beta - \alpha$. The purpose of the shift is to get rid of the distinguished position of the initial point. The correct definitions of a differentiable or regular closed curve and of a *simple closed curve* (or *Jordan curve*) are obvious.

The *opposite arc* of $z = z(t)$, $\alpha \leqq t \leqq \beta$, is the arc $z = z(-t)$, $-\beta \leqq t \leqq -\alpha$. Opposite arcs are sometimes denoted by γ and $-\gamma$, sometimes by γ and γ^{-1}, depending on the connection. A constant function $z(t)$ defines a *point curve*.

A circle C, originally defined as a locus $|z - a| = r$, can be considered as a closed curve with the equation $z = a + re^{it}$, $0 \leqq t \leqq 2\pi$. We will use this standard parametrization whenever a circle is introduced. This convention saves us from writing down the equation each time it is

needed; also, and this is its most important purpose, it serves as a definite rule to distinguish between C and $-C$.

2.2. Analytic Functions in Regions. When we consider the derivative

$$f'(z) = \lim_{h \to 0} \frac{f(z + h) - f(z)}{h}$$

of a complex-valued function, defined on a set A in the complex plane, it is of course understood that $z \in A$ and that the limit is with respect to values h such that $z + h \in A$. The existence of the derivative will therefore have a different meaning depending on whether z is an interior point or a boundary point of A. The way to avoid this is to insist that all analytic functions be defined on open sets.

We give a formal statement of the definition:

Definition 10. *A complex-valued function $f(z)$, defined on an open set Ω, is said to be analytic in Ω if it has a derivative at each point of Ω.*

Sometimes one says more explicitly that $f(z)$ is *complex analytic.* A commonly used synonym is *holomorphic.*

It is important to stress that the open set Ω is part of the definition. As a rule one should avoid speaking of an analytic function $f(z)$ without referring to a specific open set Ω on which it is defined, but the rule can be broken if it is clear from the context what the set is. Observe that f must first of all be a *function,* and hence *single-valued.* If Ω' is an open subset of Ω, and if $f(z)$ is analytic in Ω, then the restriction of f to Ω' is analytic in Ω'; it is customary to denote the restriction by the same letter f. In particular, since the components of an open set are open, it is no loss of generality to consider only the case where Ω is connected, that is to say a *region.*

For greater flexibility of the language it is desirable to introduce the following complement to Definition 10:

Definition 11. *A function $f(z)$ is analytic on an arbitrary point set A if it is the restriction to A of a function which is analytic in some open set containing A.*

The last definition is merely an agreement to use a convenient terminology. This is a case in which the set Ω need not be explicitly mentioned, for the specific choice of Ω is usually immaterial as long as it contains A. Another instance in which the mention of Ω can be suppressed is the phrase: "Let $f(z)$ be analytic at z_0." It means that a function $f(z)$ is defined and has a derivative in some unspecified open neighborhood of z_0.

Although our definition requires all analytic functions to be single-valued, it is possible to consider such multiple-valued functions as $\sqrt{z}$, $\log z$, or $\arccos z$, provided that they are restricted to a definite region in which it is possible to select a single-valued and analytic branch of the function.

For instance, we may choose for Ω the complement of the negative real axis $z \leqq 0$; this set is indeed open and connected. In Ω one and only one of the values of $\sqrt{z}$ has a positive real part. With this choice $w = \sqrt{z}$ becomes a single-valued function in Ω; let us prove that it is continuous. Choose two points $z_1, z_2 \in \Omega$ and denote the corresponding values of w by $w_1 = u_1 + iv_1$, $w_2 = u_2 + iv_2$ with $u_1, u_2 > 0$. Then

$$|z_1 - z_2| = |w_1^2 - w_2^2| = |w_1 - w_2| \cdot |w_1 + w_2|$$

and $|w_1 + w_2| \geqq u_1 + u_2 > u_1$. Hence

$$|w_1 - w_2| < \frac{|z_1 - z_2|}{u_1}$$

and it follows that $w = \sqrt{z}$ is continuous at z_1. Once the continuity is established the analyticity follows by derivation of the inverse function $z = w^2$. Indeed, with the notations used in calculus $\Delta z \to 0$ implies $\Delta w \to 0$. Therefore,

$$\lim_{\Delta z \to 0} \frac{\Delta w}{\Delta z} = \lim_{\Delta w \to 0} \frac{\Delta w}{\Delta z}$$

and we obtain

$$\frac{dw}{dz} = \frac{1}{\dfrac{dz}{dw}} = \frac{1}{2w} = \frac{1}{2\sqrt{z}}$$

with the same branch of $\sqrt{z}$.

In the case of $\log z$ we can use the same region Ω, obtained by excluding the negative real axis, and define the *principal branch* of the logarithm by the condition $|\operatorname{Im} \log z| < \pi$. Again, the continuity must be proved, but this time we have no algebraic identity at our disposal, and we are forced to use a more general reasoning. Denote the principal branch by $w = u + iv = \log z$. For a given point $w_1 = u_1 + iv_1$, $|v_1| < \pi$, and a given $\varepsilon > 0$, consider the set A in the w-plane which is defined by the inequalities $|w - w_1| \geqq \varepsilon$, $|v| \leqq \pi$, $|u - u_1| \leqq \log 2$. This set is closed and bounded, and for sufficiently small ε it is not empty. The continuous function $|e^w - e^{w_1}|$ has consequently a minimum ρ on A (Theorem 8, Corollary). This minimum is positive, for A does not contain any point $w_1 + n \cdot 2\pi i$. Choose $\delta = \min(\rho, \frac{1}{2}e^{u_1})$, and assume that

$$|z_1 - z_2| = |e^{w_1} - e^{w_2}| < \delta.$$

Then w_2 cannot lie in A, for this would make $|e^{w_1} - e^{w_2}| \geqq \rho \geqq \delta$. Neither is it possible that $u_2 < u_1 - \log 2$ or $u_2 > u_1 + \log 2$; in the former case we would obtain $|e^{w_1} - e^{w_2}| \geqq e^{u_1} - e^{u_2} > \frac{1}{2}e^{u_1} \geqq \delta$, and in the latter case $|e^{w_1} - e^{w_2}| \geqq e^{u_2} - e^{u_1} > e^{u_1} > \delta$. Hence w_2 must lie in the disk $|w - w_1| < \varepsilon$, and we have proved that w is a continuous function of z. From the continuity we conclude as above that the derivative exists and equals $1/z$.

The infinitely many values of arc cos z are the same as the values of $i \log (z + \sqrt{z^2 - 1})$. In this case we restrict z to the complement Ω' of the half lines $x \leqq -1$, $y = 0$ and $x \geqq 1$, $y = 0$. Since $1 - z^2$ is never real and $\leqq 0$ in Ω', we can define $\sqrt{1 - z^2}$ as in the first example and then set $\sqrt{z^2 - 1} = i\sqrt{1 - z^2}$. Moreover, $z + \sqrt{z^2 - 1}$ is never real in Ω', for $z + \sqrt{z^2 - 1}$ and $z - \sqrt{z^2 - 1}$ are reciprocals and hence real only if z and $\sqrt{z^2 - 1}$ are both real; this happens only when z lies on the excluded parts of the real axis. Because Ω' is connected, it follows that all values of $z + \sqrt{z^2 - 1}$ in Ω' are on the same side of the real axis, and since i is such a value they are all in the upper half plane. We can therefore define an analytic branch of $\log (z + \sqrt{z^2 - 1})$ whose imaginary part lies between 0 and π. In this way we obtain a single-valued analytic function

$$\text{arc cos } z = i \log (z + \sqrt{z^2 - 1})$$

in Ω' whose derivative is

$$D \text{ arc cos } z = i \frac{1}{z + \sqrt{z^2 - 1}} \left(1 + \frac{z}{\sqrt{z^2 - 1}}\right) = \frac{1}{\sqrt{1 - z^2}}$$

where $\sqrt{1 - z^2}$ has a positive real part.

There is nothing unique about the way in which the region and the single-valued branches have been chosen in these examples. Therefore, each time we consider a function such as $\log z$ the choice of the branch has to be specified. It is a fundamental fact that it is *impossible* to define a single-valued and analytic branch of $\log z$ in certain regions. This will be proved in the chapter on integration.

All the results of Chap. II, Sec. 1.2 remain valid for functions which are analytic on an open set. In particular, the real and imaginary parts of an analytic function in Ω satisfy the Cauchy-Riemann equations

$$\frac{\partial u}{\partial x} = \frac{\partial v}{\partial y}, \qquad \frac{\partial u}{\partial y} = -\frac{\partial v}{\partial x}.$$

Conversely, if u and v satisfy these equations in Ω, and if the partial derivatives are continuous, then $u + iv$ is an analytic function in Ω.

An analytic function in Ω *degenerates* if it reduces to a constant. In

the following theorem we shall list some simple conditions which have this consequence:

Theorem 11. *An analytic function in a region Ω whose derivative vanishes identically must reduce to a constant. The same is true if either the real part, the imaginary part, the modulus, or the argument is constant.*

The vanishing of the derivative implies that $\partial u/\partial x$, $\partial u/\partial y$, $\partial v/\partial x$, $\partial v/\partial y$ are all zero. It follows that u and v are constant on any line segment in Ω which is parallel to one of the coordinate axes. In Sec. 1.3 we remarked, in connection with Theorem 3, that any two points in a region can be joined within the region by a polygon whose sides are parallel to the axes. We conclude that $u + iv$ is constant.

If u or v is constant,

$$f'(z) = \frac{\partial u}{\partial x} - i\frac{\partial u}{\partial y} = \frac{\partial v}{\partial y} + i\frac{\partial v}{\partial x} = 0,$$

and hence $f(z)$ must be constant. If $u^2 + v^2$ is constant, we obtain

$$u\frac{\partial u}{\partial x} + v\frac{\partial v}{\partial x} = 0$$

and

$$u\frac{\partial u}{\partial y} + v\frac{\partial v}{\partial y} = -u\frac{\partial v}{\partial x} + v\frac{\partial u}{\partial x} = 0.$$

These equations permit the conclusion $\partial u/\partial x = \partial v/\partial x = 0$ unless the determinant $u^2 + v^2$ vanishes. But if $u^2 + v^2 = 0$ at a single point it is constantly zero and $f(z)$ vanishes identically. Hence $f(z)$ is in any case a constant.

Finally, if $\arg f(z)$ is constant, we can set $u = kv$ with constant k (unless v is identically zero). But $u - kv$ is the real part of $(1 + ik)f$, and we conclude again that f must reduce to a constant.

Note that for this theorem it is essential that Ω is a region. If not, we can only assert that $f(z)$ is constant on each component of Ω.

EXERCISES

1. Give a precise definition of a single-valued branch of $\sqrt{1+z} + \sqrt{1-z}$ in a suitable region, and prove that it is analytic.

2. Same problem for $\log \log z$.

3. Suppose that $f(z)$ is analytic and satisfies the condition $|f(z)^2 - 1| < 1$ in a region Ω. Show that either $\operatorname{Re} f(z) > 0$ or $\operatorname{Re} f(z) < 0$ throughout Ω.

2.3. Conformal Mapping. Suppose that an arc γ with the equation $z = z(t)$, $\alpha \leqq t \leqq \beta$, is contained in a region Ω, and let $f(z)$ be defined and continuous in Ω. Then the equation $w = w(t) = f(z(t))$ defines an arc γ' in the w-plane which may be called the *image* of γ.

Consider the case of an $f(z)$ which is analytic in Ω. If $z'(t)$ exists, we find that $w'(t)$ also exists and is determined by

$$w'(t) = f'(z(t))z'(t). \tag{1}$$

We will investigate the meaning of this equation at a point $z_0 = z(t_0)$ with $z'(t_0) \neq 0$ and $f'(z_0) \neq 0$.

The first conclusion is that $w'(t_0) \neq 0$. Hence γ' has a tangent at $w_0 = f(z_0)$, and its direction is determined by

$$\arg w'(t_0) = \arg f'(z_0) + \arg z'(t_0). \tag{2}$$

This relation asserts that the angle between the directed tangents to γ at z_0 and to γ' at w_0 is equal to $\arg f'(z_0)$. It is hence independent of the curve γ. For this reason curves through z_0 which are tangent to each other are mapped onto curves with a common tangent at w_0. Moreover, two curves which form an angle at z_0 are mapped upon curves forming the same angle, in sense as well as in size. In view of this property the mapping by $w = f(z)$ is said to be *conformal* at all points with $f'(z) \neq 0$.

A related property of the mapping is derived by consideration of the modulus $|f'(z_0)|$. We have

$$\lim_{z \to z_0} \frac{|f(z) - f(z_0)|}{|z - z_0|} = |f'(z_0)|,$$

and this means that any small line segment with one end point at z_0 is, in the limit, contracted or expanded in the ratio $|f'(z_0)|$. In other words, the linear change of scale at z_0, effected by the transformation $w = f(z)$, is independent of the direction. In general this change of scale will vary from point to point.

Conversely, it is clear that both kinds of conformality together imply the existence of $f'(z_0)$. It is less obvious that each kind will separately imply the same result, at least under additional regularity assumptions.

To be more precise, let us assume that the partial derivatives $\partial f/\partial x$ and $\partial f/\partial y$ are continuous. Under this condition the derivative of $w(t) = f(z(t))$ can be expressed in the form

$$w'(t_0) = \frac{\partial f}{\partial x} x'(t_0) + \frac{\partial f}{\partial y} y'(t_0)$$

where the partial derivatives are taken at z_0. In terms of $z'(t_0)$ this can

be rewritten as

$$w'(t_0) = \frac{1}{2}\left(\frac{\partial f}{\partial x} - i\,\frac{\partial f}{\partial y}\right) z'(t_0) + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\,\frac{\partial f}{\partial y}\right)\overline{z'(t_0)}.$$

If angles are preserved, $\arg [w'(t_0)/z'(t_0)]$ must be independent of $\arg z'(t_0)$. The expression

$$\frac{1}{2}\left(\frac{\partial f}{\partial x} - i\,\frac{\partial f}{\partial y}\right) + \frac{1}{2}\left(\frac{\partial f}{\partial x} + i\,\frac{\partial f}{\partial y}\right)\frac{\overline{z'(t_0)}}{z'(t_0)} \tag{3}$$

must therefore have a constant argument. As $\arg z'(t_0)$ is allowed to vary, the point represented by (3) describes a circle having the radius $\frac{1}{2}|(\partial f/\partial x) + i(\partial f/\partial y)|$. The argument cannot be constant on this circle unless its radius vanishes, and hence we must have

$$\frac{\partial f}{\partial x} = -i\,\frac{\partial f}{\partial y} \tag{4}$$

which is the complex form of the Cauchy-Riemann equations.

Quite similarly, the condition that the change of scale shall be the same in all directions implies that the expression (3) has a constant modulus. On a circle the modulus is constant only if the radius vanishes or if the center lies at the origin. In the first case we obtain (4), and in the second case

$$\frac{\partial f}{\partial x} = i\,\frac{\partial f}{\partial y}.$$

The last equation expresses the fact that $\overline{f(z)}$ is analytic. A mapping by the conjugate of an analytic function with a nonvanishing derivative is said to be *indirectly conformal.* It evidently preserves the size but reverses the sense of angles.

If the mapping of Ω by $w = f(z)$ is topological, then the inverse function $z = f^{-1}(w)$ is also analytic. This follows easily if $f'(z) \neq 0$, for then the derivative of the inverse function must be equal to $1/f'(z)$ at the point $z = f^{-1}(w)$. We shall prove later that $f'(z)$ can never vanish in the case of a topological mapping by an analytic function.

The knowledge that $f'(z_0) \neq 0$ is sufficient to conclude that the mapping is topological if it is restricted to a sufficiently small neighborhood of z_0. This follows by the theorem on implicit functions known from the calculus, for the Jacobian of the functions $u = u(x,y)$, $v = v(x,y)$ at the point z_0 is $|f'(z_0)|^2$ and hence $\neq 0$. Later we shall present a simpler proof of this important theorem.

But even if $f'(z) \neq 0$ throughout the region Ω, we cannot assert that the mapping of the whole region is necessarily topological. To illustrate

what may happen we refer to Fig. 3-1. Here the mappings of the sub-

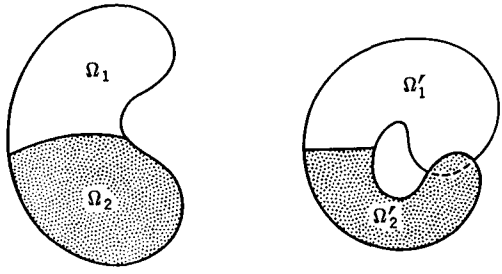

FIG. 3-1. Doubly covered region.

regions Ω_1 and Ω_2 are one to one, but the images overlap. It is helpful to think of the image of the whole region as a transparent film which partly covers itself. This is the simple and fruitful idea used by Riemann when he introduced the generalized regions now known as *Riemann surfaces.*

2.4. Length and Area. We have found that under a conformal mapping $f(z)$ the length of an infinitesimal line segment at the point z is multiplied by the factor $|f'(z)|$. Because the distortion is the same in all directions, infinitesimal areas will clearly be multiplied by $|f'(z)|^2$.

Let us put this on a rigorous basis. We know from calculus that the length of a differentiable arc γ with the equation $z = z(t) = x(t) + iy(t)$, $a \leqq t \leqq b$, is given by

$$L(\gamma) = \int_a^b \sqrt{x'(t)^2 + y'(t)^2}\, dt = \int_a^b |z'(t)|\, dt.$$

The image curve γ' is determined by $w = w(t) = f(z(t))$ with the derivative $w'(t) = f'(z(t))z'(t)$. Its length is thus

$$L(\gamma') = \int_a^b |f'(z(t))|\,|z'(t)|\, dt.$$

It is customary to use the shorter notations

$$L(\gamma) = \int_\gamma |dz|, \qquad L(\gamma') = \int_\gamma |f'(z)|\,|dz|. \tag{5}$$

Observe that in complex notation the calculus symbol ds for integration with respect to arc length is replaced by $|dz|$.

Now let E be a point set in the plane whose area

$$A(E) = \iint_E dx\, dy$$

can be evaluated as a double Riemann integral. If $f(z) = u(x,y) + iv(x,y)$ is a bijective differentiable mapping, then by the rule for changing integration variables the area of the image $E' = f(E)$ is given by

$$A(E') = \iint_E |u_x v_y - u_y v_x| \, dx \, dy.$$

But if $f(z)$ is a conformal mapping of an open set containing E, then $u_x v_y - u_y v_x = |f'(z)|^2$ by virtue of the Cauchy-Riemann equations, and we obtain

$$A(E') = \iint_E |f'(z)|^2 \, dx \, dy. \tag{6}$$

The formulas (5) and (6) have important applications in the part of complex analysis that is frequently referred to as geometric function theory.

3. LINEAR TRANSFORMATIONS

Of all analytic functions the first-order rational functions have the simplest mapping properties, for they define mappings of the extended plane onto itself which are at the same time conformal and topological. The linear transformations have also very remarkable geometric properties, and for that reason their importance goes far beyond serving as simple examples of conformal mappings. The reader will do well to pay particular attention to this geometric aspect, for it will equip him with simple but very valuable techniques.

3.1. The Linear Group. We have already remarked in Chap. 2, Sec. 1.4 that a *linear fractional transformation*

$$w = S(z) = \frac{az + b}{cz + d} \tag{7}$$

with $ad - bc \neq 0$ has an inverse

$$z = S^{-1}(w) = \frac{dw - b}{-cw + a}.$$

The special values $S(\infty) = a/c$ and $S(-d/c) = \infty$ can be introduced either by convention or as limits for $z \to \infty$ and $z \to -d/c$. With the latter interpretation it becomes obvious that S is a topological mapping of the extended plane onto itself, the topology being defined by distances on the Riemann sphere.

For linear transformations we shall usually replace the notation $S(z)$

by Sz. The representation (7) is said to be normalized if $ad - bc = 1$. It is clear that every linear transformation has two normalized representations, obtained from each other by changing the signs of the coefficients.

A convenient way to express a linear transformation is by use of homogeneous coordinates. If we write $z = z_1/z_2$, $w = w_1/w_2$ we find that $w = Sz$ if

$$\begin{aligned} w_1 &= az_1 + bz_2 \\ w_2 &= cz_1 + dz_2 \end{aligned} \tag{8}$$

or, in matrix notation,

$$\begin{pmatrix} w_1 \\ w_2 \end{pmatrix} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \begin{pmatrix} z_1 \\ z_2 \end{pmatrix}.$$

The main advantage of this notation is that it leads to a simple determination of a composite transformation $w = S_1S_2z$. If we use subscripts to distinguish between the matrices that correspond to S_1, S_2 it is immediate that S_1S_2 belongs to the matrix product

$$\begin{pmatrix} a_1 & b_1 \\ c_1 & d_1 \end{pmatrix} \begin{pmatrix} a_2 & b_2 \\ c_2 & d_2 \end{pmatrix} = \begin{pmatrix} a_1a_2 + b_1c_2 & a_1b_2 + b_1d_2 \\ c_1a_2 + d_1c_2 & c_1b_2 + d_1d_2 \end{pmatrix}.$$

All linear transformations form a group. Indeed, the associative law $(S_1S_2)S_3 = S_1(S_2S_3)$ holds for arbitrary transformations, the identity $w = z$ is a linear transformation, and the inverse of a linear transformation is linear. The ratios $z_1:z_2 \neq 0:0$ are the points of the complex projective line, and (8) identifies the group of linear transformations with the one-dimensional projective group over the complex numbers, usually denoted by $P(1,\mathbf{C})$. If we use only normalized representations, we can also identify it with the group of two-by-two matrices with determinant 1 (denoted $SL(2,\mathbf{C})$), except that there are two opposite matrices corresponding to the same linear transformation.

We shall make no further use of the matrix notation, except for remarking that the simplest linear transformations belong to matrices of the form

$$\begin{pmatrix} 1 & \alpha \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} k & 0 \\ 0 & 1 \end{pmatrix}, \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

The first of these, $w = z + \alpha$, is called a *parallel translation*. The second, $w = kz$, is a *rotation* if $|k| = 1$ and a *homothetic transformation* if $k > 0$. For arbitrary complex $k \neq 0$ we can set $k = |k| \cdot k/|k|$, and hence $w = kz$ can be represented as the result of a homothetic transformation followed by a rotation. The third transformation, $w = 1/z$, is called an *inversion*.

If $c \neq 0$ we can write

$$\frac{az + b}{cz + d} = \frac{bc - ad}{c^2(z + d/c)} + \frac{a}{c},$$

and this decomposition shows that the most general linear transformation is composed by a translation, an inversion, a rotation, and a homothetic transformation followed by another translation. If $c = 0$, the inversion falls out and the last translation is not needed.

EXERCISES

1. Prove that the reflection $z \to \bar{z}$ is not a linear transformation.

2. If

$$T_1 z = \frac{z+2}{z+3}, \qquad T_2 z = \frac{z}{z+1},$$

find $T_1 T_2 z$, $T_2 T_1 z$ and $T_1^{-1} T_2 z$.

3. Prove that the most general transformation which leaves the origin fixed and preserves all distances is either a rotation or a rotation followed by reflexion in the real axis.

4. Show that any linear transformation which transforms the real axis into itself can be written with real coefficients.

3.2. The Cross Ratio. Given three distinct points z_2, z_3, z_4 in the extended plane, there exists a linear transformation S which carries them into 1, 0, ∞ in this order. If none of the points is ∞, S will be given by

$$Sz = \frac{z - z_3}{z - z_4} : \frac{z_2 - z_3}{z_2 - z_4}. \tag{9}$$

If z_2, z_3 or $z_4 = \infty$ the transformation reduces to

$$\frac{z - z_3}{z - z_4}, \qquad \frac{z_2 - z_4}{z - z_4}, \qquad \frac{z - z_3}{z_2 - z_3}$$

respectively.

If T were another linear transformation with the same property, then ST^{-1} would leave 1, 0, ∞ invariant. Direct calculation shows that this is true only for the identity transformation, and we would have $S = T$. We conclude that S is uniquely determined.

Definition 12. *The cross ratio* (z_1, z_2, z_3, z_4) *is the image of* z_1 *under the linear transformation which carries* z_2, z_3, z_4 *into* 1, 0, ∞.

The definition is meaningful only if z_2, z_3, z_4 are distinct. A conventional value can be introduced as soon as any three of the points are distinct, but this is unimportant.

The cross ratio is invariant under linear transformations. In more precise formulation:

Theorem 12. *If z_1, z_2, z_3, z_4 are distinct points in the extended plane and T any linear transformation, then $(Tz_1,Tz_2,Tz_3,Tz_4) = (z_1,z_2,z_3,z_4)$.*

The proof is immediate, for if $Sz = (z,z_2,z_3,z_4)$, then ST^{-1} carries Tz_2, Tz_3, Tz_4 into 1, 0, ∞. By definition we have hence

$$(Tz_1,Tz_2,Tz_3,Tz_4) = ST^{-1}(Tz_1) = Sz_1 = (z_1,z_2,z_3,z_4).$$

With the help of this property we can immediately write down the linear transformation which carries three given points z_1, z_2, z_3 to prescribed positions w_1, w_2, w_3. The correspondence must indeed be given by

$$(w,w_1,w_2,w_3) = (z,z_1,z_2,z_3).$$

In general it is of course necessary to solve this equation with respect to w.

Theorem 13. *The cross ratio (z_1,z_2,z_3,z_4) is real if and only if the four points lie on a circle or on a straight line.*

This is evident by elementary geometry, for we obtain

$$\arg (z_1,z_2,z_3,z_4) = \arg \frac{z_1 - z_3}{z_1 - z_4} - \arg \frac{z_2 - z_3}{z_2 - z_4},$$

and if the points lie on a circle this difference of angles is either 0 or $\pm\pi$, depending on the relative location.

For an analytic proof we need only show that the image of the real axis under any linear transformation is either a circle or a straight line. Indeed, $Tz = (z,z_2,z_3,z_4)$ is real on the image of the real axis under the transformation T^{-1} and nowhere else.

The values of $w = T^{-1}z$ for real z satisfy the equation $Tw = \overline{Tw}$. Explicitly, this condition is of the form

$$\frac{aw + b}{cw + d} = \frac{\bar{a}\bar{w} + \bar{b}}{\bar{c}\bar{w} + \bar{d}}.$$

By cross multiplication we obtain

$$(a\bar{c} - c\bar{a})|w^2| + (a\bar{d} - c\bar{b})w + (b\bar{c} - d\bar{a})\bar{w} + b\bar{d} - d\bar{b} = 0.$$

If $a\bar{c} - c\bar{a} = 0$ this is the equation of a straight line, for under this condition the coefficient $a\bar{d} - c\bar{b}$ cannot also vanish. If $a\bar{c} - c\bar{a} \neq 0$ we can