A First Course in Geometric Topology and Differential Geometry

Ethan D. Bloch

Birkhäuser

Dedicated to my parents,

in appreciation for all they have taught me

Ethan D. Bloch

A First Course in Geometric Topology and Differential Geometry

Birkhäuser
Boston • Basel • Berlin

Ethan D. Bloch
Department of Natural Sciences and Mathematics
Bard College
Annandale, New York 12504
USA

Library of Congress Cataloging In-Publication Data

Bloch, Ethan, 1956-
A first course in geometric topology and differential geometry /
Ethan Bloch.
p. cm.
Includes bibliographical references (p. -) and index.
ISBN 0-8176-3840-7 (h : alk. paper). -- ISBN 3-7643-3840-7 (H :
alk. paper)
1. Topology. 2. Geometry, Differential. I. Title.
QA611.ZB55 1996 95-15470
516.3'63--dc20 CIP

Printed on acid-free paper

ISBN 0-8176-3840-7
ISBN 3-7643-3840-7

Typeset from author's disk in $\mathcal{AMS}$-TEX by TEXniques, Inc., Boston, MA
Illustrations by Carl Twarog, Greenville, NC
Printed and bound by Maple-Vail, York, PA
Printed in the U.S.A.
9 8 7 6 5 4 3 2 1

Contents

Introduction

This text is an introduction to geometric topology and differential geometry via the study of surfaces, and more generally serves to introduce the student to the relation of the modern axiomatic approach in mathematics to geometric intuition. The idea of combining geometry and topology in a text is, of course, not new; the present text attempts to make such a combination of subjects accessible to the junior/senior level mathematics major at a university or college in the United States. Though some of the deep connections between the topology and geometry of manifolds can only be dealt with using more advanced techniques than those presented here, we do reach the classical Gauss–Bonnet Theorem — a model theorem for the relation of topology and geometry — at the end of the book.

The notion of a surface is the unifying thread of the text. Our treatment of point set topology is brief and restricted to subsets of Euclidean spaces; the discussion of topological surfaces is geometric rather than algebraic; the treatment of differential geometry is classical, treating surfaces in $\mathbb{R}^3$. The goal of the book is to reach a number of intuitively appealing definitions and theorems concerning surfaces in the topological, polyhedral and smooth cases. Some of the goodies aimed at are the classification of compact surfaces, the Gauss–Bonnet Theorem (polyhedral and smooth) and the geodesic nature of length-minimizing curves on surfaces. Only those definitions and methods needed for these ends are developed. In order to keep the discussion at a concrete level, we avoid treating a number of technicalities such as abstract topological spaces, abstract simplicial complexes and tensors. As a result, at times some proofs seem a bit more circuitous than might be standard, though we feel that the gain in avoiding unnecessary technicalities is worthwhile.

There are a variety of ways in which this book could be used for a semester course. For students with no exposure to topology, the first three chapters, together with a sampling from Chapters IV and V, could be used as a one-semester introduction to point set and geometric topology, with a taste of smooth surfaces thrown in. Alternately, Chapters IV–VIII could be used as a quite leisurely first course on differential geometry (skipping the few instances where the previous chapters are used, and adding an intuitive discussion of the Euler characteristic for the Gauss–Bonnet Theorem). Students who have had a semester of point set topology (or a real analysis course in which either $\mathbb{R}^n$ or metric spaces are discussed), could cover a fair bit of Chapters II–VIII in one

semester, though some material would probably have to be dropped. It is also hoped that the book could be used for individual study.

This book developed out of lecture notes for a course at Bard College first given in the spring of 1991. I would like to thank Bard students Melissa Cahoon, Jeff Bolden, Robert Cutler, David Steinberg, Anne Willig, Farasat Bokhari, Diego Socolinsky and Jason Foulkes for helpful comments on various drafts of the original lecture notes. Thanks are also due to Matthew Deady, Peter Dolan, Mark Halsey, David Nightingale and Leslie Morris, as well as to the Mathematics Institute at Bar-Ilan University in Israel and the Mathematics Department at the University of Pennsylvania, who hosted me when various parts of this book were written. It is, of course, impossible to acknowledge every single topology and differential geometry text from which I have learned about the subject, and to credit the source of every definition, lemma and theorem (especially since many of them are quite standard); I have acknowledged in the text particular sources for lengthy or non-standard proofs. See the section entitled Further Study for books that have particularly influenced this text. For generally guiding my initial development as a mathematician I would like to thank my professors at Reed College and Cornell University, and in particular my advisor, Professor David Henderson of Cornell. Finally, I would like to thank Ann Kostant, mathematics editor at Birkhäuser, for her many good ideas, and the helpful staff at Birkhäuser for turning my manuscript into a finished book.

To the Student

Surfaces

Surfaces can be approached from two viewpoints, topological and geometric, and we cover both these approaches. There are three different categories of surfaces (and, more generally, "manifolds," a generalization of surfaces to all dimensions) that we discuss: topological, simplicial and smooth. In contrast to higher dimensional analogs of surfaces, in dimension two (the dimension of concern in this book), all three types of surfaces turn out to have the same topological properties. Hence, for our topological study we will concentrate on topological and simplicial surfaces. This study, called geometric topology, is covered in Chapters II–III.

Geometrically, on the other hand, the three types of surfaces behave quite differently from each other. Indeed, topological surfaces can sit very wildly in Euclidean space, and do not have sufficient structure to allow for manageable geometric analysis. Simplicial surfaces can be studied geometrically, as, for example, in Section 3.9. The most interesting, deep and broadly applicable study of the geometry of surfaces involves smooth surfaces. Our study of smooth surfaces will thus be fundamentally different in both aim and flavor than our study of topological and simplicial surfaces, focusing on geometry rather than topology, and on local rather than global results. The methodology for smooth surfaces involves calculus, rather than point-set topology. This study, called differential geometry, is studied in Chapters IV–VIII. Although apparently distinct, geometric topology and differential geometry come together in the amazing Gauss–Bonnet Theorem, the final result in the book.

Prerequisites

This text should be accessible to mathematics majors at the junior or senior level in a university or college in the United States. The minimal prerequisites are a standard calculus sequence (including multivariable calculus and an acquaintance with differential equations), linear algebra (including inner products) and familiarity with proofs and the basics of sets and functions. Abstract algebra and real analysis are not required. There are two proofs (Theorem 1.5.2 and Proposition 1.6.7) where the Least Upper Bound Property of the real numbers is

used; the reader who has not seen this property (for example, in a real analysis course) can skip these proofs. If the reader has had a course in point-set topology, or a course in real analysis where the setting is either $\mathbb{R}^n$ or metric spaces, then much of Chapter I could probably be skipped over. In a few instances we make use of affine linear maps, a topic not always covered in a standard linear algebra course; all the results we need concerning such maps are summarized in the Appendix.

Rigor vs. Intuition

The study of surfaces from topological, polyhedral and smooth points of view is ideally suited for displaying the interaction between rigor and geometric intuition applied to objects that have inherent appeal. In addition to informal discussion, every effort has been made to present a completely rigorous treatment of the subject, including a careful statement of all the assumptions that are used without proof (such as the triangulability of compact surfaces). The result is that the material in this book is presented as dictated by the need for rigor, in contrast to many texts which start out more easily and gradually become more difficult. Thus we have the odd circumstance of Section 2.2, for example, being much more abstract than some of the computational aspects of Chapter V. The reader might choose to skip some of the longer proofs in the earlier chapters upon first reading.

At the end of the book is a guide to further study, to which the reader is referred both for collateral readings (some of which have a more informal, intuitive approach, whereas others are quite rigorous), and for references for more advanced study of topology and differential geometry.

Exercises

Doing the exercises is a crucial part of learning the material in this text. A good portion of the exercises are results that are needed in the text; such exercises have been marked with an asterisk (*). Exercises range from routine computations (particularly in the chapters on differential geometry) to rather tricky proofs. No attempt has been made to rate the difficulty of the problems, since doing so is highly subjective. There are hints for some of the exercises in the back of the book.

A First Course in Geometric Topology and Differential Geometry

CHAPTER I

Topology of Subsets of Euclidean Space

1.1. Introduction

Although the goal of this book is the study of surfaces, in order to have the necessary tools for a rigorous discussion of the subject, we need to start off by considering some more general notions concerning the topology of subsets of Euclidean space. In contrast to geometry, which is the study of quantitative properties of spaces, that is, those properties that depend upon measurement (such as length, angle and area), topology is the study of the qualitative properties of spaces. For example, from a geometric point of view, a circle of radius 1 and a circle of radius 2 are quite distinct — they have different diameters, different areas, etc.; from a qualitative point of view these two circles are essentially the same. One circle can be deformed into the other by stretching, but without cutting or gluing. From a topological point of view a circle is also indistinguishable from a square. On the other hand, a circle is topologically quite different from a straight line; intuitively, a circle would have to be cut to obtain a straight line, and such a cut certainly changes the qualitative properties of the object.

While at first glance the study of qualitative properties of objects may seem vague and possibly unimportant, such a study is in fact fundamental to a more advanced understanding of such diverse areas as geometry and differential equations. Indeed, the subject of topology arose in the 19th century out of the study of differential equations and analysis. As usually happens in mathematics, once an interesting subject gets started it tends to take off on its own, and today most topologists study topology for its own sake. The subject of topology is divided into three main areas:

(1) point set topology — the most dry and formal aspect of the three, and the least popular as an area of research, but the basis for the rest of topology;

(2) geometric topology — the study of familiar geometric objects such as surfaces and their generalizations to higher dimensions by relatively concrete means, and as such the most intuitively appealing branch of topology (the author's bias);

(3) algebraic topology — the application of the methods of abstract algebra (for example, groups) to the study of topological spaces.

In this chapter we will be dealing with some aspects of point set topology; in Chapters II and III we will be dealing with geometric topology. We will not be making use of algebraic topology in this book, although for any further topological study of surfaces and other geometric objects algebraic tools are quite important.

Throughout this book we will be using the following notation. Let $\mathbb{Z}$, $\mathbb{Z}^+$, $\mathbb{Q}$, $\mathbb{R}$ denote the sets of integers, positive integers, rational numbers and real numbers respectively. Let $\mathbb{R}^n$ denote n-dimensional Euclidean space. We will let $u_1, \ldots, u_n$ denote the coordinates of $\mathbb{R}^n$, though for $\mathbb{R}^3$ we will often use x, y, z for readability. Let O_n denote the origin in $\mathbb{R}^n$. If v and w are vectors in $\mathbb{R}^n$, let $\langle v, w \rangle$ denote their inner product, and let $||v||$ denote the norm of v. Finally, we let $\mathbb{H}^n$ denote the closed upper half-space in $\mathbb{R}^n$, which is the set

$$\mathbb{H}^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n | x_n \geq 0 \right\}.$$

The boundary of $\mathbb{H}^n$ is the set $\partial\mathbb{H}^n \subset \mathbb{H}^n$ defined by

$$\partial\mathbb{H}^n = \left\{ \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix} \in \mathbb{R}^n | x_n = 0 \right\}.$$

(Observe that $\partial\mathbb{H}^n$ is just $\mathbb{R}^{n-1}$ sitting inside $\mathbb{R}^n$).

1.2. Open and Closed Subsets of Sets in $\mathbb{R}^n$

We start by recalling the concept of an interval in the real number line.

Definition. Let $a, b \in \mathbb{R}$ be any two points; we define the following sets:

Open interval:

$$(a, b) = \{x \in \mathbb{R} \mid a < x < b\}.$$

Closed interval:

$$[a, b] = \{x \in \mathbb{R} \mid a \leq x \leq b\}.$$

Half-open intervals:

$$[a, b) = \{x \in \mathbb{R} \mid a \leq x < b\},$$
$$(a, b] = \{x \in \mathbb{R} \mid a < x \leq b\}.$$

Infinite intervals:

$$[a, \infty) = \{x \in \mathbb{R} \mid a \leq x\},$$
$$(a, \infty) = \{x \in \mathbb{R} \mid a < x\},$$
$$(-\infty, b] = \{x \in \mathbb{R} \mid x \leq b\},$$
$$(-\infty, b) = \{x \in \mathbb{R} \mid x < b\},$$
$$(-\infty, \infty) = \mathbb{R}. \quad \Diamond$$

Observe that there are no intervals that are "closed" at ∞ or $-\infty$ (for example, there is no interval of the form $[a, \infty]$), since ∞ is not a real number, and therefore it cannot be included in an interval contained in the real numbers. We are simply using the symbol ∞ to tell us that an interval is unbounded.

The words "open" and "closed" here are used in a very deliberate manner, reflecting a more general concept about to be defined for all $\mathbb{R}^n$. Intuitively, an open set (in this case an interval) is a set that does not contain its "boundary" (which in the case of an interval is its endpoints); a closed set is one that does contain its boundary. A set such as a half-open interval is neither open nor closed. In dimensions higher than 1 the situation is trickier. The closest analog in $\mathbb{R}^2$ to an interval in $\mathbb{R}$ would be a rectangle; we could define an open rectangle (that is, all points $\binom{x}{y} \in \mathbb{R}^2$ such that $a < x < b$ and $c < y < d$), a closed rectangle, etc. See Figure 1.2.1. Unfortunately, whereas intervals account for a large portion of the subsets of $\mathbb{R}$ encountered on a regular basis, rectangles are not nearly so prominent among subsets of $\mathbb{R}^2$. Much more common are blob-shaped subsets of $\mathbb{R}^2$, which can still come in open, closed, or neither varieties. See Figure 1.2.2. Although the idea of openness and closedness still refers intuitively to whether a subset contains its boundary or not, one cannot characterize open and closed sets in $\mathbb{R}^n$ simply in terms of inequalities.

Let us start with open sets in $\mathbb{R}^n$; we will define closed sets later on in terms of open sets.

Definition. Let $p \in \mathbb{R}^n$ be a point, and let $r > 0$ be a number. The **open ball**

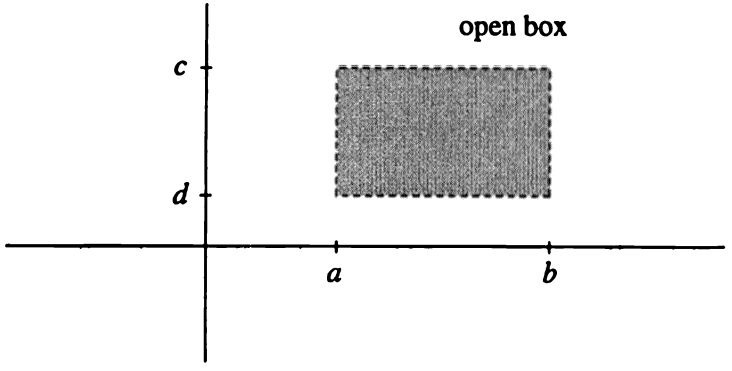


Figure 1.2.1

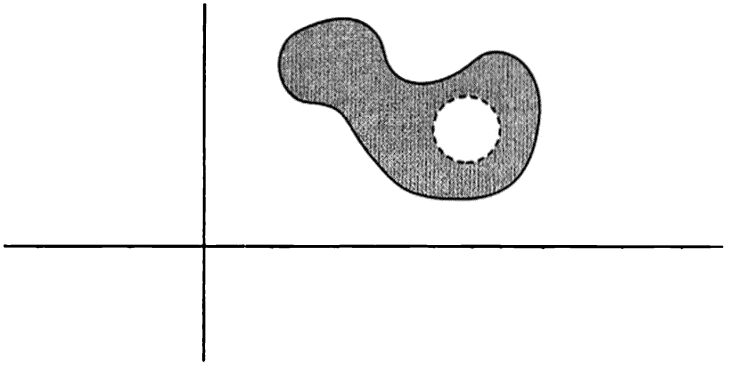


Figure 1.2.2

in $\mathbb{R}^n$ of radius r centered at p is the set $O_r(p, \mathbb{R}^n)$ defined by

$$O_r(p, \mathbb{R}^n) = \{x \in \mathbb{R}^n \mid \|x - p\| < r\}.$$

More generally, let $A \subset \mathbb{R}^n$ be any subset, let $p \in A$ be a point, and let $r > 0$ be a number. The **open ball** in A of radius r centered at p is the set $O_r(p, A)$ defined by

$$O_r(p, A) = O_r(p, \mathbb{R}^n) \cap A = \{x \in A \mid \|x - p\| < r\}.$$

The **closed ball** in A of radius r centered at p is the set $\overline{O}_r(p, A)$ defined by

$$\overline{O}_r(p, A) = \{x \in A \mid \|x - p\| \leq r\}. \quad \Diamond$$

Example 1.2.1. Let $A \subset \mathbb{R}^2$ be the square $[0, 4] \times [0, 4]$. The sets $O_1(\binom{2}{2}, A)$, $O_1(\binom{2}{0}, A)$ and $\overline{O}_1(\binom{2}{0}, A)$ are shown in Figure 1.2.3 (i). Let $B \subset \mathbb{R}^2$ be the x-axis $\mathbb{R} \times \{0\}$. The sets $O_1(\binom{0}{0}, B)$ and $\overline{O}_1(\binom{0}{0}, B)$ are shown in Figure 1.2.3 (ii). $\Diamond$

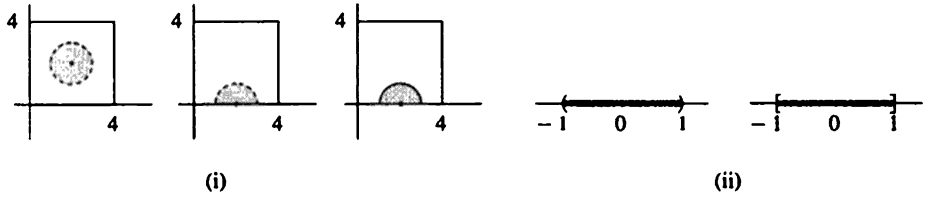


Figure 1.2.3

The following definition yields the intuitive concept we are looking for.

Definition. A subset $A \subset \mathbb{R}^n$ is an **open** subset of $\mathbb{R}^n$ if for each point $p \in A$ there is an open ball centered at p that is entirely contained in A; in other words, for each $p \in A$, there is a number $r > 0$ such that $O_r(p, \mathbb{R}^n) \subset A$. $\Diamond$

One of the simplest examples of an open subset of $\mathbb{R}^n$ is $\mathbb{R}^n$ itself. We also consider the empty set to be an open subset of every $\mathbb{R}^n$; the empty set contains no points, and therefore there is no problem assuming that for each point $p \in \emptyset$ there is an open ball centered at p which is entirely contained in $\emptyset$. (You might worry that by similar arguments one could prove almost anything about the empty set, but that isn't really an objection; further, it works out quite conveniently to have the empty set open.) A more interesting example of open sets is seen in the following lemma, in which it is proved that open balls in $\mathbb{R}^n$ are indeed open sets.

Lemma 1.2.2. *An open ball in $\mathbb{R}^n$ is an open subset of $\mathbb{R}^n$.*

Proof. Let $x \in \mathbb{R}^n$ be a point and let $r > 0$ be a number. To prove that $O_r(x, \mathbb{R}^n)$ is an open set, we need to show that for each point $y \in O_r(x, \mathbb{R}^n)$ there is a number $\epsilon > 0$ such that $O_\epsilon(y, \mathbb{R}^n) \subset O_r(x, \mathbb{R}^n)$. For given y, we choose ϵ to be any positive number such that $\epsilon < r - \|y - x\|$. See Figure 1.2.4. If $z \in O_\epsilon(y, \mathbb{R}^n)$ is any point, then using the triangle inequality we have

$$\|z - x\| \leq \|z - y\| + \|y - x\| < (r - \|y - x\|) + \|y - x\| = r.$$

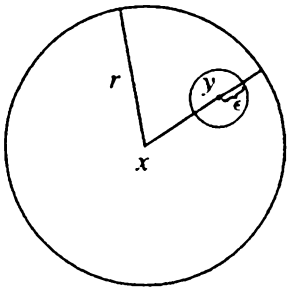


Figure 1.2.4

Hence $z \in O_r(x, \mathbb{R}^n)$, and the result is proved. □

An example of a non-open set is any closed or half-open interval in $\mathbb{R}$. Consider for example the closed interval $[a, b]$. It is seen that any open ball of the form $O_r(b, \mathbb{R})$ contains values greater than b, and so is not contained inside the interval $[a, b]$; similarly for the point a. Hence $[a, b]$ is not open.

The following lemma summarizes the most important properties of open subsets of $\mathbb{R}^n$. In the more general setting of topological spaces these properties are taken axiomatically as the properties that the collection of all open subsets of the topological space must satisfy. Observe that in part (ii) of the lemma the union may be infinite.

Lemma 1.2.3.

(i) *$\mathbb{R}^n$ and $\emptyset$ are open in $\mathbb{R}^n$.*

(ii) *The union of open subsets of $\mathbb{R}^n$ is open.*

(iii) *The intersection of finitely many open subsets of $\mathbb{R}^n$ is open.*

Proof. (i). This was dealt with above.

(ii). Let $\{U_i\}_{i \in I}$ be a collection of open subsets of $\mathbb{R}^n$, and let $x \in \bigcup_{i \in I} U_i$ be a point. Then $x \in U_k$ for some $k \in I$. By the openness of U_k there is a number $r > 0$ such that $O_r(x, \mathbb{R}^n) \subset U_k$. Hence $O_r(x, \mathbb{R}^n) \subset \bigcup_{i \in I} U_i$, and it follows that $\bigcup_{i \in I} U_i$ is open in $\mathbb{R}^n$.

(iii). Let $\{U_1, \ldots, U_m\}$ be a finite collection of open subsets of $\mathbb{R}^n$, and let $x \in \bigcap_{i=1}^m U_i$ be a point. Then $x \in U_i$ for all $i \in \{1, \ldots, m\}$. By the openness of U_i there is a number $r_i > 0$ such that $O_{r_i}(x, \mathbb{R}^n) \subset U_i$. If $r = \min\{r_1, \ldots, r_m\}$ then $O_r(x, \mathbb{R}^n) \subset U_i$ for all i, and hence $O_r(x, \mathbb{R}^n) \subset \bigcap_{i=1}^m U_i$. Thus $\bigcap_{i=1}^m U_i$ is open in $\mathbb{R}^n$. (Note where the finiteness was used.) □

Since we wish to study subsets of $\mathbb{R}^n$, such as surfaces, we need to look a little more closely at open sets. Consider the closed interval $[0, 2] \subset \mathbb{R}$. The subset $(1, 2]$ is certainly not open in $\mathbb{R}$, but consider it as a subset of $[0, 2]$. The obstacle to $(1, 2]$ being open in $\mathbb{R}$ is that there is no open ball in $\mathbb{R}$ centered at 2 and entirely contained in $(1, 2]$. If, however, we think of $[0, 2]$ as our entire universe, then we cannot really have the same complaint against $(1, 2]$, since we can have as much of an open ball centered at 2 in $(1, 2]$ as in $[0, 2]$. A set which is not open in $\mathbb{R}$ can still be viewed as open in some sense when sitting in a proper subset of $\mathbb{R}$. The same considerations hold for $\mathbb{R}^n$.

Definition. Let $A \subset \mathbb{R}^n$ be a set. A subset $S \subset A$ is a **relatively open** subset of A, often referred to simply as an **open** subset of A, if for each point $p \in S$, there is an open ball in A centered at p that is entirely contained in S. In other words, for each $p \in S$, there is a number $r > 0$ such that $O_r(p, A) \subset S$. If $p \in A$ is a point, then an **open neighborhood** in A of p is an open subset of A containing p. $\Diamond$

Example 1.2.4. The set

$$A = \{\binom{x}{y} \in \mathbb{R}^2 \mid x > 0 \text{ and } y \geq 0\}$$

is an open subset of the closed upper half-plane $\mathbb{H}^2$ (though it is not open in $\mathbb{R}^2$). The reader can supply the details. $\Diamond$

The above definition makes it clear that we cannot simply speak of an "open set" without saying in what it is open. The following lemma gives a useful characterization of relatively open sets.

Lemma 1.2.5. *Let $A \subset \mathbb{R}^n$ be a set. A subset $S \subset A$ is an open subset of A iff there exists an open subset U of $\mathbb{R}^n$ such that $S = U \cap A$.*

Proof. Suppose first that S is an open subset of A. By hypothesis, for each $p \in S$ there is a number $r_p > 0$ such that $O_{r_p}(p, A) \subset S$. It is not hard to see that

$$S = \bigcup_{p \in S} O_{r_p}(p, A).$$

Let $U \subset \mathbb{R}^n$ be defined by

$$U = \bigcup_{p \in S} O_{r_p}(p, \mathbb{R}^n).$$

By Lemmas 1.2.2 and 1.2.3 it follows that U is an open subset of $\mathbb{R}^n$. Further,

$$U \cap A = \Big[\bigcup_{p \in S} O_{r_p}(p, \mathbb{R}^n)\Big] \cap A = \bigcup_{p \in S} \big[O_{r_p}(p, \mathbb{R}^n) \cap A\big]$$
$$= \bigcup_{p \in S} O_{r_p}(p, A) = S.$$

Now suppose that there exists an open subset U of $\mathbb{R}^n$ such that $S = U \cap A$. Every point $p \in S$ is also in U, and hence for each such p there is a number $r > 0$ such that $O_r(p, \mathbb{R}^n) \subset U$. Hence

$$O_r(p, A) = O_r(p, \mathbb{R}^n) \cap A \subset U \cap A = S,$$

and it follows that S is open in A. □

As seen in Example 1.2.4, if $A \subset \mathbb{R}^n$ is a set and U is an open subset of A, then it does not necessarily follow that U is an open subset of $\mathbb{R}^n$. If, however, the set A is itself open in $\mathbb{R}^n$, then, as seen in the following lemma, everything works out as nicely as possible.

Lemma 1.2.6. *Let $A \subset B \subset C \subset \mathbb{R}^n$ be sets. If A is an open subset of B, and B is an open subset of C, then A is an open subset of C.*

Proof. By Lemma 1.2.5 there exist sets $A', B' \subset \mathbb{R}^n$ which are open in $\mathbb{R}^n$ and such that $A = A' \cap B$ and $B = B' \cap C$. Then $A = A' \cap (B' \cap C) = (A' \cap B') \cap C$. Since $A' \cap B'$ is an open subset of $\mathbb{R}^n$ by Lemma 1.2.3, it follows from Lemma 1.2.5 that A is open in C. □

The properties stated in Lemma 1.2.3 for open subsets of $\mathbb{R}^n$ also hold for open subsets of any subset of $\mathbb{R}^n$. The following lemma is proved similarly to Lemma 1.2.3, and we omit the proof.

Lemma 1.2.7. *Let $A \subset \mathbb{R}^n$ be any set.*

(i) *A and $\emptyset$ are open in A.*
(ii) *The union of open subsets of A is open in A.*
(iii) *The intersection of finitely many open subsets of A is open in A.*

The following lemma is a relative version of Lemma 1.2.5.

Lemma 1.2.8. *Let $A \subset B \subset \mathbb{R}^n$ be subsets. A subset $U \subset A$ is open in A iff there is an open subset V of B such that $U = V \cap A$.*

Proof. Suppose U is an open subset of A. By Lemma 1.2.5, there is an open subset U' of $\mathbb{R}^n$ such that $U = U' \cap A$. If we define the set V to be $V = U' \cap B$,

then V is an open subset of B, and $V \cap A = (U' \cap B) \cap A = U' \cap (B \cap A) = U' \cap A = U$ as desired.

Now suppose that U is a subset of A such that $U = V \cap A$ for some open subset V of B. Then there exists an open subset V' of $\mathbb{R}^n$ such that $V = V' \cap B$. Therefore $U = V \cap A = (V' \cap B) \cap A = V' \cap (B \cap A) = V' \cap A$, which means that U is an open subset of A. □

The following lemma discusses the behavior of open sets in products and will be technically important later on.

Lemma 1.2.9. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets.*

(i) *If $U \subset A$ and $V \subset B$ are open subsets, then $U \times V$ is an open subset of $A \times B$.*

(ii) *If $W \subset A \times B$ is an open set, then for every point $(p_1, p_2) \in W$ there are numbers $\epsilon_1, \epsilon_2 > 0$ such that $O_{\epsilon_1}(p_1, A) \times O_{\epsilon_2}(p_2, B) \subset W$.*

Proof. Exercise 1.2.7. □

We now turn to closed sets, once again starting with the non-relative case. Although we have an intuitive notion of a closed set as one that contains its boundary, the easiest way to define a closed set is as follows, entirely ignoring the notion of boundary.

Definition. A subset $C \subset \mathbb{R}^n$ is **closed** in $\mathbb{R}^n$ if the complement of C, namely $\mathbb{R}^n - C$, is an open subset of $\mathbb{R}^n$. ◊

As seen in the following example, it is important to realize that a set in $\mathbb{R}^n$ can be open, closed, both, or neither. Hence one cannot demonstrate that a set is closed by showing that it is not open.

Example 1.2.10. It is seen in Exercise 1.2.1 that the complement of a single point in $\mathbb{R}^n$ is an open subset of $\mathbb{R}^n$; hence a single point in $\mathbb{R}^n$ is a closed subset. A closed interval in $\mathbb{R}$ is a closed subset, since the complement in $\mathbb{R}$ of an interval of the form $[a, b]$ is the set $(-\infty, a) \cup (b, \infty)$, and this latter set is open in $\mathbb{R}$. A half-open interval $(a, b]$ in $\mathbb{R}$ is neither open nor closed, as the reader can verify. The set $\mathbb{R}^n$ is both open and closed in $\mathbb{R}^n$; we have already seen that it is open, and observe that $\mathbb{R}^n - \mathbb{R}^n = \emptyset$, which is also open in $\mathbb{R}^n$. ◊

The following lemma is the analog for closed sets of Lemma 1.2.3.

Lemma 1.2.11.

(i) $\mathbb{R}^n$ *and* $\emptyset$ *are closed in* $\mathbb{R}^n$.
(ii) *The union of finitely many closed subsets of* $\mathbb{R}^n$ *is closed.*
(iii) *The intersection of closed subsets of* $\mathbb{R}^n$ *is closed.*

Proof. Exercise 1.2.9. □

Based upon our experience with relatively open sets, two possible ways of defining relatively closed sets come to mind: complements of relatively open subsets and intersections with closed sets of $\mathbb{R}^n$. The following lemma shows that these two methods yield the same results.

Lemma 1.2.12. *Let* $C \subset A \subset \mathbb{R}^n$ *be sets. Then the set* $A - C$ *is open in* A *iff there exists a closed subset* D *of* $\mathbb{R}^n$ *such that* $C = D \cap A$.

Proof. $\Rightarrow$. Since $A - C$ is open in A, by Lemma 1.2.5 there exists an open subset U of $\mathbb{R}^n$ such that $A - C = U \cap A$. Observe that $A - U = C$. Let $D = \mathbb{R}^n - U$, which is closed in $\mathbb{R}^n$ by definition. Using standard properties of set operations we now have

$$D \cap A = \big[\mathbb{R}^n - U\big] \cap A = \big[\mathbb{R}^n \cap A\big] - U = A - U = C.$$

$\Leftarrow$. By hypothesis there exists a closed subset D of $\mathbb{R}^n$ such that $C = D \cap A$. Let $U = \mathbb{R}^n - D$, which is open in $\mathbb{R}^n$ by definition. Hence

$$A - C = A - \big[D \cap A\big] = A - D = A \cap \big[\mathbb{R}^n - D\big] = A \cap U,$$

where the last set is open in A by Lemma 1.2.5. □

We can now make the following definition in good conscience.

Definition. Let $A \subset \mathbb{R}^n$ be a set. A subset $C \subset A$ is a **relatively closed** subset of A, often referred to simply as a **closed** subset of A, if either of the two conditions in Lemma 1.2.12 holds. ◊

Example 1.2.13. The interval $(0, 1]$ is a closed subset of the interval $(0, 2)$, since the set $(0, 2) - (0, 1] = (1, 2)$ is open in $(0, 2)$. ◊

The properties stated in Lemma 1.2.11 for closed subsets of $\mathbb{R}^n$ also hold for closed subsets of any subset of $\mathbb{R}^n$; as before, we omit the proof.

Lemma 1.2.14. *Let $A \subset \mathbb{R}^n$ be any set.*

(i) *A and $\emptyset$ are closed in A.*
(ii) *The union of finitely many closed subsets of A is closed in A.*
(iii) *The intersection of closed subsets of A is closed in A.*

Consider the interval $(0, 1) \subset \mathbb{R}$. Certainly $(0, 1)$ is not closed in $\mathbb{R}$, though it is contained in a variety of closed subsets of $\mathbb{R}$, such as $[-17, 25.731]$. There is, however, a "smallest" closed subset of $\mathbb{R}$ containing $(0, 1)$, namely $[0, 1]$. The following definition and lemma show that there exists a smallest closed subset containing any given set.

Definition. Let $D \subset A \subset \mathbb{R}^n$ be sets. The **closure** of D in A, denoted $\overline{D}$, is defined to be the intersection of all closed subsets of A containing D. $\Diamond$

Two comments about the above definition. First, since A is closed in itself and contains D, there is at least one closed subset of A containing D, so the intersection in the definition is well-defined. Second, given any set D, the closure of D in one set containing it need not be the same as the closure of D in some other set containing it. For example, the closures of $(0, 1)$ in each of $(0, 1]$ and $[0, 1)$ are not the same. Although the set in which the closure is taking place is not mentioned in the notation $\overline{D}$, this set should always be clear from the context. That $\overline{D}$ is indeed the smallest closed set containing D is shown by the following lemma.

Lemma 1.2.15. *Let $D \subset A \subset \mathbb{R}^n$ be sets. The set $\overline{D}$ is a closed subset of A containing D and is contained in any other closed subset of A containing D.*

Proof. Exercise 1.2.14. $\square$

Exercises

1.2.1*. Show that the following sets are open in $\mathbb{R}^2$:

(1) the complement of a single point in $\mathbb{R}^n$;

(2) the open upper half-plane in $\mathbb{R}^2$ (that is, the set $\{\begin{pmatrix} x \\ y \end{pmatrix} \in \mathbb{R}^2 \mid y > 0\} = \mathbb{H}^2 - \partial\mathbb{H}^2$);

(3) the set $(1, 2) \times (5, 7)$.

1.2.2*. Find an example to show that the phrase "finitely many" is necessary in the statement of Lemma 1.2.3 (iii).

1.2.3*. Prove that the complement of a finite subset of $\mathbb{R}^n$ is open.

1.2.4*. Prove that a subset of $\mathbb{R}^n$ is open iff it is the union of open balls.

1.2.5*. Let $A \subset \mathbb{R}^n$ be a set. Show that for any point $p \in A$ and any number $r > 0$, the open ball $O_r(p, A)$ is an open subset of A.

1.2.6*. Let $A \subset \mathbb{R}^n$. Show that a subset $U \subset A$ is open in A iff for each point $p \in U$ there is an open subset $V \subset A$ containing p.

1.2.7*. Prove Lemma 1.2.9.

1.2.8. Show that the following sets are closed in $\mathbb{R}^2$:

(1) the straight line $\mathbb{R} \times \{0\}$;

(2) $\mathbb{H}^2$;

(3) the set $[1, 2] \times [5, 7]$.

1.2.9*. Prove Lemma 1.2.11.

1.2.10*. Find an example to show that the phrase "finitely many" is necessary in the statement of Lemma 1.2.11 (ii).

1.2.11*. Prove that a finite subset of $\mathbb{R}^n$ is closed in $\mathbb{R}^n$.

1.2.12*. Let $A \subset \mathbb{R}^n$ be a set of points, possibly infinite, for which there exists a number $D > 0$ such that $\|x - y\| \geq D$ for all points $x, y \in A$. Prove that A is a closed subset of $\mathbb{R}^n$. Find an example to show that the following weaker condition does not suffice to guarantee that a set is closed in $\mathbb{R}^n$: $A \subset \mathbb{R}^n$ is a set of points, possibly infinite, such that for each point $x \in A$ there exists a number $D > 0$ such that $\|x - y\| \geq D$ for all points $y \in A$.

1.2.13*. Let $A \subset B \subset C \subset \mathbb{R}^n$ be sets. Show that if A is a closed subset of B, and B is a closed subset of C, then A is a closed subset of C.

1.2.14*. Prove Lemma 1.2.15.

1.2.15*. Let $A \subset \mathbb{R}^n$ be a set. Show that for any $p \in A$ and any number $r > 0$, the closed ball $\overline{O}_r(p, A)$ is a closed subset of A.

1.2.16*. Let $A \subset \mathbb{R}^n$ be any set. If $U \subset A$ is open in A, and if $x \in U$ is any point, show that $U - \{x\}$ is open in A.

1.2.17*. Let $S \subset \mathbb{R}$ be a closed set that is bounded from above. Show that S contains its least upper bound. Similarly for greatest lower bounds.

1.2.18*. (i) Let $A \subset \mathbb{R}^2$ be a non-empty set contained in a straight line in $\mathbb{R}^2$. Show that A is not open in $\mathbb{R}^2$.

(ii) Let $B \subset \mathbb{R}^2$ be contained in a closed half-plane (that is, all points in $\mathbb{R}^2$ that are either on, or on a given side of, a straight line in $\mathbb{R}^2$). Suppose that B intersects the boundary of the closed half-plane. Show that B is not open in $\mathbb{R}^2$.

(iii) Let $R_1, \ldots, R_p \subset \mathbb{R}^2$ denote p distinct rays from the origin, and let $T_p = R_1 \cup \cdots \cup R_p$, as in Figure 1.2.5. Assume $p \geq 3$. Let $C \subset T_p \times \mathbb{R} \subset \mathbb{R}^3$ be such that C is entirely contained in $(R_1 \cup R_2) \times \mathbb{R}$, and C intersects the z-axis in $\mathbb{R}^3$. Show that C is not open in $T_p \times \mathbb{R}$. (This example may seem far-fetched, but it turns out to be useful later on.)

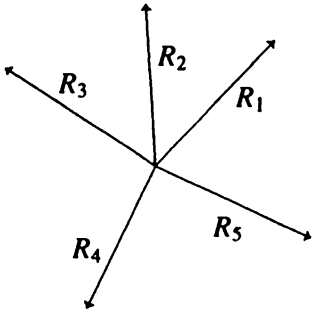


Figure 1.2.5

1.3. Continuous Maps

Continuous maps are to topology what linear maps are to vector spaces, namely maps that preserve the fundamental structures under consideration. Intuitively, continuous maps are those maps that do not "tear" their domains. A continuous map should thus have the property that if two points in the domain get closer and closer to each other, then so do their images. There are a number of equivalent rigorous definitions of continuous maps of subsets of Euclidean space; we will use the standard topological definition of continuity in terms of open sets, rather than the ϵ-δ definition used in calculus and real analysis (though we will refer to the ϵ-δ in Proposition 1.3.3, Example 1.3.4 and a few exercises).

We start with an example of a non-continuous map. Let $f: \mathbb{R} \to \mathbb{R}$ be defined by

$$f(x) = \begin{cases} x, & \text{if } x \leq 0; \\ x + 1, & \text{if } x > 0. \end{cases} \tag{1.3.1}$$

Intuitively, this function is not continuous since there is a "tear" at $x = 0$, represented by a gap in the graph of the function; see Figure 1.3.1. Now, take any open interval containing $f(0) = 0$, for example $(-\frac{1}{2}, \frac{1}{2})$, thinking of this interval as being contained in the codomain. The inverse image of this interval is

$$f^{-1}((-\frac{1}{2}, \frac{1}{2})) = (-\frac{1}{2}, 0].$$

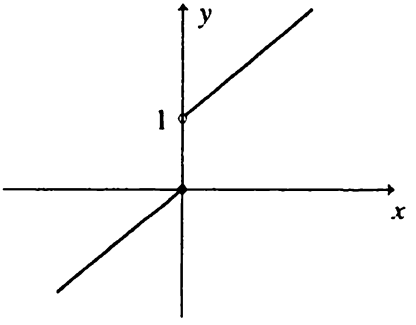


Figure 1.3.1

Observe that the inverse image of an open interval is not open. On the other hand, if we take any open interval in $\mathbb{R}$ which does not contain $f(0)$ (note that $x = 0$ is the only point of non-continuity of the function), then its inverse image is in fact an open interval. Continuity or lack thereof thus appears to be detected by looking at the openness or non-openness of inverse images of open intervals; we take this observation as the basis for the following definition. We cannot "prove" that the following definition corresponds exactly to our intuition, since we cannot prove intuitive things rigorously. The best one can hope for is that all desired intuitively reasonable properties hold, and all examples work out as expected; such is the case for the following definition.

Definition. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. The map f is **continuous** if for every open subset $U \subset B$, the set $f^{-1}(U)$ is open in A. $\Diamond$

In the above definition it is only required that if a set U is open then $f^{-1}(U)$ is open; it is not required that whenever $f^{-1}(U)$ is open that U be open.

Example 1.3.1. (1) Let $A \subset \mathbb{R}^n$ be a set, and let $f: A \to \mathbb{R}^m$ be the constant map given by $f(x) = c$ for all $x \in A$, where c is some point in $\mathbb{R}^m$. If $W \subset \mathbb{R}^m$ is any set that contains c then $f^{-1}(W) = A$, and if W does not contain c then $f^{-1}(W) = \emptyset$. Since A and $\emptyset$ are both open in A, we see that the inverse image of any open subset of the codomain is open in the domain; hence the constant map is continuous. (Note, however, that inverse images of non-open subsets of the codomain are also seen to be open in the domain, and so even for a continuous map the openness of the inverse image of a subset of the codomain does not imply that the subset itself is open.

(2) Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets. The projection maps $\pi_1: A \times B \to A$ and $\pi_2: A \times B \to B$ are both continuous maps. Let $U \subset A$ be an open set. Then $(\pi_1)^{-1}(U) = U \times B$, and Lemma 1.2.9 (i) implies that this latter set is open in $A \times B$. Hence π_1 is continuous. The other case is similar. $\Diamond$

The following lemma gives a useful variant on the definition of continuity.

Lemma 1.3.2. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. The map f is continuous iff for every closed subset $C \subset B$, the set $f^{-1}(C)$ is closed in A.*

Proof. First assume f is continuous. Let $C \subset B$ be closed. Then $B - C$ is open in B, and so $f^{-1}(B - C)$ is open by hypothesis. Using standard properties of inverse images, we have

$$f^{-1}(B - C) = f^{-1}(B) - f^{-1}(C) = A - f^{-1}(C).$$

Since $A - f^{-1}(C)$ is open, it follows that $f^{-1}(C)$ is closed. We have thus proved one of the implications in the lemma. The other implication is proved similarly. $\square$

We now show that the above definition of continuity in terms of open sets is equivalent to the ϵ-δ definition from real analysis, given in part (3) of the following proposition.

Proposition 1.3.3. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. The following statements are equivalent.*

(1) *The map f is continuous.*

(2) *For every point $p \in A$, and every open subset $U \subset B$ containing $f(p)$, there is an open subset $V \subset A$ containing p such that $f(V) \subset U$.*

(3) *For every point $p \in A$ and every number $\epsilon > 0$, there is a number $\delta > 0$ such that if $x \in A$ and $\|x - p\| < \delta$ then $\|f(x) - f(p)\| < \epsilon$.*

Proof. Statement (3) is equivalent to the following statement:

(3′) For every point $p \in A$ and every number $\epsilon > 0$, there is a number $\delta > 0$ such that

$$f(O_\delta(p, A)) \subset O_\epsilon(f(p), B).$$

We now prove (1) $\Rightarrow$ (2) $\Rightarrow$ (3′) $\Rightarrow$ (1).

(1) $\Rightarrow$ (2). Let $p \in A$ and $U \subset B$ containing $f(p)$ be given. By assumption, the map f is continuous, so that $f^{-1}(U)$ is an open subset of A. Observe that $p \in f^{-1}(U)$. By the definition of openness there is thus some open ball of the form $O_\delta(p, A)$ contained in $f^{-1}(U)$. It follows that $f(O_\delta(p, A)) \subset U$. By Exercise 1.2.5 the open ball $O_\delta(p, A)$ is an open subset of A, so let $V = O_\delta(p, A)$.

(2) $\Rightarrow$ (3′). Let $p \in A$ and $\epsilon > 0$ be given. By Exercise 1.2.5 the open ball $O_\epsilon(f(p), B)$ is an open subset of B. By assumption, there is an open subset $V \subset A$ containing p such that $f(V) \subset O_\epsilon(f(p), B)$. By the definition of openness there is some open ball of the form $O_\delta(p, A)$ contained in V. It follows that $f(O_\delta(p, A)) \subset O_\epsilon(f(p), B)$.

(3′) $\Rightarrow$ (1). Let $U \subset B$ be an open subset; we need to show that $f^{-1}(U)$ is open in A. Let $p \in f^{-1}(U)$ be any point; observe that $f(p) \in U$. Since U is open there is an open ball of the form $O_\epsilon(f(p), B)$ contained in U. By hypothesis there is a number $\delta > 0$ such that $f(O_\delta(p, A)) \subset O_\epsilon(f(p), B)$. It follows that $f(O_\delta(p, A)) \subset U$, and thus $O_\delta(p, A) \subset f^{-1}(U)$. It follows that $f^{-1}(U)$ is open in A. □

Example 1.3.4. Let m and b be real numbers such that $m \neq 0$. We will use condition (3) of Proposition 1.3.3 to show that the map $f: \mathbb{R} \to \mathbb{R}$ defined by $f(x) = mx + b$ is continuous. For each $p \in \mathbb{R}$ and each number $\epsilon > 0$ we need to find a number $\delta > 0$ such that if $x \in \mathbb{R}$ is a number and $|x - p| < \delta$, then $|f(x) - f(p)| < \epsilon$. For given p and ϵ we choose δ to be $\delta = \frac{\epsilon}{|m|}$, in which

case $|x - p| < \delta$ implies

$$|f(x) - f(p)| = |(mx + b) - (mp + b)| = |m|\,|x - p|$$
$$< |m|\,\delta = |m|\,\frac{\epsilon}{|m|} = \epsilon.$$

This proves the continuity of f. $\Diamond$

The most important method of combining functions is by composition. The following lemma shows that continuity behaves nicely with respect to composition.

Lemma 1.3.5. *Let $A \subset \mathbb{R}^n$, $B \subset \mathbb{R}^m$ and $C \subset \mathbb{R}^p$ be sets, and let $f: A \to B$ and $g: B \to C$ be continuous maps. The composition $g \circ f$ is continuous.*

Proof. Let $U \subset C$ be an open set. Then $g^{-1}(U)$ is an open subset of B, and hence $f^{-1}(g^{-1}(U))$ is an open subset of A. By a standard result concerning inverse images, we know that $(g \circ f)^{-1}(U) = f^{-1}(g^{-1}(U))$, and the lemma follows. $\square$

Suppose we have a function $f: A \to B$, and we have A broken down as a union $A = A_1 \cup A_2$, where A_1 and A_2 might or might not be disjoint. Suppose we know further that $f|A_1$ and $f|A_2$ are both continuous; can we conclude that f is continuous? The answer is no, as seen using the function given by Equation 1.3.1. As mentioned, this function is not continuous. However, we can write $\mathbb{R} = (-\infty, 0] \cup (0, \infty)$, and certainly $f|(-\infty, 0]$ and $f|(0, \infty)$ are continuous. Fortunately, as seen in the following lemma, we can rule out such annoying examples by putting some restrictions on the sets A_1 and A_2.

Lemma 1.3.6. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. Suppose that $A = A_1 \cup A_2$, and $f|A_1$ and $f|A_2$ are both continuous. If A_1 and A_2 are both open subsets of A or both closed subsets of A, then f is continuous.*

Proof. Suppose that both A_1 and A_2 are open subsets of A. Let $U \subset B$ be an open set. Then $f^{-1}(U) = (f|A_1)^{-1}(U) \cup (f|A_2)^{-1}(U)$. The set $(f|A_1)^{-1}(U)$ is an open subset of A_1, and since A_1 is open in A it follows from Lemma 1.2.6 that $(f|A_1)^{-1}(U)$ is open in A. Similarly for $(f|A_2)^{-1}(U)$. It now follows easily that $f^{-1}(U)$ is open, and this suffices to prove that f is continuous. The case where both A_1 and A_2 are closed is similar, using Exercise 1.2.13. $\square$

The following corollary is easily deducible form Lemma 1.3.6, and we omit the proof.

Corollary 1.3.7. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$, and suppose that $A = A_1 \cup A_2$, where A_1 and A_2 are both open subsets of A or both closed subsets of A. Suppose further that a function $f: A \to B$ is defined in cases by*

$$f(x) = \begin{cases} f_1(x), & \text{if } x \in A_1; \\ f_2(x), & \text{if } x \in A_2, \end{cases}$$

where $f_1: A_1 \to B$ and $f_2: A_2 \to B$ are continuous functions, and $f_1(x) = f_2(x)$ for all $x \in A_1 \cap A_2$. Then f is continuous.

The following lemma shows that the matter of the continuity of a map into a product space works out as nicely as possible. If $A_1, \ldots, A_k$ are subsets of Euclidean space, let $\pi_i: A_1 \times \cdots \times A_k \to A_i$ be the projection map for each $i \in \{1, \ldots, p\}$.

Lemma 1.3.8. *Let $A, A_1, \ldots, A_k$ be subsets of Euclidean space, and let*

$$f: A \to A_1 \times \cdots \times A_k$$

be a function. Let $f_i: A \to A_i$ be defined by $f_i = f \circ \pi_i$ for each $i \in \{1, \ldots, p\}$. Then f is continuous iff all the functions f_i are continuous.

Proof. Suppose that f is continuous. Since we know that the projection maps $\pi_i: A_1 \times \cdots \times A_k \to A_i$ are continuous (as seen in Example 1.3.1, which works with any number of factors), it follows from Lemma 1.3.5 that the functions $f_i = \pi_i \circ f$ are continuous. Now suppose that the functions f_i are continuous. We will show that f is continuous by showing that condition (2) of Proposition 1.3.3 holds. Let $p \in A$ be a point and let $U \subset A_1 \times \cdots \times A_p$ be an open set containing $f(p) = (f_1(p), \ldots, f_k(p))$. Applying Lemma 1.2.9 (which works for any number of factors) we deduce that there are numbers $\epsilon_1, \ldots, \epsilon_k > 0$ such that $O_{\epsilon_1}(f_1(p), A_1) \times \cdots \times O_{\epsilon_k}(f_k(p), A_k) \subset U$. See Figure 1.3.2. Let

$$V = f^{-1}(O_{\epsilon_1}(f_1(p), A_1) \times \cdots \times O_{\epsilon_k}(f_k(p), A_k)).$$

By a standard property of inverse images of maps, we see that

$$V = \bigcap_{i=1}^{k} (f_i)^{-1}(O_{\epsilon_i}(f_i(p), A_i)).$$

By hypothesis on the maps f_i, the sets $(f_i)^{-1}(O_{\epsilon_i}(f_i(p), A_i))$ are open in A. It follows from Lemma 1.2.7 that V is open in A. It is straightforward to see that V works as required in condition (2) of Proposition 1.3.3. □

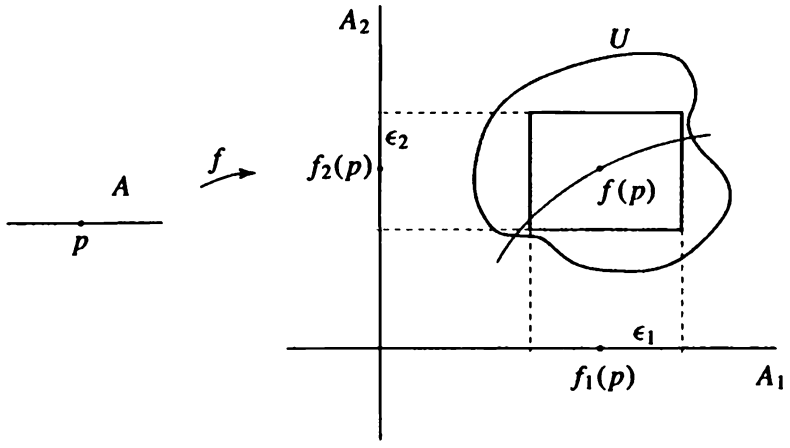


Figure 1.3.2

We take this opportunity to mention another type of map which we will need later on, though the concept is not nearly as important as continuity.

Definition. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. The map f is an **open map** if for every open subset $U \subset A$, the set $f(U)$ is open in B. The map f is a **closed map** if for every closed subset $C \subset A$, the set $f(C)$ is closed in B. $\Diamond$

As seen in Exercise 1.3.11, there exist maps that are any given combination of continuous or not, open or not, and closed or not.

Exercises

1.3.1*. Let $B \subset A \subset \mathbb{R}^n$ be sets. Show that the inclusion map $i: B \to A$ is continuous.

1.3.2. Find two discontinuous functions $f: A \to B$ and $g: B \to C$ such that the composition $g \circ f$ is continuous. (Thus the converse of the above Lemma 1.3.5 does not hold.)

1.3.3*. Let $B \subset A \subset \mathbb{R}^n$ and $C \subset \mathbb{R}^m$ be sets, and let $f: A \to C$ be a continuous map. Show that $f|B: B \to C$ is continuous.

1.3.4. Let $B \subset A \subset \mathbb{R}^n$ be sets. Show that if B is an open (respectively, closed) subset of A, then the inclusion map $i: B \to A$ is an open (respectively, closed) map.

1.3.5*. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a bijective map. Prove that f is open iff f is closed. (Note that bijectivity is crucial, since Exercise 1.3.11 shows that a non-bijective map can be open but not closed or vice-versa.)

1.3.6. Let $U \subset \mathbb{R}^{n+m} = \mathbb{R}^n \times \mathbb{R}^m$ be an open set. Show that the projection maps $\pi_1: U \to \mathbb{R}^n$ and $\pi_2: U \to \mathbb{R}^m$ are open maps.

1.3.7*. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets. A map $f: A \to B$ is called **uniformly continuous** if for every number $\epsilon > 0$ there exists a number $\delta > 0$ such that if $x, y \in A$ are any two points then $\|x - y\| < \delta$ implies $\|f(x) - f(y)\| < \epsilon$. (The point of uniform continuity is that the δ only depends upon ϵ, not upon the particular points of A.) Find an example of a continuous function that is not uniformly continuous. (See [BT, §16] for a solution.)

1.3.8. Let $f, g: \mathbb{R} \to \mathbb{R}$ be continuous functions. Define the functions $\max\{f, g\}$ and $\min\{f, g\}$ by setting

$$\max\{f, g\}(x) = \begin{cases} f(x), & \text{if } f(x) \geq g(x) \\ g(x), & \text{if } g(x) \geq f(x), \end{cases}$$

$$\min\{f, g\}(x) = \begin{cases} f(x), & \text{if } f(x) \leq g(x) \\ g(x), & \text{if } g(x) \leq f(x). \end{cases}$$

Prove that $\max\{f, g\}$ and $\min\{f, g\}$ are continuous.

1.3.9*. Let a be a non-zero real number. Show that the map $f: \mathbb{R} - \{0\} \to \mathbb{R} - \{0\}$ defined by $f(x) = \frac{a}{x}$ is continuous. (Use the ϵ-δ definition of continuity. See [SK2, Chapter 6] for a solution.)

1.3.10*. Let $F: \mathbb{R}^n \to \mathbb{R}^m$ be an affine linear map (as defined in the Appendix). Show that F is continuous. (Use the ϵ-δ definition of continuity.)

1.3.11. A map from one subset of Euclidean space to another can be any combination of continuous or not, open or not, and closed or not (there are eight such possibilities). Find one map for each of the eight types.

1.4. Homeomorphisms and Quotient Maps

Just as two vector spaces are considered virtually the same from the point of view of linear algebra if there is a linear isomorphism from one to the other, we now define the corresponding type of map between subsets of Euclidean space, the existence of which will imply that two sets are virtually the same from the point of view of topology.

Definition. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. The map f is a **homeomorphism** if it is bijective and both it and its inverse are continuous. If f is a homeomorphism, we say that A and B are **homeomorphic**, and we write $A \approx B$. $\Diamond$

A few remarks on homeomorphisms are needed. First, if two spaces are homeomorphic, there may be many homeomorphisms between the spaces. For example, the two maps $f, g: (-1, 1) \to (-2, 2)$ given by $f(x) = 2x$ and $g(x) = -2x$ are both homeomorphisms. Second, the relation of "homeomorphic" is seen to be an equivalence relation on the collection of all subsets of Euclidean spaces. Third, a reader who has seen abstract algebra should be careful not to confuse the similar sounding words "homeomorphism" and "homomorphism." *Homeomorphisms* are to topological spaces what isomorphisms are to groups; *homomorphisms* play the analogous role for groups as continuous maps do for topology.

Example 1.4.1. Any open interval (a, b) in $\mathbb{R}$ is homeomorphic to $\mathbb{R}$. We construct the desired homeomorphism in two stages, first constructing a homeomorphism $f: (a, b) \to (-\frac{\pi}{2}, \frac{\pi}{2})$, and then constructing a homeomorphism $g: (-\frac{\pi}{2}, \frac{\pi}{2}) \to \mathbb{R}$; the composition $g \circ f$ will be the desired homeomorphism $(a, b) \to \mathbb{R}$. The map f is given by the formula

$$f(x) = \frac{\pi}{b-a}x - \frac{\pi(b+a)}{2(b-a)}.$$

This map is continuous by Example 1.3.4. It is left to the reader to verify that f has an inverse, and that the inverse is also continuous. The map g is given by $g(x) = \tan x$. That g and its inverse are continuous (on the given domain) is also standard. $\Diamond$

Recall that a linear isomorphism of vector spaces is defined to be a linear map that is bijective. Any bijective map has an inverse map simply as a map

of one set to another; it can be proved that if a linear map is bijective then its inverse — which is not a priori linear — will in fact be a linear map. Does the analog hold for continuous maps? That is, if a continuous map is bijective, is the inverse map necessarily continuous? If the answer were yes, then the definition of homeomorphism would be redundant. The following example shows that the answer is no; thus continuous maps do not behave as nicely as linear maps.

Example 1.4.2. Let $g: [0, 1] \cup (2, 3] \to [0, 2]$ be defined by

$$g(x) = \begin{cases} x, & \text{if } x \in [0, 1]; \\ x - 1, & \text{if } x \in (2, 3]. \end{cases}$$

The map g slides the interval $(2, 3]$ to the left one unit. It is easy to verify that g is bijective and continuous. However, we can show that the the inverse map $g^{-1}: [0, 2] \to [0, 1] \cup (2, 3]$ is not continuous. Using Lemma 1.2.5 it can be verified that the set $U = (0, 1]$ is an open subset of the set $[0, 1] \cup (2, 3]$. However, the set $(g^{-1})^{-1}(U) = (0, 1]$ is not an open subset of $[0, 2]$. Hence g^{-1} is not a continuous map. $\Diamond$

The following lemma gives a useful characterization of homeomorphisms.

Lemma 1.4.3. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a map. Then f is a homeomorphism iff f is bijective and for every subset $U \subset B$, the set U is open in B iff $f^{-1}(U)$ is open in A.*

Proof. Exercise 1.4.2. $\square$

In addition to homeomorphisms we need to introduce another type of map, inspired by the idea of gluing things together. To make a cylinder out of a piece of paper, we cut out a rectangular strip and then glue two of the opposing sides together. See Figure 1.4.1. Although in practice one would probably have the two sides overlap a little bit if one were using glue, let us assume that the edges are glued together with no overlap (as could be done with tape).

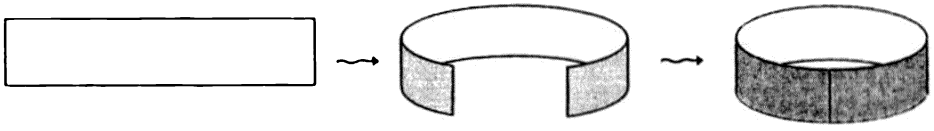


Figure 1.4.1

In the example of the cylinder each point in one of the vertical edges of the original rectangle is glued to a point on the other vertical edge, and all other points in the rectangle are left alone. One way of thinking about the relation of the cylinder to the rectangle is that corresponding pairs of points on the vertical edges in the rectangle become transformed into single points in the cylinder; points in the rectangle not in the vertical edges stay single points in the cylinder. We can thus view the gluing process as breaking up the rectangle into subcollections of points, each subcollection of points being collapsed to one point as the result of the gluing process. The following definition gives the most general way possible to break up a set into a collection of disjoint subsets.

Definition. Let $X \subset \mathbb{R}^n$ be a set. A **partition** of X is a collection $\mathcal{P} = \{A_i\}_{i \in I}$ such that $\bigcup_{i \in I} A_i = X$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$. $\Diamond$

Given a set $X \subset \mathbb{R}^n$ and a partition $\mathcal{P} = \{A_i\}_{i \in I}$ of X, we are looking for a set $Y \subset \mathbb{R}^m$ that is intuitively the result of collapsing each set A_i in the partition to a single point. We want Y to be a set such that there exists a surjective map $q: X \to Y$ with the property that if $a, b \in X$ are points, then $q(a) = q(b)$ iff a and b belong to the same set A_i; equivalently, the collection of sets of the form $q^{-1}(y)$ is the same as the original partition $\mathcal{P}$. Although this requirement on the map q is certainly necessary, it is unfortunately not sufficient. The problem, as seen in the following example, is essentially the same as that encountered in Example 1.4.2.

Example 1.4.4. Let $X = [0, 1] \cup (2, 4]$, and let $\mathcal{P}$ be the partition of X containing the set $A = [3, 4]$, with every other set in $\mathcal{P}$ a one-element set. A logical choice for a space Y and a map $q: X \to Y$ as above would be $Y = [0, 1] \cup (2, 3]$ and

$$q(x) = \begin{cases} x, & \text{if } x \in [0, 1] \cup (2, 3]; \\ 3, & \text{if } x \in [3, 4]. \end{cases}$$

It is straightforward to see that $\{q^{-1}(y) \mid y \in Y\} = \mathcal{P}$. On the other hand, let $Y_1 = [0, 2]$ and let $q_1: X \to Y_1$ be given by

$$q_1(x) = \begin{cases} x, & \text{if } x \in [0, 1]; \\ x - 1, & \text{if } x \in (2, 3]; \\ 2, & \text{if } x \in [3, 4]. \end{cases}$$

The map q_1 also has the property that $\{q_1^{-1}(y) \mid y \in Y\} = \mathcal{P}$. The set Y_1 is, however, not what we would like to call the result of collapsing the partition $\mathcal{P}$. $\Diamond$

To rule out maps such as q_1 in the above example, we use Lemma 1.4.3 as inspiration. Observe that the maps under consideration are not injective, so we cannot use the definition of homeomorphisms as a guide, since a non-injective map has no inverse.

Definition. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets. A map $q: A \to B$ is a **quotient map** if f is surjective and if, for all subsets $U \subset B$, the set U is open in B iff $q^{-1}(U)$ is open in A. If $X \subset \mathbb{R}^n$ is a set and $\mathcal{P}$ is a partition of X, a set $Y \subset \mathbb{R}^m$ is an **identification space** of X and $\mathcal{P}$ if there is a quotient map $q: X \to Y$ such that $\{q^{-1}(y) \mid y \in Y\} = \mathcal{P}$. $\Diamond$

Observe that quotient maps are automatically continuous. Not every continuous surjection is a quotient map, as can be seen from Example 1.4.2.

It is not at all evident from the above definition that for any set $X \subset \mathbb{R}^n$ and for any partition $\mathcal{P}$ of X there is an identification space of X and $\mathcal{P}$. In the general setting of arbitrary topological spaces it can be shown that identification spaces always exist, but these spaces are abstractly defined and it is not always clear whether such a space can be found sitting in some Euclidean space. We do not tackle this question in general, though we do prove in Chapter II that identification spaces do exist in the particular cases we will use to construct surfaces. The following lemma says that identification spaces are uniquely determined if they exist.

Lemma 1.4.5. *Let $X \subset \mathbb{R}^n$ be a set and let $\mathcal{P}$ be a partition of X. If $Y \subset \mathbb{R}^m$ and $Z \subset \mathbb{R}^p$ are identification spaces of X and $\mathcal{P}$, then $Y \approx Z$.*

Proof. Let $q: X \to Y$ and $r: X \to Z$ be quotient maps such that

$$\{q^{-1}(y) \mid y \in Y\} = \mathcal{P} = \{r^{-1}(z) \mid z \in Z\}.$$

Define a map $h: Y \to Z$ as follows. For each $y \in Y$, the set $q^{-1}(y)$ equals some set in $\mathcal{P}$, and this set in $\mathcal{P}$ also equals $r^{-1}(z)$ for some unique $z \in Z$; define $h(y) = z$. It is straightforward to see that $h \circ q = r$. Since r is continuous and q is a quotient map it follows from Exercise 1.4.5 that h is continuous. A similar construction with the roles of Y and Z reversed can be used to construct the analogous map $g: Z \to Y$, which is continuous and, as can be verified, is the inverse map of h. Thus h is a homeomorphism. $\square$

Example 1.4.6. Let $X = [0, 1]$ and let $\mathcal{P}$ be the partition of X containing the set $\{0, 1\}$, with all other members of $\mathcal{P}$ single-element sets. Then the identification

space of X and $\mathcal{P}$ is the circle S^1, which can be seen using the quotient map $q: [0, 1] \to S^1$ given by

$$q(x) = \begin{pmatrix} \cos 2\pi x \\ \sin 2\pi x \end{pmatrix}.$$

That q is a quotient map could be verified directly, though we will take the easy route and use a more general result given in Proposition 1.6.14. Intuitively q simply takes the interval $[0, 1]$ and glues the endpoints together. It is straightforward to verify that the sets $q^{-1}(y)$ are precisely the sets in $\mathcal{P}$. $\Diamond$

One important way of producing identification spaces is to attach two sets along homeomorphic subspaces; for example, we might wish to make a sphere out of cloth by taking two pieces of cloth shaped like unit disks and sewing them to one another along their boundaries. See Figure 1.4.2.

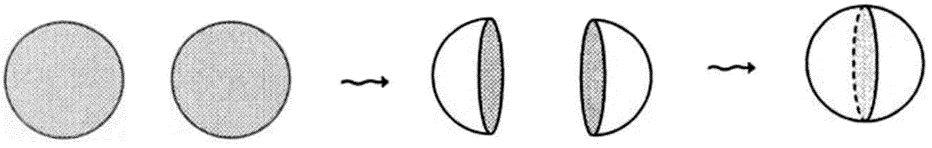


Figure 1.4.2

Definition. Let $X, Y \subset \mathbb{R}^n$ be disjoint sets. Suppose that $X' \subset X$ and $Y' \subset Y$ are sets, and $h: X' \to Y'$ is a homeomorphism. Define a partition $\mathcal{P}(h)$ on $X \cup Y$ to be the collection of all pairs $\{x, h(x)\}$ for $x \in X'$, and all single-element sets $\{z\}$ for $z \in (X - X') \cup (Y - Y')$. A set $W \subset \mathbb{R}^m$ is the result of **attaching** X and Y via the map h, denoted $X \cup_h Y$, if W is an identification space of $X \cup Y$ and $\mathcal{P}(h)$. $\Diamond$

As with identification spaces in general, it is not at all evident that for any X, Y and h as in the above definition there is an attaching space $X \cup_h Y$; again, we will prove in Section 2.6 that identification spaces do exist in the particular case we will be using to construct surfaces.

Example 1.4.7. Let $X = [-2, -1]$ and $Y = [1, 2]$, let $X' = \{-2, -1\}$ and $Y' = \{1, 2\}$, and let $h: X' \to Y'$ be defined by $h(-2) = 2$ and $h(-1) = 1$. Then the circle S^1 is an attaching space $X \cup_h Y$. One can construct the appropriate quotient map from $X \cup Y$ to S^1 by mapping X to the lower half-circle of S^1 and

Y to the upper half-circle of S^1, by a map such that -2 and 2 are sent to $\binom{-1}{0}$ and -1 and 1 are sent to $\binom{1}{0}$. $\Diamond$

Exercises

1.4.1. Show that any open disk in $\mathbb{R}^2$ is homeomorphic to $\mathbb{R}^2$. Show that any open rectangle $(a, b) \times (c, d)$ is homeomorphic to $\mathbb{R}^2$.

1.4.2*. Prove Lemma 1.4.3.

1.4.3. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a continuous bijection. Show that f is a homeomorphism iff it is an open map iff it is a closed map.

1.4.4*. Let $f: A \to B$ be a continuous bijection such that for every $a \in A$ there is an open subset $U \subset A$ containing a such that $f(U)$ is open in B and $f|U$ is a homeomorphism from U onto $f(U)$. Show that f is a homeomorphism.

1.4.5*. Let X, Y and Z be subsets of Euclidean space, and let $f: X \to Y$ and $g: Y \to Z$ be maps. Suppose that f is a quotient map. Show that g is continuous iff $g \circ f$ is continuous.

1.4.6. Find the identification space in each of the following cases (the result will be a familiar object).

(1) Let X be the unit disk in $\mathbb{R}^2$, and let $\mathcal{A}$ be the partition of X containing the unit circle as one member, with all other members of $\mathcal{A}$ single-element sets.

(2) Let $Y = \mathbb{R}^2$ and let $\mathcal{B}$ be the partition of Y containing the closed unit disk as one member, with all other members of $\mathcal{B}$ single-element sets.

(3) Let $Z = \mathbb{R}$, and let $\mathcal{C}$ be the partition of Z into subsets of the form $x + \mathbb{Z}$ for $x \in \mathbb{R}$.

1.4.7. Find the result of attaching in each of the following cases (the result will be a familiar object).

(1) Let $X = [0, 2]$ and $Y = [3, 5]$, let $X' = [1, 2]$ and $Y' = [3, 4]$, and let $h: X' \to Y'$ be defined by $h(x) = x + 2$.

(2) Let $X = O_1(\binom{1}{0}, \mathbb{R}^2)$ and $Y = O_1(\binom{4}{0}, \mathbb{R}^2)$, let $X' = \binom{4}{0}$ and $Y' = \binom{4}{0}$, and let $h: X' \to Y'$ be defined by $h(X') = Y'$.

1.4.8*. This exercise generalizes the concept of attaching via a homeomorphism; it may seem unlikely, but it will be of use in Section 2.4. Let $X \subset \mathbb{R}^n$ be a set, let $A_1, B_1, \ldots, A_p, B_p \subset X$ be sets (not necessarily disjoint) and let $h_i: A_i \to B_i$ be a homeomorphism for each $i \in \{1, \ldots, p\}$. For convenience let $h_0 = 1_X$. For each point $x \in X$ let $[x]$ be the subset of X defined by

$$[x] = \{y \in X \mid y = h_{i_1}^{\pm 1} \circ \cdots h_{i_r}^{\pm 1}(x) \text{ for some } i_1, \ldots i_r \in \{1, \ldots, p\}\}.$$

Show that the collection of sets $[x]$ for all $x \in X$ form a partition of X. This partition will be denoted $\mathcal{P}(h_1, \ldots, h_p)$.

1.4.9*. For each $i = 1, 2$, let $X_i, Y_i \subset \mathbb{R}^n$ be disjoint sets, let $X_i' \subset X_i$ and $Y_i' \subset Y_i$ be sets and let $h_i: X_i' \to Y_i'$ be homeomorphisms. Suppose that $X_1 \cup_{h_1} Y_1$ exists. Suppose further that there exist homeomorphisms $f: X_2 \to X_1$ and $g: Y_2 \to Y_1$ such that $f(X_2') = X_1'$, $g(Y_2') = Y_1'$ and $h_1 \circ f|X_2' = g|Y_2' \circ h_2$; this last condition is expressed by the commutativity of the following diagram.

$$\begin{array}{ccc} X_2' & \xrightarrow{h_2} & Y_2' \\ {\scriptstyle f}\downarrow & & \downarrow{\scriptstyle g} \\ X_1' & \xrightarrow{h_1} & Y_1' \end{array}$$

Show that $X_2 \cup_{h_2} Y_2$ exists and is homeomorphic to $X_1 \cup_{h_1} Y_1$.

1.5. Connectedness

Intuitively, a subset of Euclidean space is connected if it is made up of "one piece." The following definition nicely captures this notion.

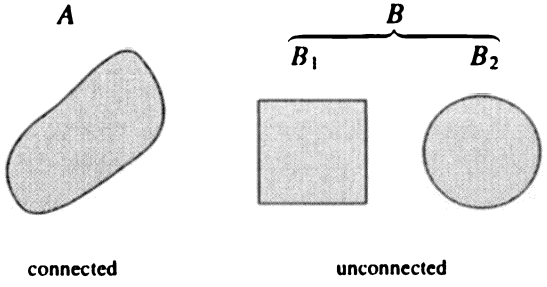


Figure 1.5.1

Definition. Let $A \subset \mathbb{R}^n$ be a set. We say that A is **connected** if it cannot be expressed as the union of two non-empty disjoint subsets each of which is open in A. If A can be expressed as the union of two non-empty disjoint open subsets, we say A is **disconnected**. $\Diamond$

The requirement for both subsets being open and non-empty in the definition of connectedness is absolutely crucial. For example, the interval $[0, 2]$ is the union of the disjoint subsets $[0, 1]$, and $(1, 2]$, the first of which is closed (though not open) in $[0, 2]$ and the second of which is open in $[0, 2]$, and yet the set $[0, 2]$ is certainly intuitively connected. (We will see shortly that $[0, 2]$ is indeed connected by our definition.) The following lemma gives some alternate characterizations of connectedness.

Lemma 1.5.1. *Let $A \subset \mathbb{R}^n$ be a set. The following are equivalent:*

(1) *A is connected;*
(2) *A cannot be expressed as the union of two non-empty disjoint subsets each of which is closed in A;*
(3) *the only subsets of A that are both open and closed in A are $\emptyset$ and A.*

Proof. Exercise 1.5.1. $\square$

The following theorem shows that not only are intervals in $\mathbb{R}$ connected, but that they are the only connected subsets of $\mathbb{R}$. The proof of this theorem makes crucial use of the Least Upper Bound Property of the real numbers; consult [HM, p. 38], or most introductory real analysis texts, for a discussion of this property. We note that an interval in the rational numbers is not connected, and it is thus necessary to use a property of the real numbers that does not hold for the rationals, of the which the Least Upper Bound Property is an example.

Theorem 1.5.2. *A non-empty subset of $\mathbb{R}$ is connected iff it is an interval (of any sort).*

Proof. First suppose that $J \subset \mathbb{R}$ is an interval. We will assume that J is not connected and derive a contradiction. By assumption, we can write $J = B_1 \cup B_2$, where B_1 and B_2 are non-empty disjoint open subsets of J. Then B_1 and B_2 are also both closed in J. Choose points $b_1 \in B_1$ and $b_2 \in B_2$. Without loss of generality assume that $b_1 < b_2$. Since J is an interval of some sort we know that $[b_1, b_2] \subset J$. Since the set $B \cap [b_1, b_2]$ is bounded above by b_2, the Least Upper Bound Property of the real numbers implies that there is a point w defined by $w = \text{lub}\,\{B_1 \cap [b_1, b_2]\}$.

Is w in B_1 or B_2? On the one hand, since $B_1 \cap [b_1, b_2]$ is closed in $[b_1, b_2]$ and hence in $\mathbb{R}$, it follows from Exercise 1.2.17 that $w \in B_1$. Since b_2 is an upper bound for $B \cap [b_1, b_2]$ it follows that $w \leq b_2$; because $b_2 \notin B_1$ it must be the case that $w < b_2$. It now follows from the definition of w as a least upper bound that $(w, b_2] \subset B_2$. However, since B_2 is closed it can be deduced that $w \in B_2$, a contradiction. Thus J must be connected.

Now suppose that $J \subset \mathbb{R}$ is connected. If J is bounded below let $a = \text{glb } J$, which exists by the Least Upper Bound property of the real numbers; if J is not bounded below let $a = -\infty$. (Infinity is not a real number, and "∞" should be treated as a symbol only.) Similarly, if J is bounded above let $b = \text{lub } J$; if J is not bounded above let $b = \infty$. The points a and b may or may not be contained in J. Note that $J \subset [a, b]$, where we leave the interval open at a or b if they are $-\infty$ or ∞ respectively. We will show that $(a, b) \subset J$, and it will then follow that J is one of (a, b), $[a, b)$, $(a, b]$ or $[a, b]$, depending upon which of a and b are contained in J.

Suppose $(a, b) \not\subset J$, so there is some point $z \in (a, b)$ that is not contained in J. Let $A_1 = (-\infty, z) \cap J$ and $A_2 = (z, \infty) \cap J$. The sets A_1 and A_2 are disjoint open subsets of J, and $A_1 \cup A_2 = J$. Neither A_1 nor A_2 is empty, since A_1 being empty would mean that z is a lower bound for J, in contradiction to the definition of a, and similarly for A_2. Hence J is not connected, a contradiction to our hypothesis on J. Thus $(a, b) \subset J$. □

There is, unfortunately, no analog of Theorem 1.5.2 for higher dimensional Euclidean spaces. The higher dimensional analogs of intervals are rectangular boxes (that is, products of intervals), and these are connected by Exercise 1.5.3, but they are certainly not the only connected subsets of Euclidean space. For example, we will see later on that $\mathbb{R}^n$ with a point removed is connected.

Though not every subset of Euclidean space is connected, every disconnected space is made up of connected pieces. See Figure 1.5.1. The following definition makes this notion precise.

Definition. Let $A \subset \mathbb{R}^n$ be a set. A subset $C \subset A$ is a **component** of A if it is non-empty, connected, and not a proper subset of a connected subset of A. ◊

For example, the set B shown in Figure 1.5.1 has two components. Observe that if a set $A \subset \mathbb{R}^n$ is connected then the only component of A is itself. Some properties of components are given in Exercise 1.5.4.

The following theorem shows that connectedness behaves quite nicely with respect to continuous maps.

Theorem 1.5.3. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a continuous map. If A is connected then so is $f(A)$.*

Proof. Suppose that $f(A)$ is not connected, so we can write $f(A) = V \cup W$, where V and W are non-empty disjoint open subsets of $f(A)$. Then $A = f^{-1}(f(A)) = f^{-1}(V) \cup f^{-1}(W)$ by standard properties of inverse images. The sets $f^{-1}(V)$ and $f^{-1}(W)$ are non-empty disjoint subsets of A, and by the continuity of f they are open subsets of A. Thus A is not connected, a contradiction. □

We can now use our results about connectivity to prove the following two important theorems. The first is familiar from calculus (where it is usually presented without proof); the second is the one-dimensional version of a result that holds in all dimensions (the two-dimensional case of which, a much more difficult result, will be proved in Section 3.6).

Theorem 1.5.4 (Intermediate Value Theorem). *Let $[a, b] \subset \mathbb{R}$ be an interval, and let $f: [a, b] \to \mathbb{R}$ be a continuous map. For any real number z between $f(a)$ and $f(b)$ there is some $c \in [a, b]$ such that $f(c) = z$.*

Proof. By Theorem 1.5.2 the interval $[a, b]$ is connected, by Theorem 1.5.3 the set $f([a, b])$ is connected, and by Theorem 1.5.2 the set $f([a, b])$ is an interval. Since $f(a)$ and $f(b)$ are both contained in $f([a, b])$, it follows that any point in $\mathbb{R}$ between $f(a)$ and $f(b)$ is also contained in $f([a, b])$. The result now follows. □

Theorem 1.5.5 (One-dimensional Brouwer Fixed Point Theorem). *Let $[a, b] \subset \mathbb{R}$ be an interval, and let $f: [a, b] \to [a, b]$ be a continuous map. Then there is a point $d \in [a, b]$ such that $f(d) = d$.*

Proof. Exercise 1.5.6. □

In both Theorems 1.5.4 and 1.5.5, we are only told that some point with certain desired properties exists; we are not told anything additional about these points, neither that they are unique (which they need not be), nor how to find them.

There is another way to approach the issue of whether a subset of Euclidean space is made up of one or more pieces. Intuitively, if a set is made up of one

piece then there should be a path from any one point in the set to any other point in it (very much like drawing something without lifting your pencil from the page).

Definition. Let $A \subset \mathbb{R}^n$ be a set and let $x, y \in A$ be points. A **path** in A from x to y is a continuous map $c\colon [0, 1] \to A$ such that $c(0) = x$ and $c(1) = y$. The set A is **path connected** if for any pair of points $x, y \in A$ there is a path in A from x to y. (Some books use the terms "pathwise connected" or "arcwise connected.") $\Diamond$

Example 1.5.6. The space $\mathbb{R}^n$ is path connected for all n. Between any two points $x, y \in \mathbb{R}^n$ there is, among many paths, the straight line path; more specifically, if

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}, \quad \text{and} \quad y = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix},$$

the straight line path in $\mathbb{R}^n$ from x to y is the map $c\colon [0, 1] \to \mathbb{R}^n$ given by

$$c(t) = t \begin{pmatrix} y_1 - x_1 \\ \vdots \\ y_n - x_n \end{pmatrix} + \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}. \quad \Diamond$$

As might be expected, path connectivity and connectivity are not unrelated.

Proposition 1.5.7. *A path connected subset of Euclidean space is connected.*

Proof. Let $A \subset \mathbb{R}^n$ be a path connected set. Assume that A is not connected. By assumption we can write A as $A = A_1 \cup A_2$, where A_1 and A_2 are non-empty disjoint open subsets of A. Let x be a point in A_1, and let y be a point in A_2. By hypothesis there exists a continuous map $c\colon [0, 1] \to A$ such that $c(0) = x$ and $c(1) = y$. Consider the subset $c([0, 1]) \subset A$. We see that

$$c([0, 1]) = \big(c([0, 1]) \cap A_1\big) \cup \big(c([0, 1]) \cap A_2\big).$$

The sets $c([0, 1]) \cap A_1$ and $c([0, 1]) \cap A_2$ are non-empty disjoint open subsets of $c([0, 1])$, so that $c([0, 1])$ is not connected. On the other hand, Theorems 1.5.2 and 1.5.3 together imply that $c([0, 1])$ is connected, a contradiction. $\square$

Using the proposition just proved and Exercise 1.5.7, we deduce that $\mathbb{R}^n$ with a point removed is connected. Although path connectedness implies connectedness, somewhat surprisingly the reverse implication does not hold in general. The following clever example is a standard one.

Example 1.5.8. Let p denote the point $\binom{0}{1} \in \mathbb{R}^2$. The "deleted comb space" is the subset $K \subset \mathbb{R}^2$ made up of p and a collection of line segments as follows:

$$K = \{p\} \cup ([0, 1] \times \{0\}) \cup \bigcup_{n=1}^{\infty} \{\frac{1}{n}\} \times [0, 1].$$

See Figure 1.5.2. Intuitively it might appear as if the point p were "isolated" in K, though in fact K is connected (but not path connected). Suppose K is not connected. We can thus write $K = A \cup B$, where $A, B \subset K$ are disjoint non-empty open subsets of K. The point p must be in one of A or B, and without loss of generality suppose that it is in A. If p is not the only point in A, then we can write $K - \{p\} = (A - \{p\}) \cup B$, and the sets $A - \{p\}$ and B are disjoint, non-empty, open subsets of $K - \{p\}$. Hence $K - \{p\}$ is not connected, which yields a contradiction, since $K - \{p\}$ is clearly path connected. The only other possibility is that p is the only point in A. The openness of A in K implies that there is some number $\epsilon > 0$ such that $O_\epsilon(p, K)$ is contained in A; if $A = \{p\}$ then $O_\epsilon(p, K) = \{p\}$, which is clearly not true from the construction of K, again a contradiction. Thus K is connected.

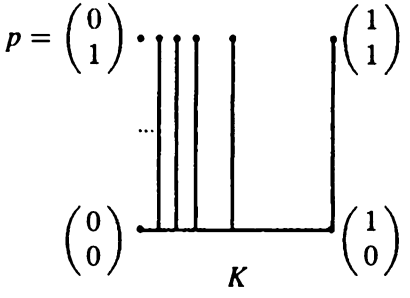


Figure 1.5.2

To see that K is not path connected, suppose otherwise. Then there is a path in K from the point p to $q = \binom{1}{1}$; let c: $[0, 1] \to K$ be such a path. It may or may not be the case that $c([0, 1])$ intersects $[0, 1] \times \{0\}$. Let us first suppose not. Consider the function $\pi_1 \circ c$: $[0, 1] \to \mathbb{R}$, where π_1 is projection from $\mathbb{R}^2$ onto the x-axis. This composition is continuous (since both π_1 and c are continuous), and we have $\pi_1 \circ c(0) = 0$ and $\pi_1 \circ c(1) = 1$. Let r be any irrational number between 0 and 1. By the Intermediate Value Theorem (Theorem 1.5.4) there is some number $z \in (0, 1)$ such that $\pi_1 \circ c(z) = r$. Hence $c(z)$ has an

irrational x-coordinate, a contradiction to the fact that the x-coordinates of all points in $K - ([0, 1] \times \{0\})$, which contains the image of c, are rational by the construction of K.

Next we need to consider the case where the image of c does intersect $[0, 1] \times \{0\}$. The set $c^{-1}([0, 1] \times \{0\})$ is a closed subset of $[0, 1]$ by the continuity of c. It follows from Exercise 1.2.17 that $c^{-1}([0, 1] \times \{0\})$ contains its greatest lower bound, denoted w. Since $c(0) = p$ we see that $w > 0$. If π_2 denotes projection from $\mathbb{R}^2$ onto the y-axis, it follows that $\pi_2 \circ c(0) = 1$ and $\pi_2 \circ c(w) = 0$. By the Intermediate Value Theorem there is a number $d \in (0, w)$ such that $\pi_2 \circ c(d) = \frac{1}{2}$. Using the definition of w we deduce that $c([0, d])$ does not intersect $[0, 1] \times \{0\}$. The same type of reasoning as in the previous paragraph can now be applied to $c|[0, d]$, again yielding a contradiction. Thus K is not path connected. $\Diamond$

Exercises

1.5.1*. Prove Lemma 1.5.1.

1.5.2*. Let $A \subset \mathbb{R}^n$ be a set, and suppose that A can be written as the union $A = \bigcup_{i \in I} A_i$ of connected sets $A_i \subset \mathbb{R}^n$, where the indexing set I is arbitrary. This hypothesis alone does not guarantee that A is connected. Show that if all the sets A_i have at least one point in common, so that $\bigcap_{i \in I} A_i \neq \emptyset$, then A is connected.

1.5.3*. Show that the product of finitely many connected subsets of Euclidean space is connected. Conclude as a corollary that, given intervals $[a_i, b_i] \subset \mathbb{R}$ for $i \in \{1, \ldots, n\}$, the box $[a_1, b_1] \times \cdots \times [a_n, b_n] \subset \mathbb{R}^n$ is connected.

1.5.4*. Let $A \subset \mathbb{R}^n$ be a set. Show that the components of A are disjoint closed subsets of A and that A is the union of its components. If A has finitely many components, show that the components are open subsets of A; give an example showing that components are not necessarily open subsets in general.

1.5.5. Is the property of being disconnected preserved by continuous maps?

1.5.6*. Prove Theorem 1.5.5.

1.5.7*. Show that the following sets are path connected, and hence connected:

(1) any open ball in $\mathbb{R}^n$, and any closed ball in $\mathbb{R}^n$;

(2) any open ball in $\mathbb{R}^n$ from which a point has been removed, when $n \geq 2$;

(3) the unit circle $S^1 \subset \mathbb{R}^2$.

1.5.8. Let $U \subset \mathbb{R}^2$ be an open, path connected set, and let $x \in U$ be a point. Show that $U - \{x\}$ is path connected. Find an example to show that the hypothesis of openness cannot be dropped.

1.5.9*. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a continuous map. Show that if A is path connected then so is $f(A)$.

1.5.10*. Let $A \subset \mathbb{R}^n$ be a set, and let $x, y, z \in A$ be points. Prove the following three properties of paths (if you are familiar with equivalence relations these properties should look familiar). Recall that, as we have defined them, paths always have domain $[0, 1]$.

(i) There is a path from x to itself.

(ii) If there is a path from x to y then there is a path from y to x.

(iii) If there is a path from x to y and a path from y to z, then there is a path from x to z.

1.5.11*. Let $A \subset \mathbb{R}^n$ be a set, and let $a \in A$ be a point. Suppose that $U \subset A$ is an open subset of A containing a such that U is path connected but $U - \{a\}$ is not path connected. Show that if $V \subset U$ is an open subset of A containing a, then $V - \{a\}$ is not path connected.

1.5.12*. Let $A \subset \mathbb{R}^n$ be a set, and let $C \subset A$ be a connected subset. Show that C is contained in a single component of A.

1.5.13*. Let $A \subset \mathbb{R}^n$ be a set. If $B \subset A$ is a subset that is both closed and open in A, show that B is the union of components of A.

1.6. Compactness

The concept of compactness, crucial in our treatment of surfaces, is less intuitively appealing than connectedness. One way of viewing compactness is as a generalization of the notion of finiteness to the topological setting, where the sets under consideration almost always have infinitely many points, but where we can define a notion of finiteness nonetheless.

To state the definition of compactness, we start with the following tools.

Definition. Let $A \subset \mathbb{R}^n$ be a set. A **cover** of A is a collection of subsets of A whose union is all of A (such a collection may be finite or infinite). If $\mathcal{U} = \{U_i\}_{i \in I}$ is a cover of A, a **subcover** of $\mathcal{U}$ is a subcollection of the sets in $\mathcal{U}$ that is itself a cover of A (any subcover is of the form $\{U_j\}_{j \in J}$ for a subset $J \subset I$). A **finite cover** is a cover of A with finitely many sets. An **open cover** of A is a cover of A such that all the sets in the cover are open subsets of A. $\Diamond$

Intuitively, if a set has an open cover with no finite subcover, then the set has something "topologically infinite" about it. The following definition, reflecting this observation, should be read with care.

Definition. Let $A \subset \mathbb{R}^n$. We say that A is **compact** if every open cover of A has a finite subcover. $\Diamond$

To show that a set is compact it is not sufficient to find some open cover of the set that has a finite subcover. It has to be shown that *any* open cover has a finite subcover, which of course is much harder since one cannot usually write down explicitly all the possible open covers of the set. Proving that a set is not compact, on the other hand, is sometimes easier since it suffices to find one open cover for which there is no finite subcover.

Example 1.6.1. (1) Any finite set of points in Euclidean space is compact. Let $A = \{p_1, \ldots, p_m\} \subset \mathbb{R}^n$, and let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of A. For each $k \in \{1, 2, \ldots, m\}$, the point p_k is contained in at least one set in $\mathcal{U}$, say U_{i_k}. Then $\bigcup_{k=1}^{m} U_{i_k} = A$, and so $\mathcal{U}$ has a finite subcover.

(2) The open interval $(0, 1)$ is not compact. Consider the open cover $\mathcal{V} = \{(0, \frac{1}{2}), (0, \frac{2}{3}), (0, \frac{3}{4}), (0, \frac{4}{5}), \ldots\}$ of $(0, 1)$. For any finite subcollection of $\mathcal{V}$ there is some positive integer n such that $(0, \frac{n}{n+1})$ is the largest interval in the subcollection, hence no finite subcollection covers the entire interval $(0, 1)$. Thus $(0, 1)$ is not compact. (There are certainly open covers of $(0, 1)$ that have finite subcovers, for example the open cover $\{(0, \frac{1}{2}), (\frac{1}{2}, 1), (0, \frac{2}{3}), (\frac{2}{3}, 1), (0, \frac{3}{4}), (\frac{3}{4}, 1), \ldots\}$ of $(0, 1)$, but to prove that a set is compact we need to show that all open covers have finite subcovers, and that is not the case for $(0, 1)$.)

Let us compare the intervals $[0, 1]$ and $(0, 1)$. Both intervals have infinitely many points, but they behave rather differently from a topological point of view. The open cover $\mathcal{V}$ that we used with $(0, 1)$ does not work with $[0, 1]$, since it misses the endpoints of $[0, 1]$. We could try the open cover $\{[0, \frac{1}{2}), [0, \frac{2}{3}), [0, \frac{3}{4}), [0, \frac{4}{5}), \ldots (a, 1]$

for any choice of a such that $0 \leq a < 1$. But no matter what our choice of a is, eventually $a < \frac{n}{n+1}$ for large enough n, and then it will be the case that

$$[0, 1] = [0, \frac{1}{2}) \cup [0, \frac{2}{3}) \cup [0, \frac{3}{4}) \cup [0, \frac{4}{5}) \cup \ldots \cup [0, \frac{n}{n+1}) \cup (a, 1],$$

which is a finite union. Of course, it might be that some more complicated method will work for $[0, 1]$, but we will later see that this is not the case. We thus see a type of "finiteness" in sets with infinitely many points.

(3) The space $\mathbb{R}^n$ is not compact for any n. Cover $\mathbb{R}^n$ with the union of all open balls of integer radius centered at the origin. Clearly this open cover has no finite subcover. $\Diamond$

More examples of compact sets will have to wait until we have proved some facts about compactness; we start with the following simple fact.

Lemma 1.6.2. *The union of finitely many compact sets is compact.*

Proof. Exercise 1.6.1. $\square$

The word "finite" cannot be dropped from Lemma 1.6.1 (see Exercise 1.6.2).

It would be nice to have a less abstract characterization of compact sets than given directly by the definition. From Example 1.6.1 (2), it should not be surprising that there is a relationship between compactness and closedness, as expressed in the following lemma. Part (ii) of this lemma does not imply that all closed sets are compact (for example $\mathbb{R}$ is a closed subset of $\mathbb{R}^2$, but it is not compact). It should also be pointed out that part (i) of the following lemma does not hold as stated for general topological spaces, though it does hold if the topological space is assumed to be Hausdorff (a certain property of topological spaces).

Lemma 1.6.3.

(i) *Let $A \subset \mathbb{R}^n$ be a set, and let $B \subset A$ be a compact subset. Then B is a closed subset of A.*

(ii) *Let $A \subset \mathbb{R}^n$ be compact, and let C be a closed subset of A. Then C is compact.*

Proof. (i). Assume that $A - B$ is non-empty (otherwise the result is trivial). We need to show that $A - B$ is open in A. Let $x \in A - B$ be a point. Using

Exercise 1.2.15 it can be verified that the collection of sets

$$\left\{B \cap \left(A - \overline{O}_{1/n}(x, A)\right)\right\}_{n \in \mathbb{Z}^+}$$

is an open cover of B. See Figure 1.6.1. By compactness of B there is some finite subcover of this open cover. Let N be the largest positive integer for which $B \cap \left(A - \overline{O}_{1/N}(x, A)\right)$ is in the finite subcover. It follows that $B \subset A - \overline{O}_{1/N}(x, A)$, and hence $O_{1/N}(x, A) \subset A - B$. Thus $A - B$ is an open subset of A.

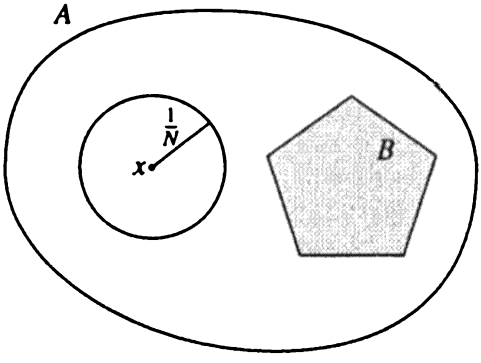


Figure 1.6.1

(ii). Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of C. By Lemma 1.2.8, for each $i \in I$ there is an open subset U_i' of A such that $U_i = U_i' \cap C$. See Figure 1.6.2. Observe that $C \subset \bigcup_{i \in I} U_i'$. It now follows that the collection of sets $\{U_i'\}_{i \in I} \cup \{A - C\}$ is an open cover of A, since $A - C$ is an open subset of A. This open cover has a finite subcover by the compactness of A. This finite subcover might or might not contain the set $A - C$; suppose that the rest of the finite subcover is $\{U_{i_1}', \ldots, U_{i_m}'\}$. It must be the case that $C \subset U_{i_1}' \cup \ldots \cup U_{i_m}'$, since even if $A - C$ were in the finite subcover, the set $A - C$ does not contribute any points in C. It is now straightforward to verify that $\{U_{i_1}, \ldots, U_{i_m}\}$ is a cover of C. Since what we did applies to any open cover $\mathcal{U}$ of C, it follows that C is compact. □

Though any compact set is closed in any set containing it, the converse is not true; to find necessary and sufficient conditions for compactness we need the following definition.

Definition. Let $A \subset \mathbb{R}^n$ be a set. Then A is **bounded** if there is some non-negative real number R such that A is contained in the open ball of

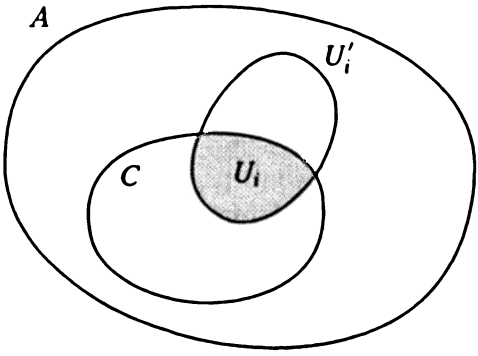


Figure 1.6.2

radius R centered at the origin. A subset of $\mathbb{R}^n$ that is not bounded is called **unbounded**. ◊

Example 1.6.4. Any open ball of the form $O_r(x, \mathbb{R}^n)$ is bounded in $\mathbb{R}^n$, since

$$O_r(x, \mathbb{R}^n) \subset O_{\|x\|+r}(0_n, \mathbb{R}^n).$$

On the other hand, $\mathbb{R}$ is definitely unbounded as a subset of itself. ◊

The property of being bounded is not preserved by homeomorphisms, not to mention arbitrary continuous maps; for example, we saw that any open interval is homeomorphic to $\mathbb{R}$, and yet finite intervals are bounded, whereas $\mathbb{R}$ is not. Boundedness is still a very useful property, an indication of which is given in the following result.

Lemma 1.6.5. *Let $A \subset \mathbb{R}^n$ be a compact set. Then A is bounded.*

Proof. For each positive integer n, let $U_n = A \cap O_n(0_n, \mathbb{R}^n)$. The collection $\{U_n\}_{n \in \mathbb{Z}^+}$ is an open cover of A. By compactness A is covered by finitely many of the sets U_n. If N is the radius of the largest of these finitely many sets U_n, then A is contained in the open ball with radius N centered at the origin. □

We have thus seen two properties of any compact set in $\mathbb{R}^n$, namely that it is closed in any set containing it (including $\mathbb{R}^n$), and that it is bounded. As long as we are only dealing with subsets of Euclidean space these two properties actually characterize compactness, where "closed" here means closed as a subset of Euclidean space, and not relatively closed in some subset of Euclidean space. This characterization of compactness definitely does not hold in the more general setting of topological spaces (or even metric spaces).

Theorem 1.6.6 (Heine–Borel Theorem). *A subset $A \subset \mathbb{R}^n$ is compact iff it is closed in $\mathbb{R}^n$ and bounded.*

A typical application of the Heine–Borel Theorem is to deduce that all closed balls in $\mathbb{R}^n$ are compact. In order to prove the Heine–Borel Theorem, we need the following two propositions, the first of which is really the crucial step. It makes use of the Least Upper Bound Property of the real numbers.

Proposition 1.6.7. *A closed interval in $\mathbb{R}$ is compact.*

Proof. Let $[a, b]$ be an interval in $\mathbb{R}$; assume $a < b$ (since the case $a = b$ is trivial). Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of $[a, b]$. Define $S \subset [a, b]$ to be

$$S = \{x \in [a, b] \mid [a, x] \text{ is covered by finitely many sets in } \mathcal{U}\}.$$

The set S is non-empty, since $a \in S$. Because $\mathcal{U}$ is a cover of $[a, b]$ there must be some U_i containing a, and thus there is some number $\epsilon > 0$ such that $[a, a + \epsilon) \subset U_i$; hence $a + \frac{\epsilon}{2} \in S$, so S contains elements greater than a. Further, the set S is bounded above by the number b. The Least Upper Bound property of the real numbers now tells us that S has a least upper bound, say z. Certainly $a < z$. We claim that $z \in S$; that is, that $[a, z]$ is covered by finitely many sets in $\mathcal{U}$. To verify this claim, let U_r be a member of $\mathcal{U}$ containing z. By the openness of U_r in $[a, b]$ it follows that the half-open interval $(z - \epsilon, z]$ is contained in U_r for some small enough number $\epsilon > 0$. Since $z = \text{lub}S$, there must be some element $y \in S$ contained in $(z - \epsilon, z]$ (or otherwise $z - \epsilon$ would be an upper bound for S). See Figure 1.6.3. By definition $[a, y)$ can be covered by finitely many sets in $U_{i_1}, \ldots, U_{i_p} \in \mathcal{U}$, and therefore $[a, z]$ can be covered by $U_r, U_{i_1}, \ldots, U_{i_p}$. Therefore $z \in S$.

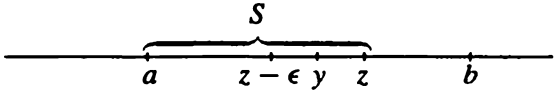


Figure 1.6.3

We now claim that in fact $z = b$, which would prove the proposition. Assume that $z \neq b$, so that $z < b$. The set U_r will then contain the open interval $(z - \epsilon, z + \epsilon)$ for some (possibly smaller) $\epsilon > 0$. It then follows that $[a, z + \frac{\epsilon}{2}]$

is covered by the sets $U_r, U_{i_1}, \ldots, U_{i_p}$, contradicting z being the least upper bound of S. Hence $z = b$. □

Proposition 1.6.8. *The product of finitely many compact subsets of Euclidean space is compact.*

Proof. Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be compact sets. We will show that the product $A \times B \subset \mathbb{R}^{n+m}$ is compact. The result for products of more than two compact sets would then follow by induction of the number of factors in the product. Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of $A \times B$. Let $a \in A$ be a point. The sets $\{U_i \cap (\{a\} \times B)\}_{i \in I}$ form an open cover of $\{a\} \times B$. The set $\{a\} \times B$ is homeomorphic to B, and hence is compact by Exercise 1.6.3. Hence some finite subcollection of the sets $\{U_i \cap (\{a\} \times B)\}_{i \in I}$ cover $\{a\} \times B$, say $U_1 \cap (\{a\} \times B), \ldots, U_p \cap (\{a\} \times B)$. Therefore $\{a\} \times B \subset U_1 \cup \ldots \cup U_p$.

We claim that there exists an open subset $W_a \subset A$ containing a such that

$$W_a \times B \subset U_1 \cup \ldots \cup U_p.$$

See Figure 1.6.4. Assuming that the claim is true for each $a \in A$, then the collection $\{W_a\}_{a \in A}$ forms an open cover of A, since $a \in W_a$ for each $a \in A$. By the compactness of A it follows that A is covered by finitely many of the W_a, say $W_{a_1}, \ldots, W_{a_r}$. Thus the sets $W_{a_1} \times B, \ldots, W_{a_r} \times B$ cover $A \times B$. By the claim each set $W_{a_i} \times B$ is contained in the union of finitely many members of $\mathcal{U}$, and hence so is $A \times B$, which proves the proposition.

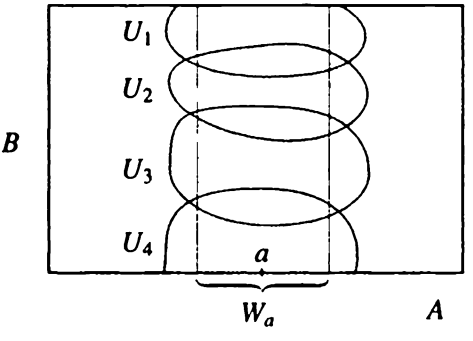


Figure 1.6.4

To prove the claim we state it more generally: If $a \in A$ is a point, and $V \subset A \times B$ is an open set containing $\{a\} \times B$, then there is an open subset

$W \subset A$ containing a such that $W \times B \subset V$. Let $a \in A$ be fixed. Using Lemma 1.2.9 it follows that for each point $(a, b) \in \{a\} \times B$ there are numbers $\delta_b, \epsilon_b > 0$ such that $O_{\delta_b}(a, A) \times O_{\epsilon_b}(b, B) \subset V$. The collection $\{O_{\epsilon_b}(b, B)\}_{b \in B}$ forms an open cover of B. By the compactness of B it follows that B is covered by finitely many of the $O_{\epsilon_b}(b, B)$, say $O_{\epsilon_{b_1}}(b_1, B), \ldots, O_{\epsilon_{b_s}}(b_s, B)$.

Let the number δ be defined by $\delta = \min\{\delta_{b_1}, \ldots, \delta_{b_s}\}$. Observe that $\delta > 0$. We now define the set W to be $W = O_\delta(a, A)$. By definition $a \in W$, so it remains to be seen that $W \times B \subset V$. Using standard results on sets we compute that

$$\begin{aligned} W \times B &= O_\delta(a, A) \times \left[O_{\epsilon_{b_1}}(b_1, B) \cup \ldots \cup O_{\epsilon_{b_s}}(b_s, B)\right] \\ &= \left[O_\delta(a, A) \times O_{\epsilon_{b_1}}(b_1, B)\right] \cup \ldots \cup \left[O_\delta(a, A) \times O_{\epsilon_{b_s}}(b_s, B)\right] \\ &\subset \left[O_{\delta_{b_1}}(a, A) \times O_{\epsilon_{b_1}}(b_1, B)\right] \cup \ldots \cup \left[O_{\delta_{b_s}}(a, A) \times O_{\epsilon_{b_s}}(b_s, B)\right]. \end{aligned}$$

Since each of the terms in the last expression is contained in V, the claim is proved. □

An example of the use of the above proposition, in combination with Proposition 1.6.7, is to show that any closed rectangle in $\mathbb{R}^2$ (that is, a set of the form $[a, b] \times [c, d]$) is compact.

Proof of Theorem 1.6.6. If the set A is compact, then it is closed and bounded by Lemmas 1.6.3 (i) and 1.6.5. Now suppose A is closed and bounded. Since A is bounded, there is some non-negative real number R such that A is contained in the open ball of radius R centered at the origin. Hence A is also contained in the set

$$\underbrace{[-R, R] \times \cdots \times [-R, R]}_{n \text{ times}} \subset \mathbb{R}^n.$$

By Propositions 1.6.7 and 1.6.8 it follows that this product of intervals is compact. Since A is closed in $\mathbb{R}^n$ it is also closed in the product of intervals by Lemma 1.2.12. Hence A is a closed subset of a compact set, and it follows from Lemma 1.6.3 (ii) that A is compact. □

Our final application of compactness is the following result.

Theorem 1.6.9 (Lebesgue Covering Lemma). *Let $A \subset \mathbb{R}^n$ be a compact set, and let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of A. Then there is a number $\epsilon > 0$*

such that for every $a \in A$ the set $O_\epsilon(a, A)$ is contained in a member of $\mathcal{U}$. (The number ϵ is called the **Lebesgue number** *of the cover $\mathcal{U}$.)*

Proof. We follow [DU]. For each point $a \in A$ there is some $i \in I$ such that $a \in U_i$ (if there is more than one such i, choose one). By the definition of openness there is some number $r(a) > 0$ such that $O_{r(a)}(a, A) \subset U_i$. The collection of sets

$$\{O_{r(a)/2}(a, A) \mid a \in A\}$$

is an open cover of A; by compactness we can find a finite subcover, which has the form

$$\{O_{r(a_1)/2}(a_1, A), \ldots, O_{r(a_p)/2}(a_p, A)\}.$$

Define the number ϵ to be

$$\epsilon = \min\{\frac{r(a_1)}{2}, \ldots, \frac{r(a_p)}{2}\},$$

which is certainly positive. To demonstrate that ϵ is as desired, let $x \in A$ be any point, and we will show that $O_\epsilon(x, A)$ is contained in one of the sets U_i. Observe that there is a number $k \in \{1, \ldots, p\}$ such that $x \in O_{r(a_k)/2}(a_k, A)$. Let $z \in O_\epsilon(x, A)$ be any point. Using the triangle inequality, we compute

$$\|z - a_k\| \leq \|z - x\| + \|x - a_k\| < \epsilon + \frac{r(a_k)}{2} \leq r(a_k).$$

Hence $z \in O_{r(a_k)}(a_k, A)$, and it follows that

$$O_\epsilon(x, A) \subset O_{r(a_k)}(a_k, A).$$

This latter set is contained in one of the sets U_i by choice of $r(a_k)$. □

Finally, we turn to the effect of continuous maps on compactness. The following theorem shows that compactness behaves nicely with respect to continuous maps, just as connectedness does. The proof of this theorem shows the power of the rather abstract definition of compactness.

Theorem 1.6.10. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a continuous map. If A is compact then so is $f(A)$.*

Proof. Let $\mathcal{U} = \{U_i\}_{i \in I}$ be an open cover of $f(A)$; we need to show that $\mathcal{U}$ has a finite subcover. By Lemma 1.2.8, for each $i \in I$ there is an open subset U_i'

of B such that $U_i = U_i' \cap f(A)$. Since f is continuous it follows that each set $f^{-1}(U_i')$ is an open subset of A. It is not hard to see that the collection

$$\mathcal{V} = \left\{f^{-1}(U_i')\right\}_{i \in I}$$

is an open cover of A. By the compactness of A we deduce that $\mathcal{V}$ has a finite subcover, so that there are indices $i_1, \ldots, i_m \in I$ such that

$$A = f^{-1}(U_{i_1}') \cup \ldots \cup f^{-1}(U_{i_m}').$$

Applying f to both sides of this equation, and using standard results concerning functions, we obtain

$$f(A) = f(f^{-1}(U_{i_1}') \cup \ldots \cup f^{-1}(U_{i_m}')) = f(f^{-1}(U_{i_1}')) \cup \ldots \cup f(f^{-1}(U_{i_m}')).$$

It can be verified that $U_i = f(f^{-1}(U_i'))$ for all $i \in I$. Hence $f(A) = U_{i_1} \cup \ldots \cup U_{i_m}$, and thus $\left\{U_{i_1}, \ldots, U_{i_m}\right\}$ is a finite subcover of $\mathcal{U}$. □

The above theorem can be used to prove the Extreme Value Theorem, used in Calculus.

Proposition 1.6.11. *Let $A \subset \mathbf{R}$ be a compact set. Then A has a maximal member and a minimal member, that is, there are points $x_1, x_2 \in A$ such that $x_1 \leq x \leq x_2$ for all $x \in A$.*

Proof. Exercise 1.6.8. □

Proposition 1.6.12. *Let $A \subset \mathbf{R}^n$ be a compact set, and let $f: A \to \mathbf{R}$ be a continuous map. Then f has a maximum value on A and a minimum value on A, that is there are points $x_{max}, x_{min} \in A$ such that $f(x_{min}) \leq f(x) \leq f(x_{max})$ for all $x \in A$.*

Proof. Combine Theorem 1.6.10 with Proposition 1.6.11. □

Theorem 1.6.13 (Extreme Value Theorem). *Let $[a, b]$ be a closed interval in $\mathbf{R}$, and let $f: [a, b] \to \mathbf{R}$ be a continuous function. Then f has a maximum value on $[a, b]$ and a minimum value on $[a, b]$.*

Proof. This follows immediately from Propositions 1.6.7 and 1.6.12. □

We saw in Example 1.4.2 that a continuous bijection need not be a homeomorphism, and a continuous surjection need not be a quotient map. The following proposition, also of use later on, shows that no such examples can be found with compact domains. (As stated, this proposition does not hold for

general topological spaces, though it does hold if the codomain is assumed to be Hausdorff.)

Proposition 1.6.14. *Let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be sets, and let $f: A \to B$ be a continuous map. Then*

(i) *if A is compact, then f is a closed map;*
(ii) *if A is compact and f is surjective, then f is a quotient map;*
(iii) *if A is compact and f is bijective, then f is a homeomorphism.*

Proof. (i). Let C be a closed subset of A; we need to show that $f(C)$ is a closed subset of B. By Lemma 1.6.3 (ii) the set C is a compact subset of A. By Theorem 1.6.10 the set $f(C)$ is a compact subset of B, and by Lemma 1.6.3 (i) we deduce that $f(C)$ is a closed subset of B.

(ii) & (iii). Because the map f is continuous, we know that if $U \subset B$ is open, then $f^{-1}(U)$ is open in A. By Lemma 1.4.3 and the definition of quotient maps, it will suffice to show that for any subset $U \subset B$, if $f^{-1}(U)$ is open in A then U is open B. So, suppose $U \subset B$ is such that $f^{-1}(U)$ is open in A. Then $A - f^{-1}(U)$ is a closed subset of A. By part (i) the map f is a closed map, so that $f(A - f^{-1}(U))$ is a closed subset of B. However, using the fact that f is surjective (in both cases (ii) and (iii)), it is not hard to show that $f(A - f^{-1}(U)) = B - U$. Therefore $B - U$ is closed in B, so that U is open in B. □

Exercise 1.6.4 shows that the hypothesis of compactness in Proposition 1.6.14 cannot be replaced with the weaker hypothesis of closedness.

Exercises

1.6.1*. Prove Lemma 1.6.2.

1.6.2. Give an infinite collection of compact sets whose union is not compact. Give an infinite collection of compact sets whose union is compact.

1.6.3*. Prove that if two subsets of Euclidean space are homeomorphic, and one is compact, then so is the other.

1.6.4. Find sets $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$, with A a closed subset of $\mathbb{R}^n$, and a continuous bijection $f: A \to B$, such that f is not a homeomorphism. Convince

yourself that this example will also show that there is a continuous surjection $f: A \to B$ with A closed in $\mathbb{R}^n$, such that f is not a quotient map.

1.6.5*. (Refer to Exercise 1.3.7.) Let $A \subset \mathbb{R}^n$ be a compact set. Prove that any continuous function $f: A \to \mathbb{R}^n$ is uniformly continuous. (See [BT, §16] for a solution.)

1.6.6. Prove that a closed interval in $\mathbb{R}$ is not homeomorphic to $\mathbb{R}$.

1.6.7*. Let $a, b, \epsilon \in \mathbb{R}$ be numbers with $a < b$ and $\epsilon > 0$. Suppose the function

$$f: [a, b] \times (-\epsilon, \epsilon) \to \mathbb{R}$$

is continuous, and $f(\binom{s}{0}) > 0$ for all $s \in [a, b]$. Show that there are numbers $M, \delta > 0$ such that $\delta \leq \epsilon$ and $f(\binom{s}{t}) \geq M$ for all $\binom{s}{t} \in [a, b] \times (-\delta, \delta)$.

1.6.8*. Prove Proposition 1.6.11.

1.6.9. If $A \subset \mathbb{R}$ is a compact connected set, show that A is a closed interval (possibly of the form $[a, a]$). If $A \subset \mathbb{R}$ is a compact set, show that it is the union of disjoint closed intervals.

1.6.10*. Let $[a, b]$ be a closed interval in $\mathbb{R}$, and let $f: [a, b] \to \mathbb{R}$ be a continuous function such that $f(a) = f(b)$. Show that there is a point $x \in (a, b)$ such that f is not injective on any open neighborhood of x.

1.6.11*. Let $A, B \subset \mathbb{R}^n$ be disjoint compact sets. Show that there is a number $m > 0$ such that $\|a - b\| \geq m$ for all $a \in A$ and $b \in B$.

1.6.12*. Let $A \subset \mathbb{R}^n$ be a compact, connected set, and let $p, q \in A$ be points. If $\mathcal{U} = \{U_i\}_{i \in I}$ is an open cover of A, show that there are sets $U_{i_1} \ldots U_{i_r}$ in $\mathcal{U}$ such that $p \in U_{i_1}$, $q \in U_{i_r}$ and $U_{i_k} \cap U_{i_{k+1}} \neq \emptyset$ for $i = 1, \ldots, r - 1$.

1.6.13*. Let $U \subset \mathbb{R}^2$ be an open set and let $c_1, c_2, \varphi_1, \varphi_2: [a, b] \to \mathbb{R}$ be continuous maps for some closed interval $[a, b]$ such that $\binom{c_1(s)}{c_2(s)} \in U$ for all $s \in [a, b]$. Show that there is some number $\epsilon > 0$ such that

$$\begin{pmatrix} c_1(s) + t\varphi_1(s) \\ c_2(s) + t\varphi_2(s) \end{pmatrix} \in U$$

for all $(s, t) \in [x, y] \times (-\epsilon, \epsilon)$.

Endnotes

Notes for Section 1.5

(A) Theorems 1.5.4 and 1.5.5 tell us of the existence of some point with certain desired properties; we are not told how to find these points, and as such these theorems are what are known as "existence theorems." Existence theorems, one of the hallmarks of modern mathematics, were given particular prominence when D. Hilbert proved such a theorem in 1888 in connection with algebraic geometry. (This theorem can be found in [KN, pp. 119–120].) There are some mathematicians who do not accept existence proofs, though they are in the minority.

(B) In Example 1.5.8, the proof that K is not path connected makes direct use of the Least Upper Bound Property of the real numbers (via Exercise 1.2.17). It is possible to give a proof that K is not path connected without directly invoking the Least Upper Bound Property, but our proof is more straightforward intuitively. See [MU2, §3-2] for an alternate proof.

Notes for Section 1.6

In Proposition 1.6.8 we restricted our attention to finite products of compact sets. The same result also holds for infinite products, and is known as the Tychonoff Theorem. The proof of the Tychonoff Theorem is substantially more difficult than the proof in the finite product case, making use of the axiom of choice (see [MU2, §5-1] for a good discussion).

CHAPTER II

Topological Surfaces

2.1 Introduction

If we wish to be able to make interesting geometric statements about subsets of Euclidean space, we need to restrict our attention to a reasonable class of geometric objects. One of the most widely studied type of geometric objects are manifolds; the two-dimensional version of a manifold is a surface, which we will define rigorously in the next section.

Two of the most well-known examples of surfaces are the plane $\mathbb{R}^2$ and the unit sphere in $\mathbb{R}^3$; this sphere is denoted S^2 and is defined by

$$S^2 = \{x \in \mathbb{R}^3 \mid \|x\| = 1\}.$$

Since any two objects that are homeomorphic to one another are essentially interchangeable from a topological viewpoint, we will refer to any subset of Euclidean space homeomorphic to S^2 as a sphere. If we let S^1 denote the unit circle in $\mathbb{R}^2$, that is

$$S^1 = \{x \in \mathbb{R}^2 \mid \|x\| = 1\},$$

then $S^1 \times \mathbb{R} \subset \mathbb{R}^3$ is an infinite right circular cylinder, which is a surface; see Figure 2.1.1. Another important surface is the torus, denoted T^2, which is the surface of a bagel; the torus is hollow, like an inner tube. See Figure 2.1.2. This surface will be described analytically in Section 5.3 as a surface of revolution. We will refer to any subset of Euclidean space homeomorphic to T^2 as a torus. It can be seen that $S^1 \times S^1 \subset \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4$ is a torus. Of course, not everything in $\mathbb{R}^3$ is a surface, for example the objects pictured in Figure 2.1.3.

There is, actually, more than one definition of a surface, depending upon which types of propertie s one is interested in discussing. We will discuss the three main types of surfaces: topological, simplicial and smooth. Topological surfaces will be treated in the present chapter, simplicial surfaces will be treated in the next chapter, and smooth surfaces will be treated in Chapters V–VIII. In a certain sense, to be made more precise later on, every surface of one of the three types can be converted into either of the other types (something that does not hold in higher dimensions).

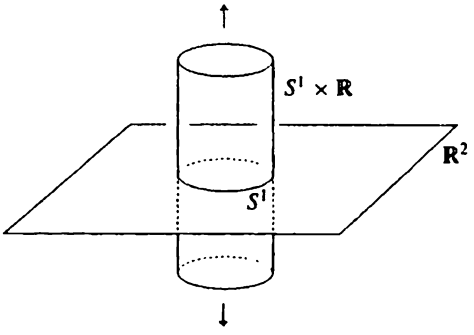


Figure 2.1.1

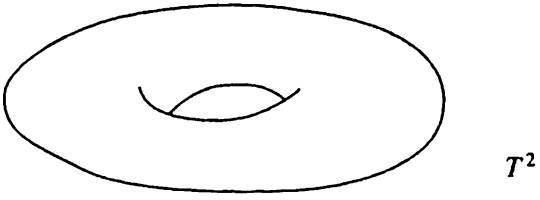


Figure 2.1.2

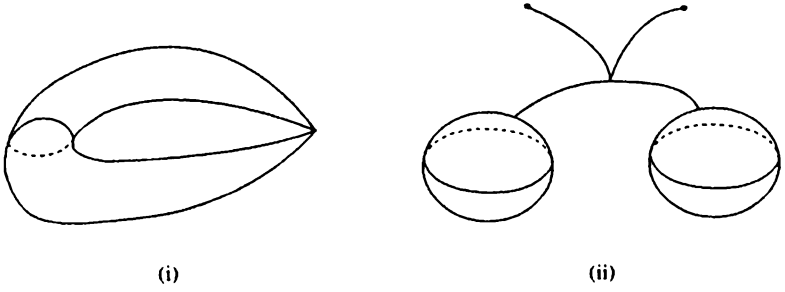


Figure 2.1.3

Our study of surfaces, similar to a botanist's study of plants, occurs on two levels: macro (what are all the types of plants in the world, and how does one identify them) and micro (how does an individual plant operate). On the micro level we study geometric properties of smooth surfaces in $\mathbb{R}^3$. On the macro level we wish to find a list of all possible surfaces, and find a convenient way to distinguish between different surfaces. We start with the macro question,

known as *classification,* which will be discussed in this chapter and the next. The micro questions will be addressed in the rest of the book. Before proceeding to surfaces, we discuss some crucial topological tools in Section 2.2.

2.2 Arcs, Disks and 1-Spheres

The three types of objects studied in this section are all fundamental building blocks of surfaces. A rigorous analysis of these objects requires two of the most important (and difficult) theorems on geometric topology, Invariance of Domain and the Schönflies Theorem; we will only give references for the proofs of these results. In contrast to much of our discussion in Chapter I, where analogs of most of our concepts and results hold in the more general setting of topological spaces, the first of these two theorems holds only in $\mathbb{R}^n$, and the second holds only in $\mathbb{R}^2$.

We start with some notation: Let D^2 and int D^2 denote the standard closed and open unit disks in $\mathbb{R}^2$, that is,

$$D^2 = \{x \in \mathbb{R}^2 \mid \|x\| \leq 1\}$$
$$\text{int } D^2 = \{x \in \mathbb{R}^2 \mid \|x\| < 1\} = O_1(O_2, \mathbb{R}^2).$$

The disk D^2 is the union of two disjoint subsets, namely int D^2 and S^1, which we will refer to as the interior and boundary of D^2.

Definition. A subset of $\mathbb{R}^n$ that is homeomorphic to the closed interval $[-1, 1]$ is an **arc**; a subset of $\mathbb{R}^n$ that is homeomorphic to the disk D^2 is a **disk**; a subset of $\mathbb{R}^n$ that is homeomorphic to the unit circle S^1 is a **1-sphere** (also known as a **simple closed curve**).

Given that D^2 has a well-defined interior and boundary, it would be reasonable to expect that any disk $B \subset \mathbb{R}^n$ also has an interior and a boundary. If $h: D^2 \to B$ is a homeomorphism, it would be plausible to define the interior and boundary of B to be the sets $h(\text{int } D^2)$ and $h(S^1)$ respectively. Since there are many homeomorphisms $D^2 \to B$, we would need to verify that the definition of interior and boundary of B does not depend upon the choice of homeomorphism; to do so we need the following theorem, that is of fundamental importance in geometric topology. Consider a subset of $\mathbb{R}^2$ that is homeomorphic to $\mathbb{R}^2$, such as the interior of a square; intuitively, any such set appears to be open in $\mathbb{R}^2$. (By

contrast, the interior of a square sitting in $\mathbb{R}^3$ is not open in $\mathbb{R}^3$.) The following theorem shows that our intuition is correct.

Theorem 2.2.1 (Invariance of Domain). *Let $U \subset \mathbb{R}^n$ be homeomorphic to $\mathbb{R}^n$. Then U is open in $\mathbb{R}^n$.*

Proofs of Invariance of Domain can be found in [MU3], [MS2] or [H-W, p. 95]. The first two proofs cited use algebraic topology; the third is more elementary (though not necessarily simpler). Different references use different (though equivalent) statements of Invariance of Domain.

The converse to Invariance of Domain is not true; there are many open subsets of $\mathbb{R}^n$ which are not homeomorphic to $\mathbb{R}^n$. An immediate corollary of Invariance of Domain is the following theorem, which may seem obvious, but is not trivial to prove.

Theorem 2.2.2. *Let n and m be positive integers that are not equal. Then $\mathbb{R}^n \not\approx \mathbb{R}^m$.*

Proof. Without loss of generality assume that $n < m$. We consider $\mathbb{R}^n$ to be a subset of $\mathbb{R}^m$ by identifying $\mathbb{R}^n$ with the space of all vectors of the form

$$\begin{pmatrix} x_1 \\ \vdots \\ x_n \\ 0 \\ \vdots \\ 0 \end{pmatrix} \in \mathbb{R}^m.$$

It can be verified directly from the definition of openness that $\mathbb{R}^n$ is not an open subset of $\mathbb{R}^m$. By Invariance of Domain a subset of $\mathbb{R}^m$ that is not open in $\mathbb{R}^m$ cannot be homeomorphic that $\mathbb{R}^m$. □

The following lemma is the result we wanted concerning disks.

Lemma 2.2.3. *Let $B \subset \mathbb{R}^n$ be a disk, and let $h_1, h_2 : D^2 \to B$ be homeomorphisms. Then $h_1(\text{int } D^2) = h_2(\text{int } D^2)$ and $h_1(S^1) = h_2(S^1)$.*

Proof. Since h_1 and h_2 are bijections it suffices to show that $h_1(\text{int } D^2) = h_2(\text{int } D^2)$, and showing this fact is equivalent to showing that $h_2^{-1} \circ h_1(\text{int } D^2) = \text{int } D^2$. The map $h_2^{-1} \circ h_1 : D^2 \to D^2$ is a homeomorphism, and thus the set

$h_2^{-1} \circ h_1(\text{int } D^2)$ is homeomorphic to int D^2. Since int D^2 is homeomorphic to $\mathbf{R}^2$ (by Exercise 1.4.1), it follows from Invariance of Domain that $h_2^{-1} \circ h_1(\text{int } D^2)$ is open in $\mathbf{R}^2$. Suppose that $h_2^{-1} \circ h_1(\text{int } D^2) \cap S^1 \neq \emptyset$; let $x \in h_2^{-1} \circ h_1(\text{int } D^2) \cap S^1$ be any point. It is not hard to verify that any open ball centered at x must contain points outside of D^2, and hence it must contain points outside of $h_2^{-1} \circ h_1(\text{int } D^2)$. This conclusion would contradict the openness of $h_2^{-1} \circ h_1(\text{int } D^2)$, and hence it must be the case that $h_2^{-1} \circ h_1(\text{int } D^2) \cap S^1 = \emptyset$; thus $h_2^{-1} \circ h_1(\text{int } D^2) \subset \text{int } D^2$.

By reversing the roles of h_1 and h_2 one could also conclude that $h_1^{-1} \circ h_2(\text{int } D^2) \subset \text{int } D^2$. Applying the appropriate inverse maps to both sides of this inclusion it follows that $\text{int } D^2 \subset h_2^{-1} \circ h_1(\text{int } D^2)$. Combining this inclusion with the result of the previous paragraph gives the desired result. □

The analog for arcs of the above lemma is given in Exercise 2.2.3. We are now able to make the following definition.

Definition. Let $B \subset \mathbf{R}^n$ be a disk. The **interior** and **boundary** of B, denoted int B and ∂B respectively, are the sets $h(\text{int } D^2)$ and $h(S^1)$ respectively for any homeomorphism $h: D^2 \to B$. Let $J \subset \mathbf{R}^n$ be an arc. The **interior** and **boundary** of J, denoted int J and ∂J respectively, are the sets $h((-1, 1))$ and $h(\{-1\} \cup \{1\})$ respectively for any homeomorphism $h: [-1, 1] \to J$.

The following figure shows some disks and arcs together with their boundaries.

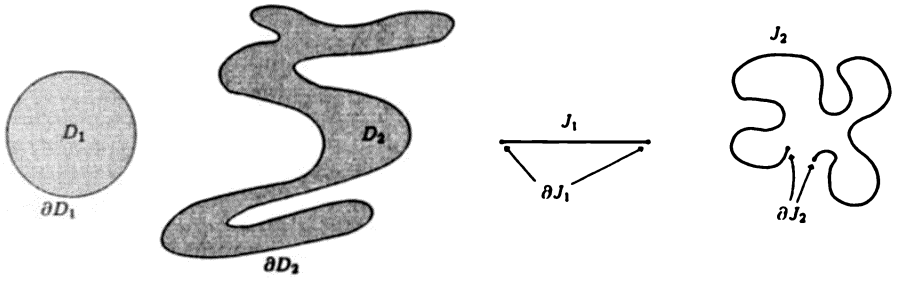


Figure 2.2.1

Observe that the boundary of any disk is a 1-sphere. Does the converse hold? That is, for any 1-sphere $C \subset \mathbf{R}^n$ is there a disk $B \subset \mathbf{R}^n$ such that $C = \partial B$? The answer in general is no. Consider the curve $C \subset \mathbf{R}^3$ shown in Figure 2.2.2. This curve is known as the trefoil knot, and it can be shown using

the techniques of knot theory that there is no disk $B \subset \mathbb{R}^3$ such that ∂B is this knot; see [RO, p. 52] for a proof. By contrast, it turns out that any 1-sphere in $\mathbb{R}^2$ is the boundary of a disk in $\mathbb{R}^2$; this result should not be taken as entirely obvious, since an arbitrary 1-sphere in $\mathbb{R}^2$ can be quite complicated, as in Figure 2.2.3. We will deduce this fact from the following result, the Schönflies Theorem, stated below.

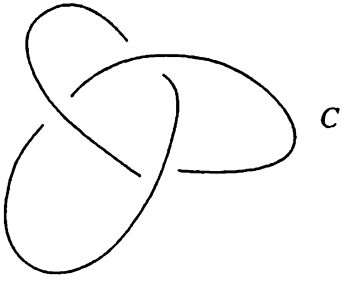


Figure 2.2.2

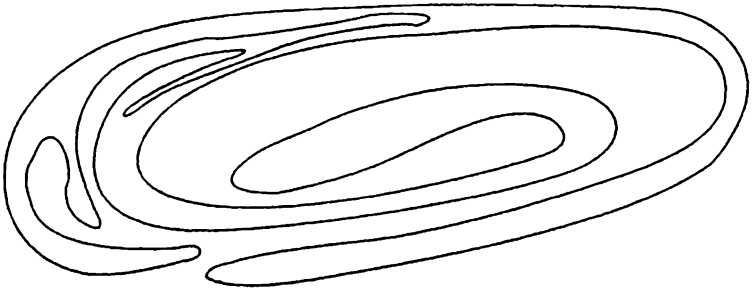


Figure 2.2.3

Definition. Let $h: \mathbb{R}^2 \to \mathbb{R}^2$ be a homeomorphism. The function h is **the identity map outside a disk** if there is some disk $A \subset \mathbb{R}^2$ such that $h|(\mathbb{R}^2 - \text{int } A)$ is the identity map.

Theorem 2.2.4 (Schönflies Theorem). *Let $C \subset \mathbb{R}^2$ be a 1-sphere. Then there is a homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H(S^1) = C$ and H is the identity map outside a disk.*

See [BI], [CA2], [MO] or [TH] for proofs; different books state the Schönflies Theorem differently, though all variants, including the version we use, are equiv-

alent. From the Schönflies Theorem we can now deduce the following result, the first part of which is the well known Jordan Curve Theorem (one of those results in topology that seem obvious intuitively but are surprisingly hard to prove), and the second part answers our question about 1-spheres in $\mathbb{R}^2$ being the boundaries of disks. Actually, it is not quite fair to claim that we have deduced the Jordan Curve Theorem from the Schönflies Theorem, since proofs of the latter theorem usually make use of the former.

Corollary 2.2.5. *Let $C \subset \mathbb{R}^2$ be a 1-sphere.*

(i) *(Jordan Curve Theorem) The set $\mathbb{R}^2 - C$ has precisely two components, one of which is bounded and one of which is unbounded.*

(ii) *The union of C and the bounded component of $\mathbb{R}^2 - C$ is a disk, of which C is the boundary.*

Proof. Exercise 2.2.5. □

Another useful corollary to the Schönflies Theorem is the following.

Corollary 2.2.6. *Let $B_1, B_2 \subset \mathbb{R}^2$ be disks. Then there is a homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H(B_1) = B_2$ and H is the identity map outside a disk containing B.*

Proof. Since ∂B_i is a 1-sphere for each $i = 1, 2$, it follows from the Schönflies Theorem that for each i there is a homeomorphism $H_i: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H_i(S^1) = \partial B_i$ and H_i is the identity map outside a disk. It follows from Exercise 2.2.6 that $H_i(D^2) = B_i$. The map $H = H_2 \circ (H_1)^{-1}$ is thus a homeomorphism of $\mathbb{R}^2$ to itself such that $H(B_1) = B_2$. Since the H_i are both the identity maps outside disks, it follows that $(H_1)^{-1}$ is the identity map outside a disk, and it follows from Exercise 2.2.7 that H is the identity map outside a disk. □

Exercises

2.2.1*. Let $A \subset \mathbb{R}^n$ be any set. Let $V \subset U \subset A$ be sets such that $U \approx \mathbb{R}^m \approx V$. If U is open in A, then show that V is open in A.

2.2.2*. Let $U \subset \mathbb{R}^n$ be open, and let $h: U \to \mathbb{R}^n$ be a homeomorphism from U onto its image. Show that $h(U)$ is open in $\mathbb{R}^n$.

2.2.3*. Let $J \subset \mathbb{R}^n$ be an arc, and let $h_1, h_2: [0, 1] \to J$ be homeomorphisms. Show that $h_1((0, 1)) = h_2((0, 1))$ and $h_1(\{0\} \cup \{1\}) = h_2(\{0\} \cup \{1\})$.

2.2.4*. Let $B_1 \subset \mathbb{R}^n$ and $B_2 \subset \mathbb{R}^m$ be disks, and let $h: B_1 \to B_2$ be a homeomorphism. Show that $h(\text{int } B_1) = \text{int } B_2$ and $h(\partial B_1) = \partial B_2$.

2.2.5*. Prove Corollary 2.2.5.

2.2.6*. Let $B_1, B_2 \subset \mathbb{R}^2$ be disks, and let $H: \mathbb{R}^2 \to \mathbb{R}^2$ be a homeomorphism such that $H(\partial B_1) = \partial B_2$. Show that $H(B_1) = B_2$.

2.2.7*. For each $i = 1, 2$ let $h_i: \mathbb{R}^2 \to \mathbb{R}^2$ be a homeomorphism which is the identity map outside a disk (not necessarily the same disk for both values of i). Show that $h_2 \circ h_1$ is the identity map outside a disk.

2.2.8. Let $B \subset \text{int } D^2$ be a disk. Show that there is a point $x \in \partial B$ such that the radial line segment in $\mathbb{R}^2$ starting at x and ending at the point of distance 1 from the origin intersects ∂B and S^1 in precisely one point each (a radial line segment is a line segment that when extended contains the origin); see Figure 2.2.4. Show that there must be at least two such points.

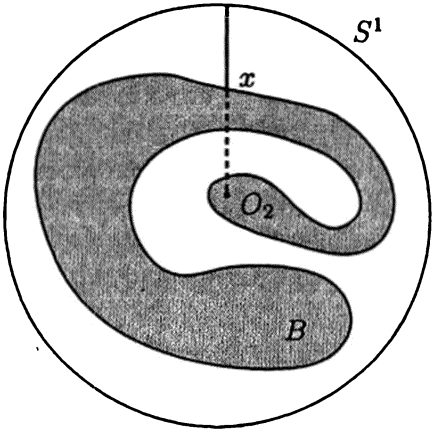


Figure 2.2.4

2.2.9. This exercise proves the Annulus Theorem in dimension 2. (This theorem was shown to be true in all dimensions other than 4 by [KI], though the the proof is extremely difficult, due to the lack of the analog of the Schönflies

Theorem in dimensions higher than 2.) Let $B_1, B_2 \subset \mathbb{R}^2$ be disks with $B_2 \subset \text{int } B_1$. Show that $B_1 - \text{int } B_2$ is homeomorphic to the washer-shaped set

$$A = \{v \in \mathbb{R}^2 \mid 1 \leq \|v\| \leq 2\}.$$

2.2.10*. Let $A \subset \mathbb{R}^m$ be a set that is homeomorphic to $\mathbb{R}^n$. Show that $n \leq m$, and that if $n < m$ then A is not open in $\mathbb{R}^m$.

2.2.11*. Show that $\mathbb{H}^n \not\approx \mathbb{R}^n$ for all $n \geq 1$.

2.2.12*. Let $B \subset \mathbb{R}^n$ be a disk, and let $J \subset B$ be an arc such that int $J \subset \partial B$. Show that $J \subset \partial B$.

2.2.13. Show that no proper subset of S^1 is homeomorphic to S^1.

2.2.14*. A subset of $\mathbb{R}^n$ is called a **theta-curve** if it is homeomorphic to the set

$$\Theta = S^1 \cup ([-1, 1] \times \{0\}) \subset \mathbb{R}^2.$$

State and prove the analog for theta-curves of both parts of Corollary 2.2.5.

2.3 Surfaces in $\mathbb{R}^n$

What is it that distinguishes sets such as $\mathbb{R}^2$, S^2 and T^2 from sets such as those pictured in Figure 2.1.3? Consider Figure 2.1.3 (i), referred to as a pinched torus since it can be obtained by taking a torus and pinching a loop around it to a point. If we draw a small ball around the pinch point, and cut out the part of the pinched torus inside the ball, we obtain after some stretching an object that looks like two open disks glued together at a single point. See Figure 2.3.1 (i). By contrast, if we draw a small ball around any point on the torus, or any other point on the pinched torus, and cut out the neighborhood of the point inside the ball, we obtain after some stretching an object that looks like one open disk. See Figure 2.3.1 (ii). In other words, a small neighborhood of any point on the torus (or the sphere or the plane) looks like an open disk, whereas on the pinched torus there is a point with a different type of neighborhood.

Definition. A subset $Q \subset \mathbb{R}^n$ is called a **topological surface**, or just **surface** for short, if each point $p \in Q$ has an open neighborhood that is homeomorphic to the open unit disk int D^2.

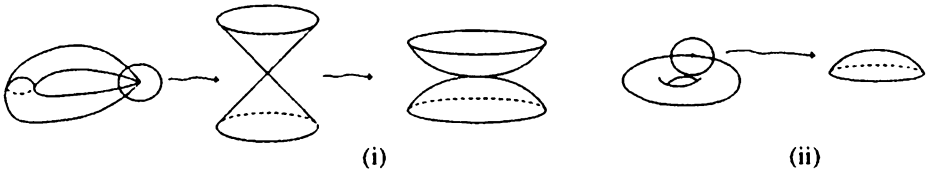


Figure 2.3.1

Since any open disk in $\mathbb{R}^2$ is homeomorphic to any other open disk in $\mathbb{R}^2$, and to $\mathbb{R}^2$ itself, it would suffice to show that each point p as in the above definition has an open neighborhood that is homeomorphic to some open disk in $\mathbb{R}^2$ or to $\mathbb{R}^2$.

Example 2.3.1. (1) The plane $\mathbb{R}^2$ is a surface, as is any open subset of $\mathbb{R}^2$, since any point in an open subset of $\mathbb{R}^2$ is contained in an open disk inside the subset.

(2) The infinite cylinder $S^1 \times \mathbb{R} \subset \mathbb{R}^3$ is a surface. Intuitively it is easy to see that every point on the infinite cylinder has an open neighborhood that is homeomorphic to an open disk; we leave it to the reader to write down a formula for such a homeomorphism. An open cylinder of the form $S^1 \times (a, b)$ is also a surface, whereas a closed cylinder $S^1 \times [a, b]$ is not a surface according to our definition, since points on the boundary do not have the required type of neighborhoods. See Figure 2.3.2. $\Diamond$

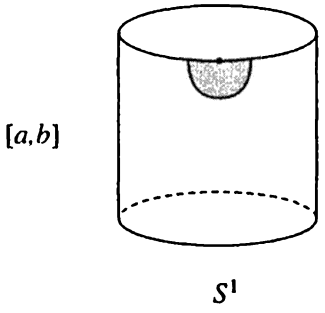


Figure 2.3.2

We are allowing our surfaces to sit in any $\mathbb{R}^n$, not just $\mathbb{R}^3$, although the latter is certainly most convenient. Some surfaces we will encounter later on (such as the Klein bottle) do not fit into $\mathbb{R}^3$. In more advanced treatments, where the full

generality of topological spaces is used, one can define surfaces as creatures unto themselves, which do not sit a priori in any surrounding space. It turns out, however, that all abstractly defined surfaces (and, more generally, manifolds) are in fact homeomorphic to surfaces (or manifolds) that reside in Euclidean space. Thus we are not losing any generality by restricting our attention to surfaces that are by definition in Euclidean space.

Since one of our goals is to distinguish between surfaces, we need to clarify what it means for two surfaces to be "the same" or not. For example, a sphere of radius 1 and a sphere of radius 2, though different from the point of view of geometry, are indistinguishable from the point of view of topology, being homeomorphic. For the duration of this chapter and the next we will consider homeomorphic surfaces as "the same." For example, the two surfaces pictured in Figure 2.3.3 (referred to as an unknotted torus and a knotted torus, respectively) are homeomorphic (even though it is not possible to deform one surface into the other without cutting or tearing while staying in $\mathbb{R}^3$). To construct a homeomorphism between the surfaces, consider Figure 2.3.4, in which the surface in part (i) of Figure 2.3.3 is cut along a 1-sphere, is knotted, and is finally re-glued. The map that takes each point in the unknotted torus and maps it to its final location after the cutting and re-gluing maneuver is the desired homeomorphism. Although the surface was cut during this construction so we do not have a continuous deformation, the resulting map is continuous since the surface is re-glued exactly where it was cut. The difference between the knotted and unknotted tori (plural for torus) is simply the way in which they sit in $\mathbb{R}^3$.

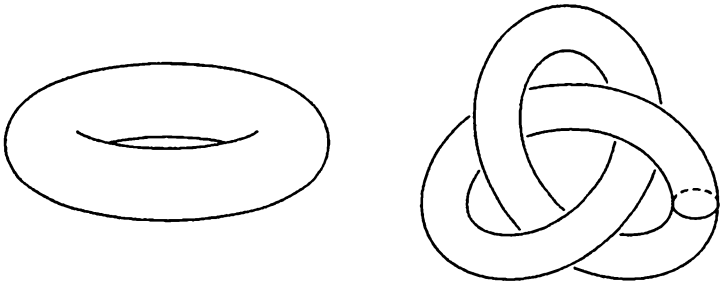
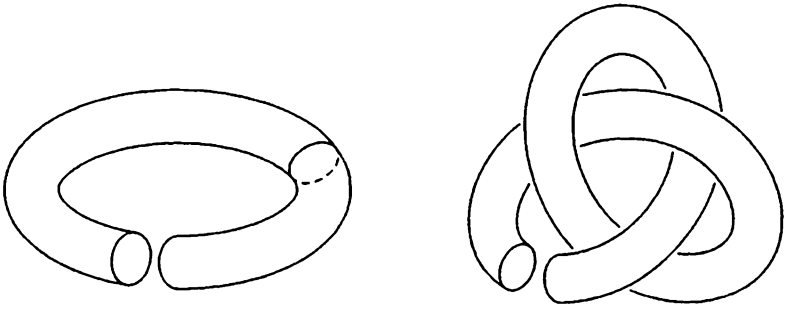


Figure 2.3.3



cut re-glue

Figure 2.3.4

Exercises

2.3.1. Which of the following are surfaces?

(i) S^2 – point,

(ii) $(a, b) \times (b, c) \subset \mathbb{R}^2$,

(iii) $(a, b) \times [c, d] \subset \mathbb{R}^2$,

(iv) $\big(D^2 \times \{0, 1\}\big) \cup \big(S^1 \times [0, 1]\big) \subset \mathbb{R}^3$.

2.3.2. Using reasoning of the sort used in Figure 2.3.4, which surfaces shown in Figure 2.3.5 are homeomorphic to one another?

2.3.3*. Let $Q \subset \mathbb{R}^n$ be a topological surface, and let $p \in Q$ be a point. Show that for any number $\epsilon > 0$ there are subsets $U, B \subset Q$ such that U is homeomorphic to int D^2 and contains p, the set B is a disk containing p in its interior, and $U, B \subset O_\epsilon(p, Q)$.

2.3.4. Let $Q \subset \mathbb{R}^n$ be a surface, and let $U \subset Q$ be a set that is homeomorphic to an open subset of $\mathbb{R}^2$. Show that U is open in Q.

2.3.5*. If $p, q \in S^2$ are any two distinct points, show that $S^2 - \{p\} \approx \mathbb{R}^2$ and $S^2 - \{p, q\} \approx S^1 \times \mathbb{R}$.

2.3.6. Show that a surface from which a closed subset has been removed is still a surface.

2.3.7*. Let $Q_1 \subset \mathbb{R}^n$ and $Q_2 \subset \mathbb{R}^m$ be surfaces, and let $B_i \subset Q_i$ be a disk for $i = 1, 2$. Suppose that $Q_1 - \text{int } B_1 \approx Q_2 - \text{int } B_2$. Show that $Q_1 \approx Q_2$.

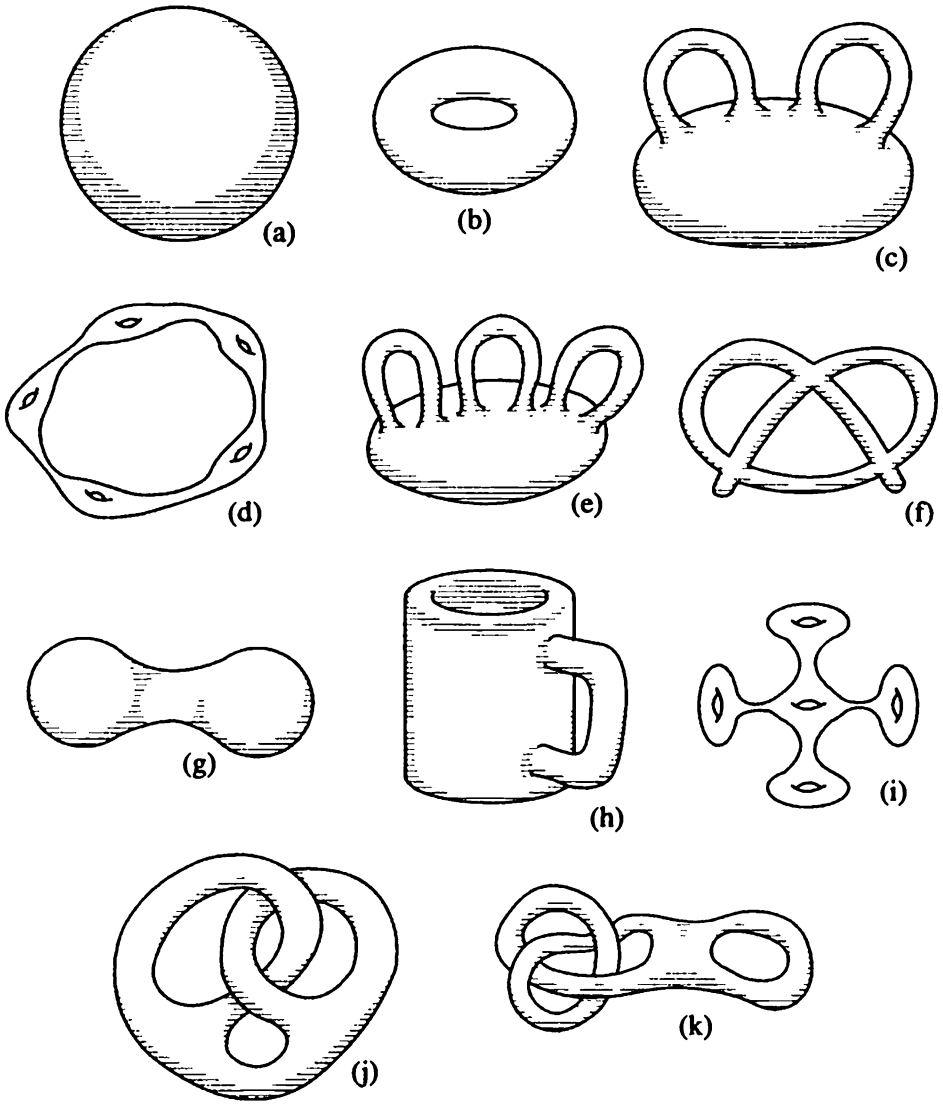


Figure 2.3.5

2.4 Surfaces via Gluing

Our next step is to develop a technique called *gluing* for constructing more surfaces. Recall the discussion of constructing a cylinder from a rectangular piece

of paper in Section 1.4. Though a compact cylinder is not strictly speaking a surface, the same idea of gluing can be used to obtain true surfaces. For example, starting with a disk, dividing its boundary into two semi-circles, and then gluing these semi-circles as in Figure 2.4.1 yields a sphere, albeit a somewhat "calzone-shaped" one. (In this and in other constructions, it is best to think of the surfaces as made out of cloth or rubber rather than paper.)

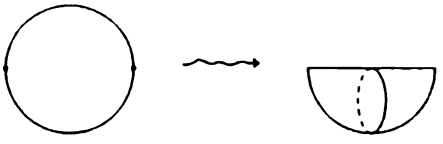


Figure 2.4.1

Now take a square, and label the sides as in Figure 2.4.2; the labeling indicates that sides labeled with the same letter are to be glued, with arrows matching up. Try to figure out what is obtained before reading on.

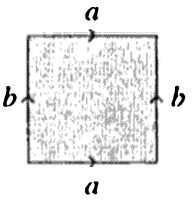


Figure 2.4.2

The easiest way to see what is obtained is to glue the edges in two stages. After gluing the edges labeled a, we obtain a cylinder. Notice that the arrows on the sides labeled b become arrows on the edges of the cylinder. We next glue the edges b, and the result is a torus. See Figure 2.4.3. A general procedure for this sort of construction is given in the following definition.

Figure 2.4.3

Definition. A **polygonal disk** is a disk that sits in some plane in $\mathbb{R}^n$, the boundary of which is a polygon. If D is a polygonal disk, a **gluing scheme** S for the edges of D is a labeling of each edge of D with an arrow and a letter, where each letter used in the labeling appears on precisely two edges.

For a polygonal disk to have a gluing scheme it must have an even number of edges. Some examples of gluing schemes appear in Figure 2.4.4. Observe that the gluing schemes in parts (i) and (ii) of the figure are not the same, since the direction of one of the arrows differs in the two figures. On the other hand, the gluing schemes in parts (i) and (iii) are essentially the same, since the arrows on both edges labeled a are reversed in (iii) as compared to (i). Although we have yet to give a formal definition of gluing the edges of a polygonal disk via a gluing scheme, intuitively the idea is just as in the case of gluing the edges of the square used above to obtain a torus. For example, the result of gluing the edges of Figure 2.4.4 (i) is seen in Figure 2.4.5.

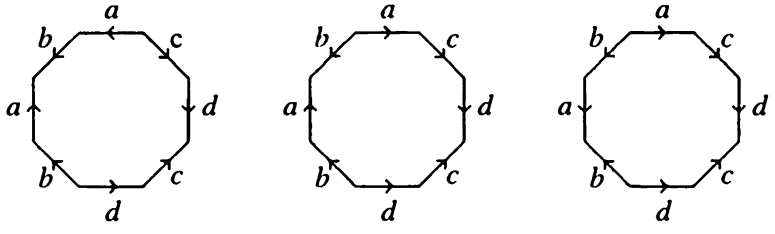


Figure 2.4.4

Figure 2.4.5

If we start with any polygonal disk and any gluing scheme for the edges of the disk, do we obtain a surface in some $\mathbb{R}^n$? To answer this question we need a rigorous definition of what it means for something to be the result of such a

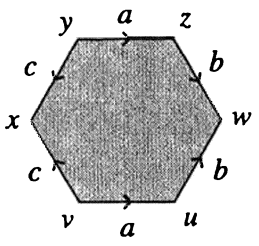
gluing. We use the concept of an identification space, as discussed in Section 1.4. Given a polygonal disk D and a gluing scheme S for the edges of D, we will construct a partition of D induced by the gluing scheme, and the result of gluing the edges of D according to a gluing scheme will simply be an identification space for D and this partition of D, whenever such an identification space exists.

Consider the gluing scheme shown in Figure 2.4.2, used to construct T^2. It glues all four vertices of the square to each other, glues the points in the interiors of the edges of the square in pairs, and does not glue the points in the interior of the square to anything. More generally, for any polygonal disk and gluing scheme, points in the interiors of the edges of the polygonal disk get glued in pairs and points in the interior of the polygonal disk do not get glued to anything; the vertices of the polygonal disk get glued to one another in collections of various sizes.

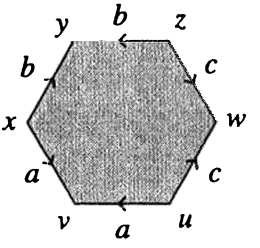
Definition. Let D be a polygonal disk, and let S be a gluing scheme for the edges of D. The gluing scheme S divides up the edges of D into pairs, called **edge-sets**, such that two edges are in the same edge-set iff they are identified under S; let $E_1, \ldots E_k$ denote the edge-sets. For each E_i, let L_{E_i} denote the unique affine linear map that takes one of the edges in E_i to the other so that their endpoints are matched up according to the arrows on the edges given by the gluing scheme (see Lemma A.7); there are two such maps, depending upon which of the two edges is the domain and which is the codomain, so choose one map. The **induced partition** of D by S, denoted $\mathcal{P}(S)$, is the partition of D given by $\mathcal{P}(L_{E_1}, \ldots, L_{E_k})$, using the notation of Exercise 1.4.8. The sets in this partition that contain vertices consist only of vertices, and these collections of vertices are called **vertex-sets**.

Example 2.4.1. See Figure 2.4.6 for two examples of gluing schemes on a polygonal disk with eight edges, and the associated vertex-sets. Note that the vertex-sets for the two different gluing schemes are quite different in their sizes, even though the same size polygonal disk was used in both cases, and both gluing schemes yield a 2-sphere. There is, in general, no way of knowing a priori the number and sizes of the vertex-sets, in contrast to the edge-sets — all of which contain two edges, and of which there are half as many as the number of edges of D. $\Diamond$

We can now finally state what it means rigorously to glue the edges of a polygonal disk by a gluing scheme.



vertex sets: {x}, {y, v}
{z, u}, {w}



vertex sets: {w}, {v}, {y}
{u, x, z}

Figure 2.4.6

Definition. Let D be a polygonal disk, and let S be a gluing scheme for the edges of D. A subset $X \subset \mathbb{R}^n$ is **obtained** from D and S if X is an identification space of D and $\mathcal{P}(S)$; that is, there is a quotient map $q: D \to Q$ such that if $x, y \in D$ are points, then $q(x) = q(y)$ iff x and y are in the same set in $\mathcal{P}(S)$.

Example 2.4.2. Consider the gluing scheme used to construct T^2. If we start with the square in Figure 2.4.2, we want to find a quotient map from this square onto the torus with the desired properties. The map from the square to the torus can be obtained simply by taking every point in the square and mapping it to where it ends up in the torus at the end of the gluing process shown in Figure 2.4.3. The fact that this map is a quotient map follows from the compactness of the square and the continuity of the map, using Proposition 1.6.14 (ii). The requirement on the inverse images of points under the map can be seen straightforwardly. Thus, the torus is indeed obtained from the square and the gluing scheme shown in Figure 2.4.2 according to the above definition. ◊

As mentioned in Section 1.4, it is not clear whether for any set in Euclidean space and any partition of the set there exists an identification space also sitting in some Euclidean space. We then ask whether there is some set in Euclidean space, not to mention a surface, obtained from every polygonal disk and every gluing scheme for the edges of the disk? Conversely, is every compact surface obtained from some polygonal disk and gluing scheme? The following result takes care of all these questions.

Theorem 2.4.3. (i) *Let D be a polygonal disk, and let S be a gluing scheme for the edges of D. Then there is a surface $Q \subset \mathbb{R}^n$ that is obtained from D and S.*
(ii) *Let $Q \subset \mathbb{R}^n$ be a compact connected surface. Then there is a polygonal disk D and a gluing scheme S for the edges of D such that Q is obtained from D and S.*

The rather lengthy proof of part (i) is given in Appendix 2A.1, to avoid interrupting the development of the material. The proof of part (ii) is delayed until Section 3.4, where we will have more tools at our disposal.

We conclude this section with some very important examples of surfaces constructed by gluing. First, consider a compact cylinder, obtained by gluing two opposite edges of a rectangle. Suppose we put in half a twist prior to gluing this time. The result will be the well-known Möbius strip, denoted M^2. See Figure 2.4.7 (i)–(ii). The Möbius strip is not a surface as we have defined it, though it is a "surface with boundary." The two unlabeled edges in the rectangle shown in Figure 2.4.7 (i) are glued end-to-end in the Möbius strip, where they form a 1-sphere, as in Figure 2.4.7 (iii); this 1-sphere is the boundary of the Möbius strip, and is denoted ∂M^2. We will use the term Möbius strip to refer to any subset of Euclidean space homeomorphic to the standard Möbius strip.

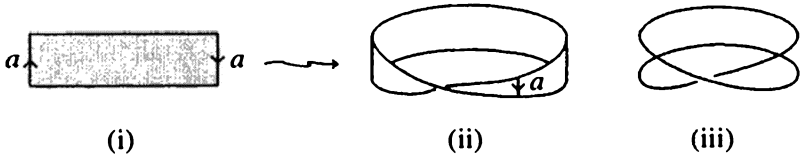


Figure 2.4.7

The Möbius strip has only "one side." In contrast to a cylinder, which can have one side painted red and the other side painted black, if we start painting anywhere on a Möbius strip we eventually cover everywhere on the surface with that single color. Actually, surfaces don't really have "sides", since they have no thickness. Imagine a Möbius strip with no thickness (and thus transparent), on which there is a rightward-facing person, as in Figure 2.4.8. If the person went all the way around the Möbius strip, she would come back facing left. Such a reversal could never happen on a cylinder. (Whereas the properties of one-sidedness and figure-reversing coincide for surfaces sitting in $\mathbb{R}^3$, they need not coincide in more general circumstances; see [WE, Chapter 8] for a very nice

discussion.) While you are at it, take a cylinder and a Möbius strip and cut them down the middle lengthwise; before cutting each object try to figure out how many pieces you will obtain.

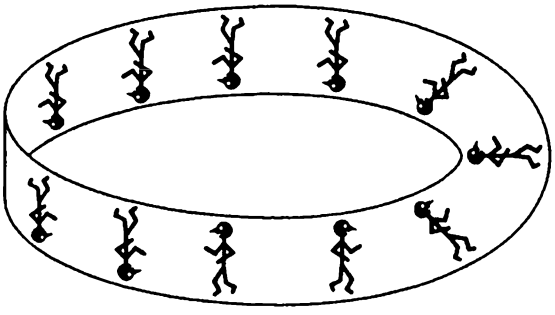


Figure 2.4.8

We can obtain surfaces that contain Möbius strips as follows. Consider the two gluing schemes for squares shown in Figure 2.4.9. Inside each of the surfaces obtained by these gluing schemes sits a Möbius strip, since inside each square sits a strip (shaded in the figure) with opposing edges glued appropriately.

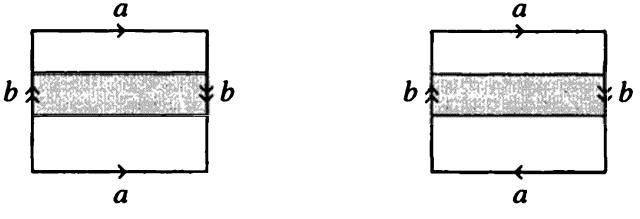


Figure 2.4.9

If we glue the sides labeled a in Figure 2.4.9 (i) we get a cylinder, as in Figure 2.4.10 (i). Observe that the sides labeled b in Figure 2.4.10 (i) have arrows facing in opposite directions. In order to glue these sides so that their arrows match we need to pass one end of the tube through its side, as in Figure 2.4.10 (ii) and (iii). It is, of course, not possible to pass a physical object through

itself, though mathematically there is really no problem with the concept of self-intersection. (Still, it is nicer if self-intersections can be avoided, since an object with self-intersections is not a surface in Euclidean space as we have defined it.) We can also get around the problem of this surface passing through itself by going into $\mathbb{R}^4$. Draw a 1-sphere around the part of the surface where the self-intersection occurs, labeled C in Figure 2.4.10 (iii), and then push the interior of the disk bounded by the 1-sphere "up" into $\mathbb{R}^4$. Such a move gets the disk out of the way, and thus there is no self-intersection. (If you have not thought about four-dimensional space previously such a maneuver may seem somewhat baffling, but it really works.) This argument about placing the surface in $\mathbb{R}^4$ in such a way that it has no self-intersections can be made rigorous, though we will not take the trouble here. The result of this process yields a surface known as the Klein bottle, Figure 2.4.10 (iii), denoted K^2. As usual, the term Klein bottle will apply to any subset of Euclidean space homeomorphic to the standard Klein bottle.

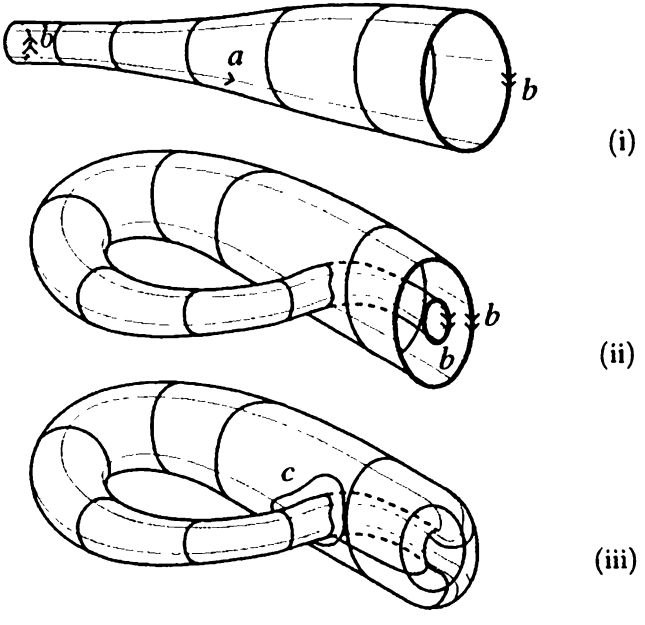


Figure 2.4.10

The surface obtained by the gluing indicated in Figure 2.4.9 (ii) is a bit trickier to visualize; it also needs to be placed in $\mathbb{R}^4$ to avoid self-intersection.

For convenience, we rotate the original square as in Figure 2.4.11 (i), and label the corners as shown; observe that if the edges are glued as indicated then the points labeled A and A' will be glued to one another, as will the points labeled B and B'. We start by simply gluing these pairs of points, yielding Figure 2.4.11 (ii). Look closely at how the edges now need to be glued in pairs. We can certainly glue one of the pairs of edges, say those labeled a, so that their arrows match. However, to glue the other pair of edges one would have to pass the surface through itself if we stayed in $\mathbb{R}^3$, yielding something like Figure 2.4.11 (iii); in $\mathbb{R}^4$ the self-intersection can be avoided. This surface is known as the projective plane, denoted P^2. We could obtain P^2 by gluing the boundary of a disk as shown in Figure 2.4.12.

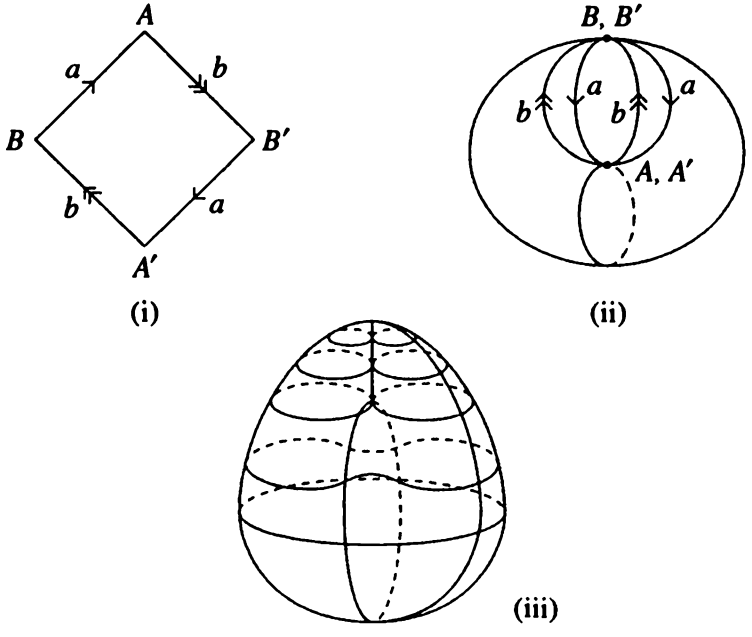


Figure 2.4.11

The surfaces K^2 and P^2 will play important roles in our study of surfaces. We noted before that both K^2 and P^2 contain Möbius strips; we can now make more precise the nature of these inclusions. If we take two disks of the same size made of cloth and glue their boundaries together, we obtain sphere; this notion of gluing can be made rigorous by using the notion of attaching via a

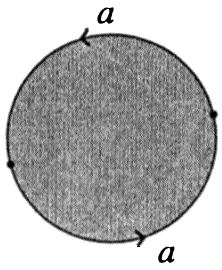


Figure 2.4.12

homeomorphism discussed in Section 1.4, where the homeomorphism in this case is any homeomorphism between the boundaries of the two disks. Given that the boundary of a Möbius strip is a 1-sphere, as is the boundary of a disk, we ask what would happen if we glue the boundary of a Möbius strip to the boundary of a disk, and what would happen if we glue the boundaries of two Möbius strips together? The next two lemmas answer these questions.

Lemma 2.4.4. *Let $B \subset P^2$ be a disk; then $P^2 - \text{int } B \approx M^2$. Thus, P^2 can be obtained by attaching a Möbius strip and a disk via a homeomorphism of their boundaries.*

Proof. The second sentence in the lemma follows straightforwardly from the first. For the first sentence, we start by observing that Proposition A2.2.6 implies that the choice of disk B makes no difference, so we can chose a disk that is convenient. A standard method of proof uses a cutting and pasting method, as pictured in Figure 2.4.13. This procedure starts with the disk shown in Figure 2.4.12; a smaller disk in the interior is chosen, and after removing the interior of the inner disk some rearrangement is done until we end up with a Möbius strip. □

Lemma 2.4.5. *The Klein bottle can be obtained by attaching two Möbius strips via a homeomorphism of their boundaries.*

Proof. A pictorial proof of this result can be given by cutting an appropriate model of K^2 in half, as shown in Figure 2.4.14. A proof using cutting and pasting, similar to the proof of the previous lemma, is left to the reader in Exercise 2.4.2. □

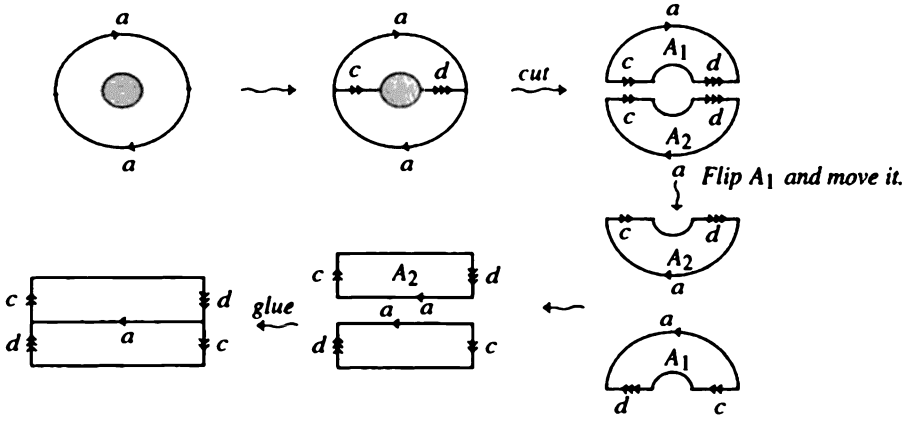


Figure 2.4.13

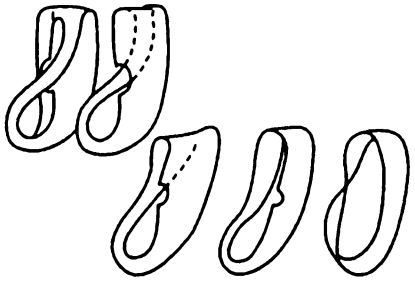


Figure 2.4.14

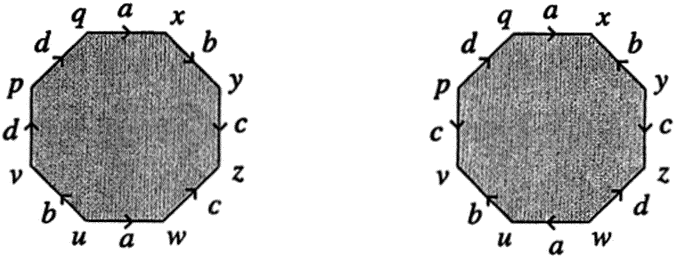
Exercises

2.4.1. Give a polygonal disk and a gluing scheme that will yield the surface pictured in Figure 2.4.15.

Figure 2.4.15

2.4.2*. Prove Lemma 2.4.5.

2.4.3. Find the vertex-sets for each of the polygonal disks and gluing schemes shown in Figure 2.4.16.



(i) (ii)

Figure 2.4.16

2.5 Properties of Surfaces

We discuss a number of properties, some familiar and some new, which a given surface may or may not have. One of the most important of these properties is compactness. The surfaces S^2, T^2, K^2 and P^2 are all compact, whereas $\mathbb{R}^2$ is not compact. In fact, any surface obtained by gluing the edges of a polygonal disk will be compact, since the surface is the image of a compact set (the polygonal disk) under a continuous map (the quotient map from the polygonal disk to the surface). For the most part we will restrict our attention to compact surfaces, since non-compact surfaces can be much more complicated topologically, as in Figure 2.5.1.

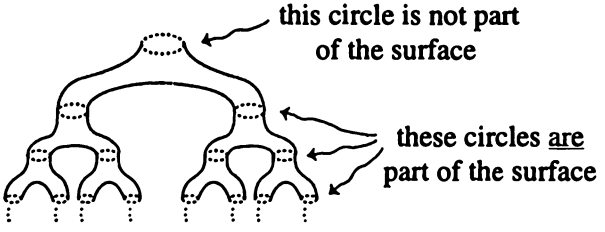


Figure 2.5.1

We will also need to apply the notion of connectedness to surfaces. In general, a surface need not be connected, for example, the union of two spheres that do not touch. Whereas in Section 1.5 we saw that the concepts of connectedness and path connectedness do not coincide in general, we now see that they do coincide for surfaces.

Proposition 2.5.1. *A surface in $\mathbb{R}^n$ is connected iff it is path connected.*

Proof. That a path connected surface is connected follows from Theorem 1.5.7. Now assume that $Q \subset \mathbb{R}^n$ is a connected surface; we will show that Q is path connected. Pick any point $p \in Q$. Let $A \subset Q$ be the set defined by

$$A = \{q \in Q \mid \text{there is a path in } Q \text{ from } p \text{ to } q\}.$$

We will show that A is both open and closed in Q; since A is non-empty (it contains p) it will then follow from the connectedness of Q and Lemma 1.5.1 that $A = Q$. Hence Q must be path connected, since for any two points $q_1, q_2 \in Q$ we can find paths from p to each of q_1 and q_2, and we can then apply Exercise 1.5.10.

To show that A is both open and closed, we start with the following observation. For any point $q \in Q$ there is an open subset $W \subset Q$ containing q that is homeomorphic to the open disk int D^2. By Exercise 1.5.7 int D^2 is path connected, and hence W is path connected by Exercise 1.5.9. Thus any point in Q is contained in an open subset of Q that is path connected.

We now return to the set A. Let q be any point in A, and let $W \subset Q$ be a path connected open set containing q. Since there is a path from p to q, and since there is a path from q to any point in W, it follows that there is a path from p to any point in W. Hence $W \subset A$. Since this result holds for any $q \in A$, it follows from Exercise 1.2.6 that A is open in Q.

Now let $s \in Q - A$ be a point, so that there is no path from p to s. Let $V \subset Q$ be a path connected open set containing s. If any point in V were connected by a path to p, then since that point is connected by a path to s, it would follow that there is a path from p to s, a contradiction. Thus $V \subset Q - A$. It follows that $Q - A$ is an open subset of Q, and therefore A is closed in Q. □

Another property a surface may possess concerns the difference mentioned in the previous section between the cylinder and the Möbius strip. Rather than try to characterize more rigorously what this difference is (as is done in more advaced treatments), we observe that if a surface contains a Möbius strip

then it will certainly have the direction-reversing property of the Möbius strip. Conversely, it seems plausible that if a surface has this reversing property then it would contain a Möbius strip, and we are thus led to the following definition.

Definition. A surface is **orientable** if it does not contain a Möbius strip, and it is **non-orientable** if it does contain a Möbius strip.

We saw in the previous section that K^2 and P^2 contain Möbius strips, and are thus non-orientable. On the other hand $\mathbb{R}^2$, S^2 and T^2 can all be shown to be orientable. One way to see this fact intuitively is that each of these surfaces can be colored with one color on one side and a different color on the other side. That being the case, none of them could contain a Möbius strip, since otherwise a Möbius strip in one of these surfaces would inherit this coloring. A more rigorous demonstration that these three surfaces are orientable would require a more advanced definition of orientability. Actually, any surface in $\mathbb{R}^3$ that is a closed subset of $\mathbb{R}^3$ is orientable (see [SA]); surfaces in higher-dimensional $\mathbb{R}^n$ need not be orientable.

An issue we have already touched on is the distinction between a property inherent in a surface and one that is dependent upon how the surface sits in Euclidean space; a property of the former type is called *intrinsic*, whereas the latter type is called *extrinsic*. The compactness of a surface is intrinsic to the surface and does not depend upon how the surface sits in Euclidean space, since if two surfaces in Euclidean space are homeomorphic, then either both are compact or neither are. We also saw the example of the knotted and unknotted tori in Figure 2.3.3. These surfaces are homeomorphic as mentioned, so knottedness in an extrinsic property. In fact, if the knotted torus is placed in $\mathbb{R}^4$, it can be continuously deformed into an unknotted torus, so we have even more evidence that the issue of knottedness only depends upon how a surface is sitting in a certain Euclidean space. Another way to express the extrinsic nature of knottedness is to image a bug that lives on a torus and that cannot see anything off of the torus (and in particular cannot look through three-dimensional space from one part of the torus to another, no matter how close the other part of the torus is). Such a bug would not be able to tell if the torus it is on were knotted or not.

Exercises

2.5.1. Which of the following surfaces are compact?

(1) $S^2 - \{\text{point}\}$;

(2) $S^1 \times \mathbb{R}$;

(3) the surface in Figure 2.5.2.

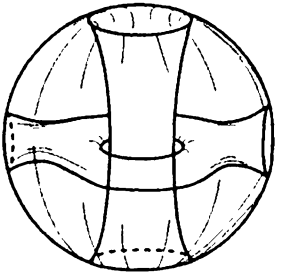


Figure 2.5.2

2.5.2*. Prove that an orientable surface cannot be homeomorphic to a non-orientable surface.

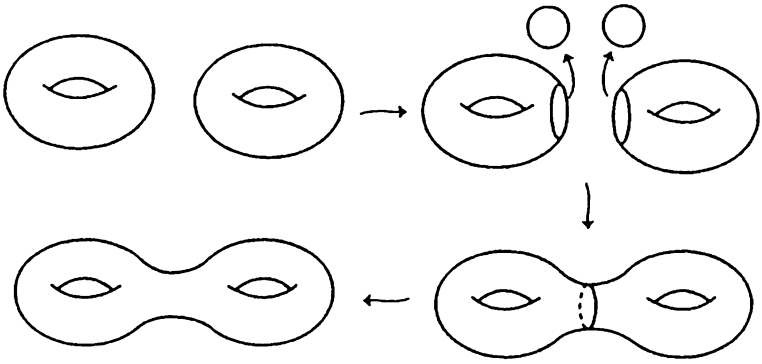
2.5.3*. Let $Q \subset \mathbb{R}^n$ be a connected surface. Suppose that each point in Q has an open neighborhood in Q that is contained in a plane in $\mathbb{R}^n$. Show that Q is contained in a plane. Show that the analogous result with spheres replacing planes also holds.

2.6 Connected Sum and the Classification of Compact Connected Surfaces

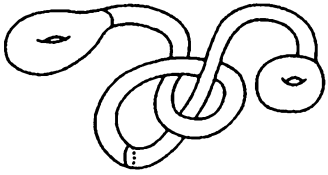
Our present goals are to make a complete list of compact connected surfaces up to homeomorphism and to find an easy method for distinguishing between such surfaces. It is not obvious that this can be accomplished. Some surfaces can appear to be quite complicated, as in Figure 2.5.2. Additionally, as seen in Figure 2.3.3 and Exercise 2.3.2, some surfaces that may appear distinct are in fact simply sitting differently in Euclidean space.

In order to state our main result we first need to introduce a systematic method for constructing new surfaces out of old ones. The idea is to take two surfaces, remove the interior of a small disk from each one, and then glue what remains as shown in Figure 2.6.1 (i). Equivalently, we could remove the interiors

of the two disks, and run a tube from one hole to the other (the result is the same up to homeomorphism); see Figure 2.6.1 (ii). The result of this construction certainly appears to be a surface, and thus we can make new surfaces in this way. To define this construction rigorously we use the notion of attaching via a homeomorphism, as discussed in Section 1.4. Observe that a disk can always be found in any surface, since a disk can always be found inside int D^2.



(i)



(ii)

Figure 2.6.1

Definition. Let $Q_1, Q_2 \subset \mathbb{R}^n$ be compact connected surfaces. For $i = 1, 2$ let $B_i \subset Q_i$ be a disk, and let $h: \partial B_1 \to \partial B_2$ be a homeomorphism. The **connected sum** of Q_1 and Q_2, denoted $Q_1 \# Q_2$, is the the attaching space $(Q_1 - \text{int } B_1) \cup_h (Q_2 - \text{int } B_2)$.

A number of questions arise here. First, as mentioned in Section 1.4, it is not clear that we can always find a subset of Euclidean space that is the result of any given attempt at attaching. Second, even if we can always form an attaching

space $(Q_1 - \text{int } B_1) \cup_h (Q_2 - \text{int } B_2)$, how do we know that we always obtain a surface? Finally, supposing that we do always obtain a surface, what about the choices we made in the construction, namely the choice of disks $B_i \subset Q_i$ and the choice of homeomorphism h; if for different choices we were to obtain different surfaces, then our construction would not be well-defined. Fortunately, everything works out as well as possible.

Proposition 2.6.1. *Let $Q_1, Q_2 \subset \mathbb{R}^n$ be compact connected surfaces. Let $B_i \subset Q_i$ be a disk for $i = 1, 2$ and let $h: \partial B_1 \to \partial B_2$ be a homeomorphism. Then the attaching space $(Q_1 - \text{int } B_1) \cup_h (Q_2 - \text{int } B_2)$ exists and is a surface in some $\mathbb{R}^n$. Any two surfaces obtained in this way are homeomorphic.*

The cleanest way to prove this result uses more advanced techniques; see [RO], [HE] and [MI1]. We give an accessible, though somewhat involved, proof of this proposition in Appendix A2.1.

The following lemma gives a few properties of connected sum, which we state without proof; the first two properties are straightforward, and the third follows from Exercise 2.3.6.

Lemma 2.6.2. *Let A, B and C be compact connected surfaces. Then*

(i) $A \# B \approx B \# A$,
(ii) $(A \# B) \# C \approx A \# (B \# C)$,
(iii) $A \# S^2 \approx A$.

These properties of connected sum make it appear as if connected sum acts analogously to addition and multiplication of numbers. However, unlike those two operations, there are no inverses with respect to connected sum (with S^2 playing the role of the identity element), as seen in the following proposition. The proof of the proposition involves a method known as the "Mazur swindle."

Proposition 2.6.3. *Let A and B be compact connected surfaces such that $A \# B \approx S^2$. Then $A \approx B \approx S^2$.*

Proof. Consider the infinite connected sum

$$X = A \# B \# A \# B \# A \# B \# \ldots .$$

(To be completely rigorous we would need to define the notion of infinite series using connected sums, though we will not go into that here.) Because connected

sum is associative, we can regroup the summands in X, giving us two different computations of the value of X. First, we have

$$X = (A \# B) \# (A \# B) \# (A \# B) \# \ldots \approx S^2 \# S^2 \# \ldots \approx S^2.$$

On the other hand, we also have

$$X = A \# (B \# A) \# (B \# A) \# \ldots \approx A \# S^2 \# S^2 \# \ldots \approx A.$$

Combining these two calculations for the value of X we see that $A \approx S^2$; similar reasoning shows that $B \approx S^2$ as well. □

Two useful examples of connected sums are the following lemmas.

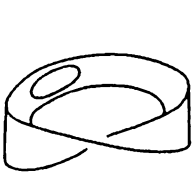
Lemma 2.6.4. $P^2 \# P^2 \approx K^2$.

Proof. Exercise 2.6.1. □

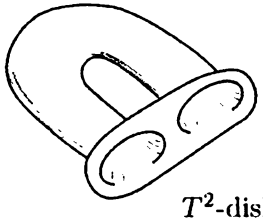
Lemma 2.6.5. $P^2 \# T^2 \approx P^2 \# P^2 \# P^2$.

Proof. The previous lemma shows that it suffices to prove that $P^2 \# T^2 \approx P^2 \# K^2$. Suppose that for some disk $B \subset P^2$ we could show that $(P^2 - \text{int } B) \# T^2 \approx (P^2 - \text{int } B) \# K^2$; the result would then follow using Exercise 2.3.7. Using Lemma 2.4.4, it thus suffices to prove that $M^2 \# T^2 \approx M^2 \# K^2$. (We have not discussed connected sums involving a non-surface such as M^2, but as long as we stay away from ∂M^2 there is no problem.) The reason we go from P^2 to the M^2 strip is to make the proof visualizable.

To form the connected sum of M with each of T^2 and K^2, we need to know what all these objects look like with the interior of a disk cut out; by Proposition 2.6.1 we can use the disks of our choice. Cutting out the interior of a disk from M^2 leaves a Möbius strip with a hole cut out — not very exciting. If we cut the interior of a disk out of T^2 we can deform what remains as in Figure 2.6.2.



M^2-disk



T^2-disk

Figure 2.6.2

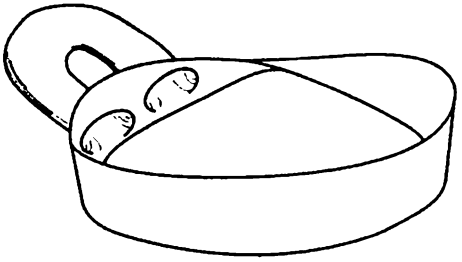


Figure 2.6.3

We can now form $M^2 \# T^2$ by attaching the appropriate 1-spheres as in Figure 2.6.3.

Turning to K^2, we may as well cut out the interior of a disk that is convenient to visualize, as in Figure 2.6.4. We form $M^2 \# K^2$ as in Figure 2.6.5. To show that $M^2 \# T^2 \approx M^2 \# K^2$, we need to see that the objects in Figures 2.6.3 and 2.6.5 are homeomorphic. This homeomorphism is demonstrated in Figure 2.6.6. □

Figure 2.6.4

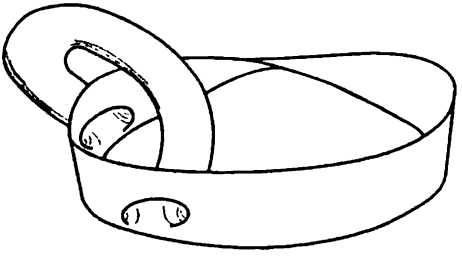


Figure 2.6.5

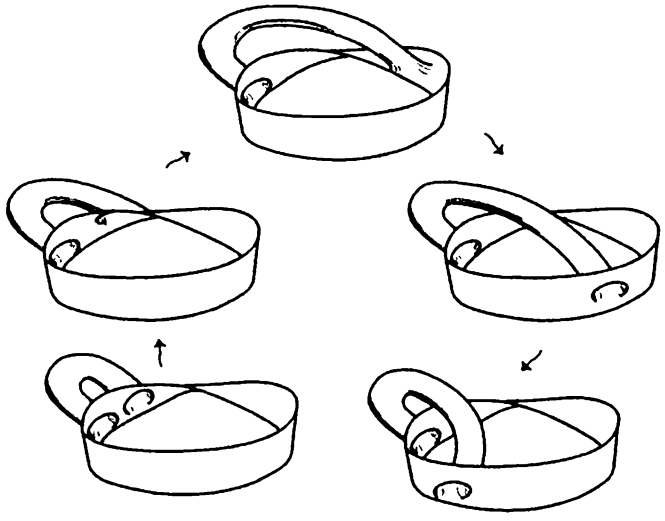


Figure 2.6.6

Lemma 2.6.5 shows that the operation # has no cancelation property. The following proposition clarifies the relation of orientability and connected sum. The proof of the proposition is trickier than might be expected, and we will have to gloss over one technical detail (a more satisfying proof would make use of algebraic topology).

Proposition 2.6.6. *Let Q_1 and Q_2 be compact connected surfaces in $\mathbb{R}^n$. Then $Q_1 \# Q_2$ is orientable iff both Q_1 and Q_2 are orientable.*

Proof. We prove the proposition by showing that the falseness of either statement implies the falseness of the other. Assume first that one of Q_1 or Q_2 is non-orientable; without loss of generality assume it to be Q_1. Hence Q_1 contains a Möbius strip, denoted M. Since M is not a surface by itself, the surface Q_1 must contain some point q not in M. Using the compactness of M and $\{q\}$ and Exercises 1.6.11 and 2.3.3, it follows that there is a disk $B \subset Q_1 - M$. By using the disk B in the construction of $Q_1 \# Q_2$ (which we are at liberty to do by Proposition 2.6.1), it follows that $M \subset Q_1 \# Q_2$. Thus $Q_1 \# Q_2$ is non-orientable.

Now assume that $Q_1 \# Q_2$ contains a Möbius strip M; we will show that one of Q_1 or Q_2 contains a Möbius strip. When the connected sum of Q_1 and Q_2 is formed, the interior of a disk was cut out of each surface, and the surfaces were attached along the boundary 1-sphere in each of the surfaces. After the

attaching, there remains on the connected sum a 1-sphere C where the two boundary circles were attached. Evidently this 1-sphere C separates $Q_1 \# Q_2$ into two pieces, one piece formerly from Q_1 and the other piece formerly from Q_2. If M does not intersect C then it is entirely contained in either Q_1 or Q_2, in which case the proof will be complete; hence assume that M intersects C. By using some technicalities we will not go into, it can be shown that M can be deformed so that every time it intersects C it actually crosses it, and does so in finitely many places; see Figure 2.6.7. Label these intersections $I_1, \ldots, I_r$ in order along M. Observe that r must be an even integer.

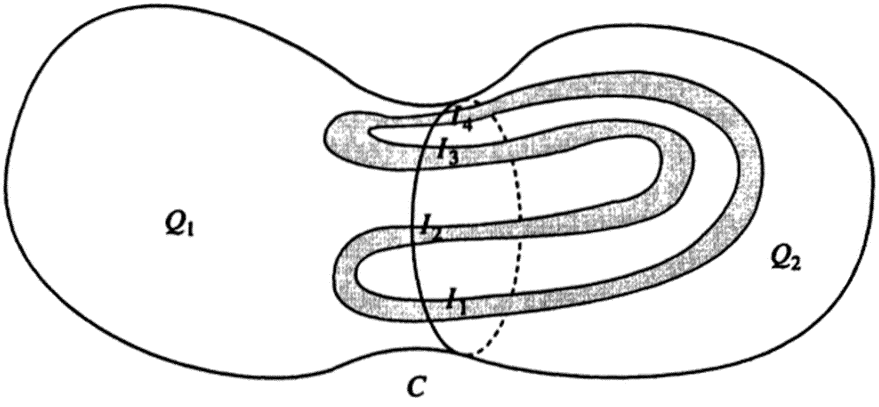


Figure 2.6.7

Take an arrow along C at I_1 whose width is the width of M, as in Figure 2.6.7. Think of running the arrow along the length of M, starting and ending at I_1, and passing through every other intersection I_k along the way exactly once. Because M is a Möbius strip, if one runs the arrow along the length of M it will be pointing the other way along C when it reaches I_1 again. As we run the arrow along M, each time it passes through an intersection I_k it points in one of two possible directions along C. For each $k \in \{1, \ldots, r\}$, as we go from intersection I_k to intersection I_{k+1} (where addition is mod r), the arrow either reverses direction along C or it does not. See Figure 2.6.8.

Suppose that every time we go from I_k to I_{k+1} the arrow reversed direction along C. Since r is an even number, it would follow that after going along all of M the arrow would end up not being reversed, a contradiction. Hence it must be the case that for some value of $j \in \{1, \ldots, r\}$ the transition from I_j to I_{j+1}

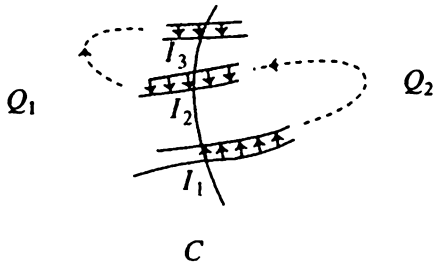


Figure 2.6.8

does not reverse the direction of the arrow along C. See Figure 2.6.9. The part of M between I_j and I_{j+1} is entirely contained in one of Q_1 or Q_2; without loss of generality assume it is in Q_1. Let N be a strip in Q_1 running along C between I_j and I_{j+1}, as in Figure 2.6.9 (it does not matter which such strip we choose). Let M' be the union of N and the part of M between I_j and I_{j+1}. By construction M' is entirely contained in Q_1, and it is seen that M' is a Möbius strip. □

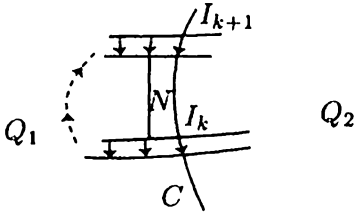


Figure 2.6.9

We now return to our main problem, namely finding all compact connected surfaces. Using connected sums there appear to be infinitely many different surfaces, since, for example, we can take the connected sum of an arbitrary number of tori. See Figure 2.6.10. The surfaces obtained this way are indeed all distinct, as seen in the following theorem. We will have to wait until Section 3.6 to prove this result.

Theorem 2.6.7 (Classification of Compact Connected Surfaces). *Any compact connected surface is homeomorphic to the sphere, a connected sum of tori,*

or a connected sum of projective planes, that is, is one member of the list

$$S^2$$
$$T^2,\ T^2 \# T^2,\ T^2 \# T^2 \# T^2,\ \ldots$$
$$P^2,\ P^2 \# P^2,\ P^2 \# P^2 \# P^2,\ \ldots .$$

The surfaces in this list are all distinct.

Figure 2.6.10

This remarkable theorem tells us that we know what all the compact connected surfaces are; simple criteria for distinguishing between such surfaces will be found during the course of the proof of the classification theorem. At first glance it appears as if we are missing the Klein bottle from the list in the theorem; Lemma 2.6.4 indicates where to find the Klein bottle in the list.

Exercises

2.6.1*. Prove Lemma 2.6.4.

2.6.2. Where on the list in Theorem 2.6.7 are the surfaces $K^2 \# P^2$, $K^2 \# K^2$ and $T^2 \# T^2$? Where is the surface shown in Figure 2.6.11?

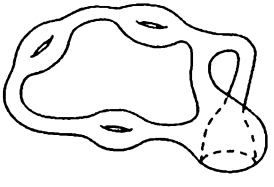


Figure 2.6.11

2.6.3. Show that none of the surfaces in the list

$$S^2,\ T^2,\ T^2 \# T^2,\ T^2 \# T^2 \# T^2,\ \ldots$$

is homeomorphic to any of the surfaces in the list

$$P^2,\ P^2 \# P^2,\ P^2 \# P^2 \# P^2, \ldots .$$

Appendix A2.1 Proof of Theorem 2.4.3 (i)

We start with some preliminaries. Recall the definition of edge-sets and vertex-sets in Section 2.4. The following lemma, though seemingly simple, is the heart of the proof of Theorem 2.4.3 (i).

Lemma A2.1.1. *Let D be a polygonal disk and let S be a gluing scheme for the edges of D.*

(i) *Let v be a vertex of D. The vertex-set containing v is the single-element set $\{v\}$ iff both edges of D containing v are glued to one another by the gluing scheme S.*

(ii) *If a vertex-set contains k vertices ($k \geq 1$), the vertices in the vertex-set can be labeled as $w_1, \ldots, w_k$ such that for each $i = 1, \ldots, k$ one of the edges containing w_i is identified by the gluing scheme with one of the edges containing w_{i+1} (where the addition is mod k).*

Proof. (i). This is straightforward, and we leave the details to the reader.

(ii). The proof is by induction on k. The case $k = 1$ follows from part (i). Now suppose that $k > 1$ and that the result holds true for all vertex-sets with fewer than k vertices in all polygonal disks and for all gluing schemes. Let W be a vertex-set of D with k members; let v be a vertex in this vertex-set. We now take the polygonal disk D, cut out a wedge containing v and the two edges adjacent to v, and close up the wedge to form a new polygonal disk D_1 with two fewer edges than D. See Figure A2.1.1.

We define a gluing scheme S_1 for D_1 as follows. From part (i) of this lemma it follows that the two edges of D containing v are not identified with one another under S; suppose these edges are labeled a and b. Let S_1 be defined by using S on all edges labeled other than a and b, and identifying the other edges of D_1 labeled a and b with one another with their given arrows. Let W_1 be the vertex-set of D_1 and S_1 containing all the vertices in W other than v. Since W_1 has $k - 1$ members, the induction hypothesis holds with respect to

W_1. Hence we can label the vertices of W_1 as $w_1, \ldots, w_{k-1}$ such that for each $i = 1, \ldots, k-1$ one of the edges containing w_i is identified by the gluing scheme with one of the edges containing w_{i+1}. Without loss of generality we could choose the labeling of the vertices so that w_{k-1} is contained in the edge of D_1 labeled a and w_1 is contained in the edge of D_1 labeled b. If we set $w_k = v$ it is not hard to see that the labeling $w_1, \ldots, w_k$ of the vertices of W works as desired. □

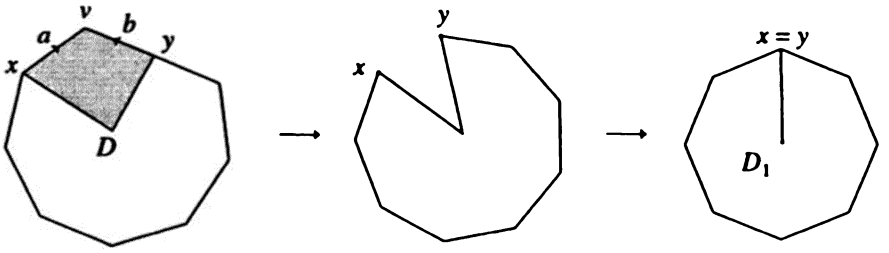


Figure A2.1.1

We are now ready for our main proof.

Proof of Theorem 2.4.3 (i). The outline of the proof is as follows. First, we will construct maps from various parts of the disk D onto certain disks in $\mathbf{R}^2$. We will then use these maps to construct a continuous map H from D into some $\mathbf{R}^m$, where m depends upon the number of edges of D and the gluing scheme S. Next, we will show that $H(x) = H(y)$ for points $x, y \in D$ iff x and y are in the same set in $\mathcal{P}(S)$. Finally, we will show that the image of this map is a surface. Since the disk D is compact and the map is continuous, it will then follow from Proposition 1.6.14 (ii) that the map is a quotient map onto its image. This outline is admittedly vague, and indeed this whole proof is the sort where one simply has to make sure that each step is logical, taking it on faith that by the end one will actually end up where expected — as indeed turns out to be the case.

Suppose that D has n edges; recall that the n must be an even integer. We may assume without loss of generality that $n \geq 4$; if not there must be precisely two edges, in which case one can divide each edge in two, and the appropriately defined gluing scheme for the divided edges will yield the same result topologically. See Figure A2.1.2 for the two possible cases with two

edges. (The theorem is true with $n = 2$, but the proof is simpler assuming $n \geq 4$.) Next, we may also assume without loss of generality that D is the regular polygonal disk with n sides centered at the origin in $\mathbb{R}^2$ with inscribed radius 1. See Figure A2.1.3. It is easy to calculate that each side of D has length $2 \tan \frac{\pi}{n}$; for convenience we will let A_n denote half this length, that is $A_n = \tan \frac{\pi}{n}$.

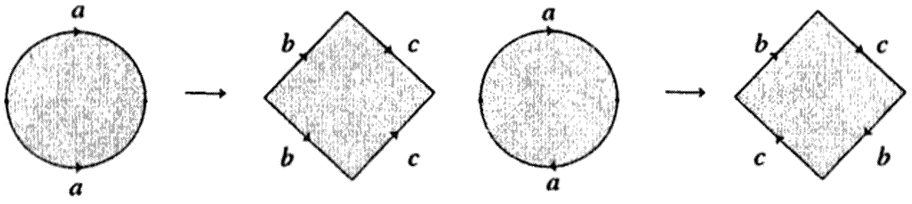


Figure A2.1.2

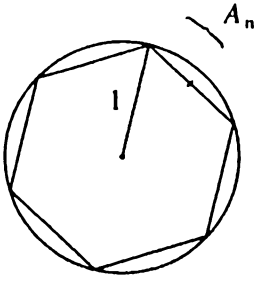


Figure A2.1.3

Let E be an edge-set of D with respect to gluing scheme S. Let a and a' be the edges in E, and let p, p' be their respective midpoints. We define the set U_E to be

$$U_E = O_{A_n}(p, D) \cup O_{A_n}(p', D).$$

The set U_E is the union of two half-disks with parts of their boundaries. See Figure A2.1.4. We construct a map $g_E: U_E \to O_{A_n}(O_2, \mathbb{R}^2)$ by mapping the half-disk $O_{A_n}(p, D)$ rigidly onto the upper half of the disk $O_{A_n}(O_2, \mathbb{R}^2)$, and mapping the half-disk $O_{A_n}(p', D)$ rigidly onto the lower half of the disk $O_{A_n}(O_2, \mathbb{R}^2)$, making sure that the boundaries of the two half-disks, match up as prescribed

by the gluing scheme S (it may be necessary to flip over one of the half-disks). It is seen by the construction that the map g_E is injective everywhere except on a and a' (where it is two-to-one), and that $g_E(x) = g_E(y)$ for any two points $x, y \in U_E$ iff x and y are in the same set in $\mathcal{P}(S)$. The image of g_e is $O_{A_n}(O_2, \mathbb{R}^2)$. There is a map g_E for each edge-set E.

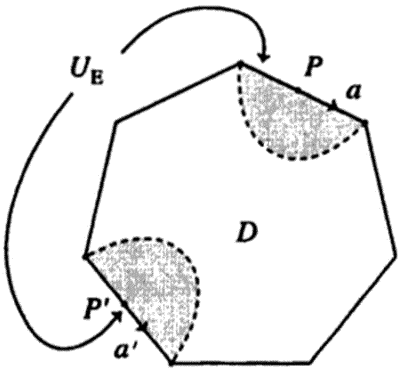


Figure A2.1.4

Now let W be a vertex-set of D with respect to S; W contains one or more vertices. We define sets U_W analogously to the sets U_E, namely

$$U_W = \bigcup_{w \in W} O_{A_n}(w, D).$$

The set U_W is the union of a collection of wedge-shaped pieces. See Figure A2.1.5. We wish to construct a map $g_W: U_W \to O_{A_n}(O_2, \mathbb{R}^2)$ with properties analogous to the maps g_E. Suppose that there are k vertices in W. If $k = 1$, that is, if there is one vertex $w \in W$, then the map g_W is defined by mapping w to the origin, and wrapping the single wedge $U_W = O_{A_n}(w, D)$ around until its two edges overlap and it covers $O_{A_n}(O_2, \mathbb{R}^2)$.

Now assume $k \geq 2$. Begin by dividing the disk $O_{A_n}(O_2, \mathbb{R}^2)$ into k equal wedges, labeled $F_1, \ldots, F_k$, as in Figure A2.1.6. Label the vertices in W as $w_1, \ldots, w_k$, as in the conclusion of Lemma A2.1.1 (ii). The map g_W is defined so that it takes all the vertices w_i to the origin, and it takes each wedge $O_{A_n}(w_i, D)$ onto the wedge F_i (taking edges to edges affine linearly, possibly squeezing or stretching the wedges). There are actually two ways we could define the map g_W on each $O_{A_n}(w_i, D)$, depending on whether we flip the wedge over or not; having arbitrarily chosen how to map $O_{A_n}(w_1, D)$, then by

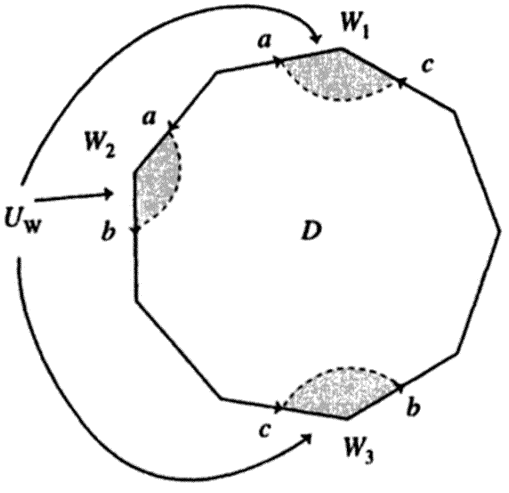


Figure A2.1.5

Lemma A2.1.1 (ii) the map is determined on the rest of the wedges if we want to insure that edges that are glued together by the gluing scheme get mapped by g_W to the same image. See Figure A2.1.6. It is seen by the construction that the map g_W is injective everywhere except on the vertices and edges of the wedges $O_{A_n}(w_i, D)$, and that $g_W(x) = g_W(y)$ for any two points $x, y \in U_W$ iff x and y are in the same set in $\mathcal{P}(S)$. The image of g_W is $O_{A_n}(O_2, \mathbb{R}^2)$. There is a map g_W for each vertex-set W.

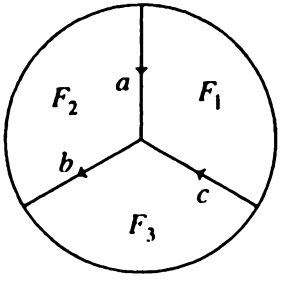


Figure A2.1.6

For the sake of maintaining the analogy (which will make things simpler notationally later on) we define U_D to be the disk int D^2, which is just the open inscribed disk in D. We define the map g_D to be the identity map $U_D \to$ int D^2, which is both continuous and injective.

Let $\mathcal{E}$ denote the collection of all the edge-sets of D with respect to S, and let $\mathcal{V}$ denote the collection of all the vertex-sets of D with respect to S. It is not hard to see that the collection of sets

$$\{U_E\}_{E\in\mathcal{E}} \cup \{U_W\}_{W\in\mathcal{V}} \cup \{U_D\}$$

is an open cover of D (the sets are relatively open in D). We will also need two slightly shrunken, concentric versions of these sets, such that both collections of the shrunken sets still form open covers of D. Choose two very small numbers $\epsilon_2 > \epsilon_1 > 0$, and define the following sets:

$$U'_E = O_{A_n-\epsilon_1}(p, D) \cup O_{A_n-\epsilon_1}(p', D)$$

and

$$U''_E = O_{A_n-\epsilon_2}(p, D) \cup O_{A_n-\epsilon_2}(p', D)$$

for each $E \in \mathcal{E}$;

$$U'_W = \bigcup_{w\in W} O_{A_n-\epsilon_1}(w, D)$$

and

$$U''_W = \bigcup_{w\in W} O_{A_n-\epsilon_2}(w, D)$$

for each $W \in \mathcal{V}$;

$$U'_D = O_{1-\epsilon_1}(O_2, \mathbb{R}^2), \qquad U''_D = O_{1-\epsilon_2}(O_2, \mathbb{R}^2).$$

For a small enough choice of ϵ_2 and ϵ_1, the collections of sets

$$\{U'_E\}_{E\in\mathcal{E}} \cup \{U'_W\}_{W\in\mathcal{V}} \cup \{U'_D\}$$

and

$$\{U''_E\}_{E\in\mathcal{E}} \cup \{U''_W\}_{W\in\mathcal{V}} \cup \{U''_D\}$$

are open covers of D. Moreover, note that $U''_E \subset U'_E$ for each $E \in \mathcal{E}$, that $U''_W \subset U'_W$ for each $W \in \mathcal{V}$, and that $U''_D \subset U'_D$, where the inclusions are proper. See Figure A2.1.7. We have now completed the first stage of the proof of the theorem, as outlined at the start of the proof.

For the next stage of the proof, we start by constructing some auxiliary functions. First, let $\lambda, \mu: [0, \infty) \to [0, \infty)$ be functions with graphs as in Figure A2.1.8. Next, for each $E \in \mathcal{E}$ define a real-valued function $\phi_E: D \to \mathbb{R}$

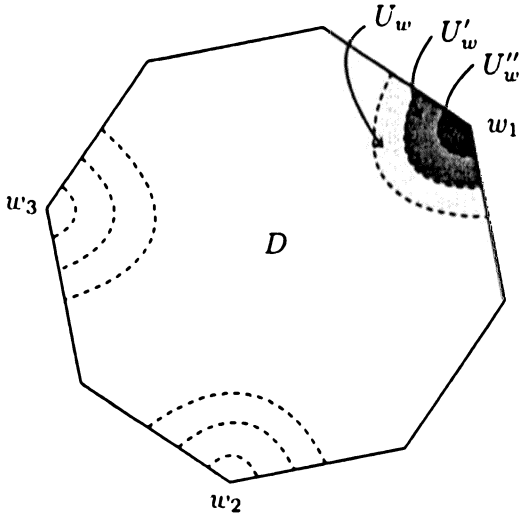


Figure A2.1.7

by

$$\phi_E(x) = \begin{cases} \lambda(\|x - p\|), & \text{if } x \in \overline{O_{\Lambda_n - \epsilon}(p, D)}; \\ \lambda(\|x - p'\|), & \text{if } x \in \overline{O_{\Lambda_n - \epsilon}(p', D)}; \\ 0, & \text{if } x \in D - U'_E, \end{cases}$$

where p and p' are as before. It is straightforward to see that ϕ_E is well-defined, and using Corollary 1.3.7 the map is seen to be continuous. It is important to observe that if two points $x, y \in D$ are identified by the gluing scheme S then $\phi_E(x) = \phi_E(y)$. Further, for any point $x \in U'_E$ we have $\phi_E(x) > 0$. For each $W \in \mathcal{V}$ we define a function $\phi_W : D \to \mathbb{R}$ completely analogously to the definition of ϕ_E; the functions ϕ_W have properties analogous to the ϕ_E. Finally, we define a function $\phi_D : D \to \mathbb{R}$ by $\phi_D(x) = \mu(\|x\|)$, and once again this function has properties analogous to the ϕ_E and ϕ_W.

To save writing, we define the set Δ to be the collection

$$\Delta = \mathcal{E} \cup \mathcal{V} \cup \{D\};$$

for each $\delta \in \Delta$ we thus have sets U_δ, U'_δ, and U''_δ and functions g_δ and ϕ_δ as defined above. Observe that since the sets U'_δ cover D, it follows that for each point $x \in D$ there is at least one $\delta \in \Delta$ such that $\phi_\delta(x) > 0$. Next, for each

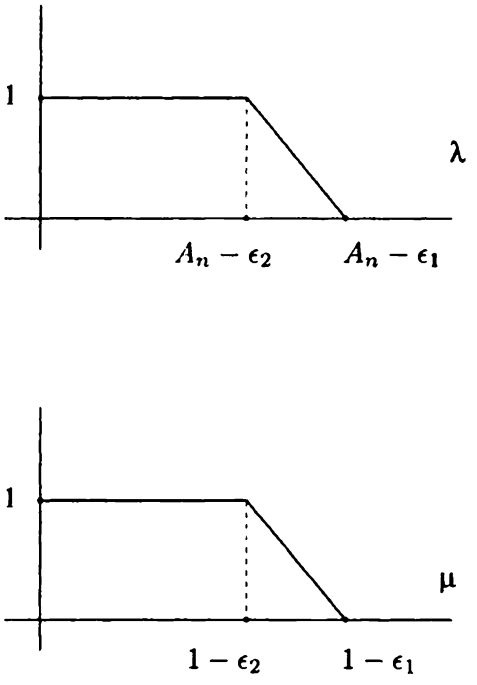


Figure A2.1.8

$\delta \in \Delta$ define a map $h_\delta \colon D \to \mathbb{R}^2$ by

$$h_\delta(x) = \begin{cases} \phi_\delta(x) \cdot g_\delta(x), & \text{if } x \in U_\delta; \\ O_2, & \text{if } x \in D - \overline{U'_\delta}. \end{cases}$$

To see that h_δ is a well-defined function, we note that the overlap of the domain of both cases of the definition is $U_\delta \cap (D - \overline{U'_\delta}) = U_\delta - \overline{U'_\delta}$, and that both cases of the definition have value O_2 on this region. That h_δ is continuous follows from Corollary 1.3.7, observing that h_δ is continuous on each of the two open subsets U_δ and $D - \overline{U'_\delta}$.

We are now ready for the home stretch. Suppose that the set Δ has d elements in it, labeled $\delta_1, \ldots, \delta_d$. (The number d depends upon the number of edges of D and the number of edge-sets and vertex-sets, which in turn depends upon the gluing scheme.) We define a map

$$H \colon D \to \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{k \text{ times}} \times \underbrace{\mathbb{R}^2 \times \cdots \times \mathbb{R}^2}_{k \text{ times}} = \mathbb{R}^{3d}$$

by

$$H(x) = \big(\phi_{\delta_1}(x), \ldots, \phi_{\delta_d}(x), h_{\delta_1}(x), \ldots, h_{\delta_d}(x)\big).$$

The map H is the one promised in the outline of the proof. Since each of the component maps of H is continuous, so is H by Lemma 1.3.8. As mentioned in the outline, the compactness of D and Proposition 1.6.14 (ii) together imply that H is a quotient map onto its image. To complete the theorem, we need to show that (1) $H(D)$ is a surface, and (2) $H(x) = H(y)$ for points $x, y \in D$ iff x and y are in the same set in $\mathcal{P}(S)$.

Let us start with item (2). Let $x, y \in D$ be any point, and suppose that $H(x) = H(y)$. If $x = y$ there is nothing to show, so assume $x \neq y$. From the definition of H it follows that $\phi_{\delta_i}(x) = \phi_{\delta_i}(y)$ and $h_{\delta_i}(x) = h_{\delta_i}(y)$ for all $i \in \{1, \ldots, d\}$. As remarked previously there is at least one $\delta_j \in \Delta$ for which $\phi_{\delta_j}(x) > 0$; it follows that $\phi_{\delta_j}(y) > 0$ as well. By the definition of the map ϕ_{δ_j} we see that $x, y \in U_{\delta_j}$. Since $h_{\delta_j}(x) = h_{\delta_j}(y)$ it follows that $\phi_{\delta_j}(x) \cdot g_{\delta_j}(x) = \phi_{\delta_j}(y) \cdot g_{\delta_j}(y)$. Hence $g_{\delta_j}(x) = g_{\delta_j}(y)$. There are now three cases, depending upon whether δ_j is an edge-set, a vertex-set or D. If δ_j is D, then g_{δ_j} is injective, and so $x = y$, a contradiction; hence $\delta_j \neq D$. If δ_j is an edge-set or a vertex-set, then, as mentioned above, the map g_{δ_j} is injective on $U_{\delta_j} \cap D$, and $g_{\delta_j}(x) = g_{\delta_j}(y)$ for any two points $x, y \in U_{\delta_j}$ iff x and y are in the same set in $\mathcal{P}(S)$. Putting all this information together demonstrates item (2). (One can now see why H was defined as it was.)

Finally, we need to show that $H(D)$ is a surface. If $H(x)$ is any point in $H(D)$, we need to show that there is a open subset of $H(D)$ containing $H(x)$ that is homeomorphic to int D^2. For convenience, we will find a subset of $H(D)$ that contains $H(x)$ and is homeomorphic to a disk in $\mathbb{R}^2$ by a homeomorphism that maps $H(x)$ to a point in the interior of the disk. Since the sets $\{U''_\delta\}_{\delta \in \Delta}$ cover D, there is some $\eta \in \Delta$ such that $x \in U''_\eta$. To complete the proof it will suffice to show that there is a homeomorphism between $H(\overline{U''_\eta})$ and a closed disk in $\mathbb{R}^2$. We proceed by defining a map $\pi: H(\overline{U''_\eta}) \to \mathbb{R}^2$ as follows: For any point $y \in H(\overline{U''_\eta})$, let $\pi(y) = g_\eta(z)$, where z is any point in $\overline{U''_\eta}$ such that $y = H(z)$. It needs to be verified that this definition makes sense, that is, that $g_\eta(z)$ is independent of the choice of z such that $H(z) = y$. Observe that if $y = H(z_1) = H(z_2)$, then as we saw above z_1 and z_2 are identified by the gluing scheme. It follows that $g_\eta(z_1) = g_\eta(z_2)$, and so π is well-defined.

We need to verify that π is injective. The proof is the backward version of what we just did. Say $\pi(y_1) = \pi(y_2)$ for $y_1, y_2 \in H(\overline{U''_\eta})$. Then $g_\eta(z_1) =$

$g_\eta(z_2)$, where $z_1, z_2 \in \overline{U''_\eta}$ are such that $y_1 = H(z_1)$ and $y_2 = H(z_2)$. Since $g_\eta(z_1) = g_\eta(z_2)$, it follows that z_1 and z_2 are identified under the gluing scheme, and therefore $H(z_1) = H(z_2)$; thus $q_1 = q_2$, so π is injective. Next, we need to verify that π is continuous. Observe that we can express the map π explicitly in terms of coordinates as follows. The domain of π is a subset of

$$\mathbb{R}^{3d} = \underbrace{\mathbb{R} \times \cdots \times \mathbb{R}}_{k \text{ times}} \times \underbrace{\mathbb{R}^2 \times \cdots \times \mathbb{R}^2}_{k \text{ times}}.$$

If $\eta = \delta_i$ then it can be seen, using the definition of the map H, that π is given by

$$\big(x_1, \ldots, x_d, (a_1, b_1), \ldots, (a_d, b_d)\big) \xmapsto{\pi} \left(\frac{a_i}{x_i}, \frac{b_i}{x_i}\right).$$

This formula makes sense, since the map ϕ_η is positive on $\overline{U''_\eta}$, hence the value of x_i is always positive in the domain of π, namely $H(\overline{U''_\eta})$. It follows from Exercise 1.3.9 and Lemma 1.3.8 that π is continuous.

Finally, observe that the set $\overline{U''_\eta}$ is closed and bounded, and hence compact by the Heine–Borel Theorem (Theorem 1.6.6). Since H is a continuous map, it follows from Theorem 1.6.10 that $H(\overline{U''_\eta})$ is compact as well. Since π is continuous and injective, it is a homeomorphism onto its image by Proposition 1.6.14 (iii). The image of π is equal to $g_\eta(\overline{U''_\eta})$, which is simply the disk $\overline{O}_{A_\eta - \epsilon_2}(O_2, \mathbb{R}^2)$ or $\overline{O}_{1-\epsilon_2}(O_2, \mathbb{R}^2)$, depending upon whether $\eta \in \mathcal{E} \cup \mathcal{V}$ or $\eta = D$. Hence the map π is a homeomorphism from $H(\overline{U''_\eta})$ to a disk in $\mathbb{R}^2$, and this is what we needed to show. □

Appendix A2.2 Proof of Proposition 2.6.1

We essentially follow the method of advanced texts such as [RO], [HE] and [MI1], though we take a rather circuitous route in order to avoid some technical difficulties. The bulk of our work will be to prove Proposition A2.2.8 below, which states that connected sum is well-defined for a certain class of surfaces. This class of surfaces will then be shown to include all surfaces, and it will follow that connected sum is well-defined for all compact connected surfaces.

Most of this section is taken up with a number of technical issues concerning 1-spheres and disks. We start with the notion of a homeomorphism of S^1 to itself being orientation preserving or reversing. Intuitively, the 1-sphere S^1 can be "oriented" in one of two ways, either clockwise or counterclockwise, as

shown by arrows in Figure A2.2.1. It seems plausible that any homeomorphism $h: S^1 \to S^1$ either preserves or reverses orientation. For example, a rotation preserves orientation, whereas reflection in the y-axis reverses orientation.

clockwise counterclockwise

Figure A2.2.1

Let $x, y \in S^1$ be any two distinct points. The points x and y divide S^1 into two arcs; we let $\overrightarrow{xy}$ denote the arc that runs from x to y in the counterclockwise direction, and we let $\overrightarrow{yx}$ denote the other arc. Now let $h: S^1 \to S^1$ be a homeomorphism. Observe that $S^1 - \{x, y\}$ consists of precisely two components, as does $S^1 - \{h(x), h(y)\}$. Since a homeomorphism takes components to components, it must be the case that h takes the arc $\overrightarrow{xy}$ onto one of the arcs $\overrightarrow{h(x)h(y)}$ or $\overrightarrow{h(y)h(x)}$. See Figure A2.2.2. The following lemma clarifies what might happen.

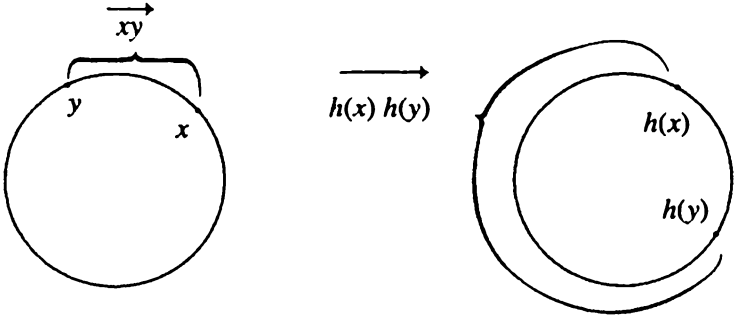


Figure A2.2.2

Lemma A2.2.1. *Let $h: S^1 \to S^1$ be a homeomorphism. Suppose that for some pair of distinct points $x, y \in S^1$ it is the case that $h(\overrightarrow{xy}) = \overrightarrow{xy}$. Then $h(\overrightarrow{x'y'}) = \overrightarrow{h(x')h(y')}$ for all pairs of distinct points $x', y' \in S^1$.*

Proof. There are a number of cases to check, depending upon which of x, x', y and y' are equal to one another; we will do the case where all these points are distinct, the other cases being similar. Either $y' \in \overrightarrow{xy}$ or $y' \in \overrightarrow{yx}$; suppose first that the former holds; since $h(\overrightarrow{xy}) = \overrightarrow{h(x)h(y)}$, it follows that $h(\overrightarrow{xy'}) \subset \overrightarrow{h(x)h(y')}$, and hence $h(\overrightarrow{xy'})$ must be the arc in S^1 that runs counterclockwise from $h(x)$ to $h(y')$. Thus $h(\overrightarrow{xy'}) = \overrightarrow{h(x)h(y')}$. A similar argument holds if $y' \in \overrightarrow{yx}$. If we now keep y' fixed and use the same argument while replacing x with x', we deduce that $h(\overrightarrow{x'y'}) = \overrightarrow{h(x')h(y')}$. □

We can now make the following definition.

Definition. Let $h: S^1 \to S^1$ be a homeomorphism. Then h is **orientation preserving** if for some pair of distinct points $x, y \in S^1$ it is the case that $h(\overrightarrow{xy}) = \overrightarrow{h(x)h(y)}$; otherwise h is **orientation reversing.**

To apply the concept of orientation preserving and reversing to any 1-sphere in $\mathbb{R}^n$, where there is no notion of "clockwise," we proceed by pulling everything back to S^1.

Definition. Let $C \subset \mathbb{R}^n$ be a 1-sphere, and let $h: C \to C$ be a homeomorphism. Then h is **orientation preserving** (respectively **orientation reversing**) if, for any homeomorphism $f: S^1 \to C$, the map $f^{-1} \circ h \circ f$ is an orientation preserving (resp. orientation reversing) homeomorphism of S^1 to itself.

It is verified in Exercise A2.2.5 that the choice of the homeomorphism f in the above definition does not affect the definition.

We now turn to homeomorphisms of disks. As a first step we show that any homeomorphism of S^1 to itself can be extended to a homeomorphism of D^2.

Lemma A2.2.2. *Let $h: S^1 \to S^1$ be a homeomorphism. Then there is a homeomorphism $H: D^2 \to D^2$ such that $H|S^1 = h$.*

Proof. There are many ways to define the map H, but the simplest is to set $H(O_2) = O_2$, and for each point $x \in S^1$ to map the line segment from O_2 to x linearly onto the line segment from O_2 to $h(x)$. See Figure A2.2.3. One can

give a formula for this map:

$$H(x) = \begin{cases} \|x\| h\left(\frac{x}{\|x\|}\right), & \text{if } x \in D^2 - \{O_2\}; \\ O_2, & \text{if } x = O_2. \end{cases}$$

It can be verified that this map is a homeomorphism. □

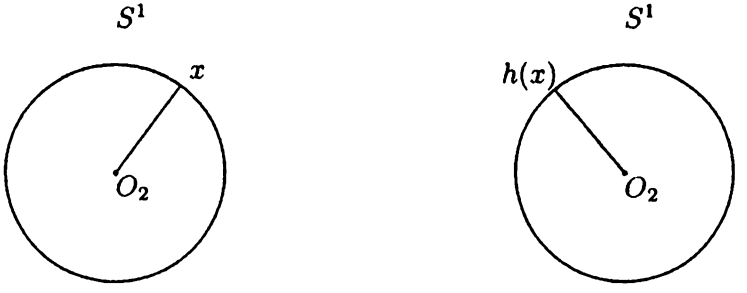


Figure A2.2.3

We now turn to a harder question. Let $B \subset \mathbb{R}^2$ be a disk, and suppose we are given a homeomorphism $h: \partial B \to \partial B$. As seen in Exercise A2.2.3, the map h can always be extended to a homeomorphism of $\mathbb{R}^2$ to itself; that is, there is always a homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H|\partial B = h$. Can we insure that H is the identity outside some larger disk containing B? In general the answer is no; one example (proved using algebraic topology) is given by reflecting S^1 in the y-axis. Certain maps h can be extended as desired, however, as shown in the following proposition.

Proposition A2.2.3. *Let $B \subset \mathbb{R}^2$ be a disk, and let $h: \partial B \to \partial B$ be a homeomorphism. If h is orientation preserving then there is a homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H(B) = B$, $H|\partial B = h$ and H is the identity map outside a disk containing B.*

Proof. First, suppose that we could prove the result in the case of the disk D^2; we show that the result would then hold for all disks $B \subset \mathbb{R}^2$, and all orientation preserving homeomorphisms $h: \partial B \to \partial B$. Let such a disk and such a homeomorphism be given. Since ∂B is a 1-sphere, it follows from Theorem 2.2.4 (the Schönflies Theorem) that there is a homeomorphism $G: \mathbb{R}^2 \to \mathbb{R}^2$ such that $G(S^1) = \partial B$ and G is the identity map outside a disk. The map $(G|S^1)^{-1} \circ h \circ G|S^1$ is an orientation preserving homeomorphism of S^1 to

itself, so by hypothesis there is a homeomorphism $F: \mathbb{R}^2 \to \mathbb{R}^2$ such that $F|S^1 = (G|S^1)^{-1} \circ h \circ G|S^1$ and F is the identity map outside a disk. It can then be verified that the map $H = G \circ F \circ G^{-1}$ is a homeomorphism of $\mathbb{R}^2$ to itself such that $H|\partial B = h$ and H is the identity map outside a disk (for the latter, use Exercise 2.2.7). It follows from Exercise 2.2.6 that $H(B) = B$.

Now comes the hard part, proving the theorem in the case of the disk D^2. Let $h: S^1 \to S^1$ be an orientation preserving homeomorphism. To define our homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ with the desired properties, we break up $\mathbb{R}^2$ into four closed regions, and we will define H on each of these regions. The four regions are the disk D^2; the washer-shaped region between the circles of radius 1 and 2 centered at the origin (including the 1-spheres), which we denote A_1; the washer-shaped region between the circles of radius 2 and 3 centered at the origin (including the 1-spheres), which we denote A_2; and the region outside the open disk of radius 3 centered at the origin. See Figure A2.2.4.

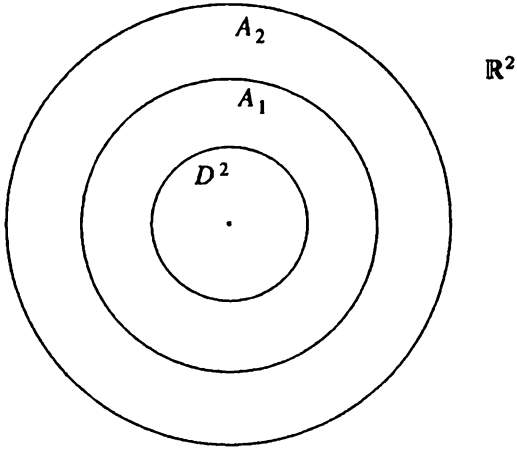


Figure A2.2.4

We define the map $H|D^2$ to be the homeomorphism of D^2 to itself given by Lemma A2.2.2. To define the map $H|A_1$ we start by using Exercise A2.2.6 to find a pair of antipodal points $p, q \in S^1$ such that $h(p)$ and $h(q)$ are antipodal. (We could do the proof without this exercise, but it's simpler, and more fun, this way.) Let Θ be the angle from the line segment $\overline{O_2 p}$ to the line segment $\overline{O_2 h(p)}$; it must also be the case that Θ is the angle between the analogous line segments with q replacing p. See Figure A2.2.5. We now define $H|A_1$ to be

the homeomorphism of A_1 to itself, which is h on S^1, and which takes each concentric circle to itself essentially by the map h followed by a rotation, so that at the circle of radius 2 the rotation is by angle $-\Theta$. If R_ϕ denote rotation centered at the origin of $\mathbb{R}^2$ by angle ϕ, we can give a formula for $H|A_1$ by

$$H|A_1(x) = \|x\| \, R_{(1-\|x\|)\Theta} \circ h\left(\frac{x}{\|x\|}\right).$$

It can be verified that $H|A_1$ fixes the points $2p$ and $2q$, and that $H|A_1$ restricted to the circle of radius 2 centered at the origin is an orientation preserving homeomorphism. Hence $H|A_1$ maps each of the arcs $\overrightarrow{2p\,2q}$ and $\overrightarrow{2q\,2p}$ homeomorphically to themselves.

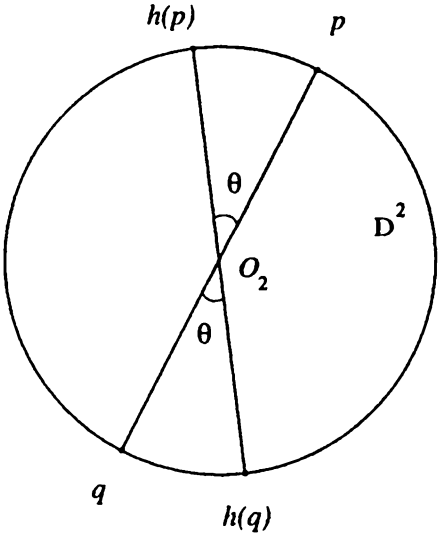


Figure A2.2.5

The map $H|A_2$ will be a homeomorphism of A_2 to itself. We start by setting $H|A_2$ equal to $H|A_1$ on the circle of radius 2 centered at the origin (the intersection of A_1 and A_2). We now divide up A_2 into two disks B_a and B_b as in Figure A2.2.6. We now apply Exercise A2.2.7 to each of these two disks. Although Exercise A2.2.7 is stated in terms of the rectangle $[-1, 1] \times [0, 1]$, it applies just as well to any other disk, and in particular to each of B_a and B_b, with the arcs $\overrightarrow{2p\,2q}$ and $\overrightarrow{2q\,2p}$ taking the role of $[-1, 1] \times \{0\}$. It follows from the exercise that there are homeomorphisms of each of B_a and B_b to themselves

that equal $H|A_1$ on the arcs $\overrightarrow{2p\,2q}$ and $\overrightarrow{2q\,2p}$, and are the identity maps on the rest of the boundaries of B_a and B_b. We define $H|A_2$ by piecing these homeomorphisms together.

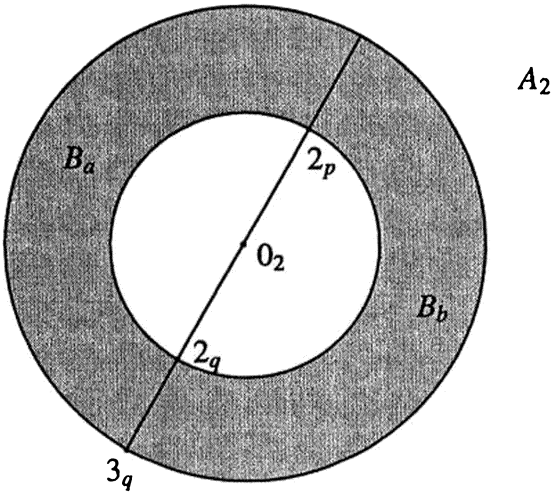


Figure A2.2.6

Observe that $H|A_2$ is the identity maps on the circle of radius 3 centered at the origin. Define H on the region outside the open disk of radius 3 centered at the origin to be the identity map. We have thus defined H on all of $\mathbb{R}^2$. Since H is a homeomorphism of each of the four regions used to itself, and agrees on the intersections of the regions, it is seen that H is a well-defined homeomorphism of $\mathbb{R}^2$ to itself. By construction $H|S^1 = h$, and H is the identity map outside the disk of radius 3 centered at the origin. □

We now have a number of other results about disks and homeomorphisms of surfaces.

Proposition A2.2.4. *Let $Q \subset \mathbb{R}^n$ be a surface, and let $B \subset Q$ be a disk. If $h\colon D^2 \to B$ is a homeomorphism, then there is a map $H\colon \overline{O}_2(0_2, \mathbb{R}^2) \to Q$ that is a homeomorphism onto its image and such that $H|D^2 = h$.*

Proof. We define the map H to equal h on D^2, so we need to define H on the washer-shaped region A_1 (as defined in the proof of Proposition A2.2.3); the problem is doing so injectively. By Proposition 1.6.14 (iii) and the compactness

of $\overline{O}_2(O_2, \mathbb{R}^2)$, any injective map defined on this set is automatically a homeomorphism onto its image, so injectivity is indeed the crucial property.

By the definition of a surface, every point $b \in B$ is contained in an open subset U_b of Q that is homeomorphic to int D^2. The sets $h^{-1}(U_b)$ for all $b \in B$ form an open cover of D^2. By Theorem 1.6.9 and the compactness of D^2 there is a number $\epsilon > 0$ such that for each point $x \in D^2$ the set $O_\epsilon(x, D^2)$ is contained in one of the sets $h^{-1}(U_b)$. We can thus divide S^1 into arcs $\alpha_1 \ldots \alpha_n$ for some sufficiently large integer n so that any three adjacent disks D_{i-1}, D_i and D_{i+1} as in Figure A.2.2.7 (i) are contained in a single set $h^{-1}(U_b)$. The annulus A_1 is then divided up into corresponding disks $E_1 \ldots E_n$ as in Figure A2.2.7 (ii), and the map H will be defined on the disks E_i one at a time, starting with the disk E_1.

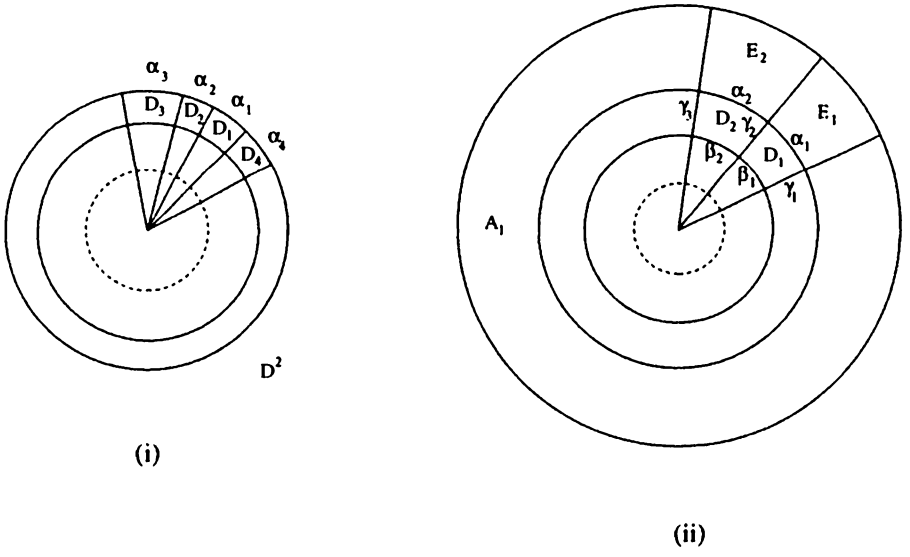


Figure A2.2.7

We start with the following preparation. The set $h(D_n \cup D_1 \cup D_2)$ is contained in the set U_b for some $b \in B$; fix this b. Since U_b is homeomorphic to int D^2 it is also homeomorphic to $\mathbb{R}^2$, and let $f: U_b \to \mathbb{R}^2$ be a homeomorphism. Since $D_n \cup D_1 \cup D_2$ is a disk, the set $f(h(D_n \cup D_1 \cup D_2))$ is a disk in $\mathbb{R}^2$. Let the rectangle $D_n' \cup D_1' \cup D_2'$ be as shown in Figure A2.2.8. Pick a homeomorphism $g: \partial f(h(D_n \cup D_1 \cup D_2)) \to \partial(D_n' \cup D_1' \cup D_2')$, where the arcs α_n,

$\alpha_1, \alpha_2, \beta_n, \beta_1$ and β_2 are taken to the line segments $\alpha_1' \ldots \beta_2'$, respectively, and the line segments γ_n and γ_3 are taken affine linearly to the line segments γ_n' and γ_3' (see the Appendix for a discussion of affine linear maps). By Exercise 2.2.7 there is a homeomorphism $G: \mathbb{R}^2 \to \mathbb{R}^2$ such that $G|\partial f(h(D_n \cup D_1 \cup D_2))$ and $G(f(h(D_n \cup D_1 \cup D_2))) = D_n' \cup D_1' \cup D_2'$ (ignore the points b_i in the exercise).

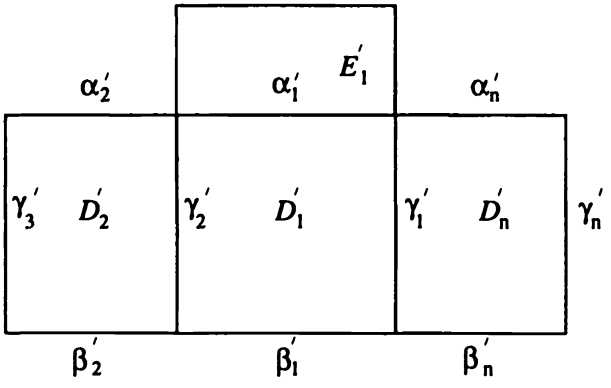


Figure A2.2.8

Choose some number $\delta > 0$, and let E_1' be the rectangle as shown in Figure A2.2.8. Define the map $p: E_1 \to E_1'$ by letting $p|\alpha_1 = G \circ f \circ h|\alpha_1$, and then having each radial line segment in E_1 be taken affine linearly to the corresponding vertical line segment in E_1'. We now extend H from D^2 to $D^2 \cup E_1$ by letting $H|E_1 = f^{-1} \circ G^{-1} \circ p$. It is straightforward to verify that H so defined is a continuous map on $D^2 \cup E_1$. With an arbitrary choice of δ the map H might not be injective, but H will be injective if δ is chosen to be small enough (but still positive), as we now show.

For any δ the map H is injective on each of D^2 and E_1, so the question is whether $H(x) = H(y)$ for some $x \in E_1$ and $y \in D^2 - \alpha_1$. The sets $H(\overline{D^2 - (D_n \cup D_1 \cup D_2)})$ and $H(\alpha_1)$ are compact and disjoint, so by Exercise 1.5.11 there is a number $m > 0$ such that $\|x - y\| \geq m$ for all $x \in H(\alpha_1)$ and $y \in H(\overline{D^2 - (D_n \cup D_1 \cup D_2)})$. By the compactness of α_1, it can be verified (using Exercise 1.5.6) that if $\delta > 0$ is chosen small enough then the sets $H(\overline{D^2 - (D_n \cup D_1 \cup D_2)})$ and $H(E_1)$ will be disjoint. Fix such a choice of δ. Then $H(x) \neq H(y)$ for any $x \in E_1$ and $y \in \overline{D^2 - (D_n \cup D_1 \cup D_2)}$. However, from the construction of H it can be verified that H is injective on all of $D_n \cup D_1 \cup D_2 \cup E_1$, hence $H(x) \neq H(y)$ for any $x \in E_1$ and

$y \in (D_n \cup D_1 \cup D_2) - \alpha_1$. It now follows that H is injective.

Extending H to $E_2 \ldots E_n$ is done similarly, with slight variations taking into account those E_i on which H has been defined; details are left to the reader. The number δ may have to be chosen smaller with each step, but since there are only finitely many steps, the number δ can always be chosen to be positive. This completes the proof. □

The following corollary can be derived straightforwardly from the above lemma and Exercise 2.3.4, and we omit the proof.

Corollary A2.2.5. *Let $Q \subset \mathbb{R}^n$ be a surface, and let $B \subset Q$ be a disk. Then B is contained in an open subset of Q homeomorphic to int D^2.*

To see the significance of this corollary, observe that by contrast not every subset of a surface is contained in an open subset of the surface that is homeomorphic to int D^2. For example, the 1-sphere $S^1 \times \{0\} \subset S^1 \times \mathbb{R}$ is not contained in a subset of $S^1 \times \mathbb{R}$ which is homeomorphic to int D^2. The proof of this fact is outlined in Exercise 3.8.4.

Proposition A2.2.6. *Let $Q \subset \mathbb{R}^n$ be a path connected surface and let B_1, $B_2 \subset Q$ be disks. Then there is a homeomorphism $H: Q \to Q$ such that $H(B_1) = B_2$.*

Proof. The setup takes more time than the actual argument. Choose points $p \in \text{int}\, B_1$ and $q \in \text{int}\, B_2$. Let $c: [0, 1] \to Q$ be a path from p to q, that is $c(0) = p$ and $c(1) = q$. By the definition of a surface, each point $x \in c([0, 1])$ is contained in an open subset of Q which is homeomorphic to int D^2. By the compactness and connectedness of $c([0, 1])$ it follows from Exercise 1.5.12 that there are finitely many of these open sets, say $U_1 \ldots U_r$, such that $p \in U_1$, that $q \in U_r$ and that $U_k \cap U_{i+1} \neq \emptyset$ for $k = 1, \ldots, r-1$. Using Corollary A2.2.5 there are open subsets of Q, called U_0 and U_{r+1} for convenience, such that both these sets are homeomorphic to int D^2 and contain B_1 and B_2 respectively. See Figure A2.2.9.

For each $k = 1, \ldots, r-1$ choose some point $x_k \in U_k \cap U_{k+1}$. For convenience let $x_0 = p$ and $x_r = q$. It is straightforward to see that by Exercise 2.3.3 there are disks $D_0, \ldots D_r \subset Q$ such that $x_k \in \text{int}\, D_k$ and $D_k \subset U_k \cap U_{k+1}$ for $k = 0, \ldots, r$. Let $D_{-1} = B_1$ and $D_{r+1} = B_2$ for convenience. See Figure A2.2.9.

We now make repeated use of Corollary 2.2.6; although Corollary 2.2.6 takes place in $\mathbb{R}^2$, it works just as well inside any of the sets U_k, each of which

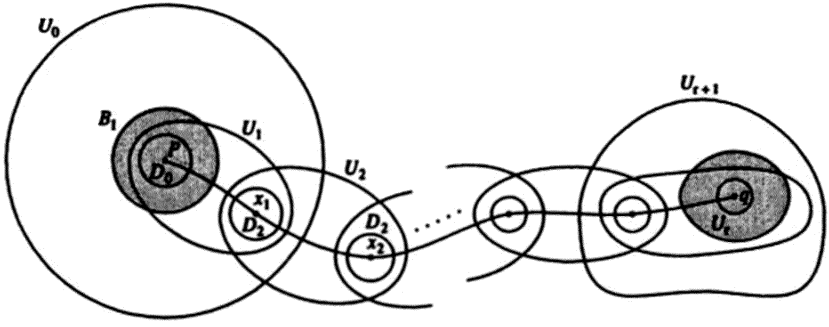


Figure A2.2.9

is homeomorphic to $\mathbf{R}^2$. Let $k \in \{0, \ldots, r+1\}$ be any number. The set U_k contains the disks D_{k-1} and D_k. Applying Corollary 2.2.6, there is a disk $C_k \subset U_k$ and a homeomorphism $G_k: U_k \to U_k$ such that $G_k(D_{k-1}) = D_k$ and G_k is the identity on $U_k - C_k$. By the continuity of G_k it is not hard to show that G_k must also be the identity on ∂C_k. Define a map $H_k: Q \to Q$ defined by

$$H_k(x) = \begin{cases} G_k(x), & \text{if } x \in C_k; \\ x, & \text{if } x \in Q - \operatorname{int} C_k. \end{cases}$$

Using Corollary 1.3.7 it can be shown that H_k is a homeomorphism; certainly $H_k(D_{k-1}) = D_k$. It can now be verified that the map $H = H_{r+1} \circ \cdots \circ H_0$ is the desired homeomorphism. □

Proposition A2.2.6 is not true if the hypothesis of path connectivity (or equivalently connectivity, by Proposition 2.5.1) is dropped. For example, if the surface Q consists of the union of a disjoint torus and sphere (as in Figure A2.2.10), then there can be no homeomorphism as in the conclusion of the theorem if B_1 is contained in the torus and B_2 is contained in the sphere; any homeomorphism must take components to components, and, as we shall see, the sphere is not homeomorphic to the torus.

Proposition A2.2.7. *Let $Q \subset \mathbf{R}^n$ be a surface, let $B \subset Q$ be a disk and let $h: \partial B \to \partial B$ be a homeomorphism. If h is orientation preserving then there is a homeomorphism $H: Q \to Q$ such that $H(B) = B$ and $H|\partial B = h$.*

Proof. Exercise A2.2.8.

We are now ready to discuss connected sums. We start by defining a convenient category of surfaces.

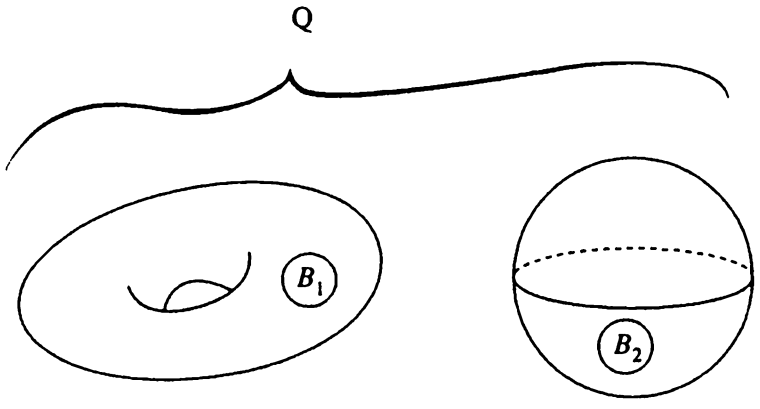


Figure A2.2.10

Definition. Let $Q \subset \mathbb{R}^n$ be a compact connected surface. The surface Q is **disk-reversible** if there is a disk $B \subset Q$ and a homeomorphism $H: Q \to Q$ such that $H(B) = B$ and $H|\partial B$ is an orientation reversing homeomorphism of ∂B.

We leave it to the reader in Exercise A2.2.9 to show that if a surface is disk-reversible, then the criterion in the definition is in fact satisfied with respect to any disk in the surface. Our main technical result on connected sums is the following. Recall the definition of connected sum given in Section 2.6.

Proposition A2.2.8. *Let $Q_1, Q_2 \subset \mathbb{R}^n$ be compact connected surfaces.*

(i) *Let $B_i \subset Q_i$ be a disk for $i = 1, 2$ and let $h: \partial B_1 \to \partial B_2$ be a homeomorphism. Then the attaching space $(Q_1 - int\, B_1) \cup_h (Q_2 - int\, B_2)$ exists and is a surface in some $\mathbb{R}^n$.*

(ii) *There are at most two distinct surfaces (depending only upon Q_1 and Q_2) to which all the surfaces $(Q_1 - int\, B_1) \cup_h (Q_2 - int\, B_2)$ are homeomorphic.*

(iii) *If at least one of Q_1 or Q_2 is disk-reversible, then all surfaces $(Q_1 - int\, B_1) \cup_h (Q_2 - int\, B_2)$ are homeomorphic.*

Proof. (i) & (ii). We start by constructing the two surfaces mentioned in part (ii), and then use them to demonstrate part (i). We start with some initial set-up. Using Theorem 2.4.3 (ii) there is, for each $i = 1, 2$, a polygonal disk D_i and a gluing scheme S_i for the edges of D_i such that Q_i is obtained from D_i

and S_i; let $q_i: D_i \to Q_i$ be an appropriate quotient map. For each $i = 1, 2$ choose a triangular disk $T_i \subset \text{int } D_i$, where T_i has vertices a_i, b_i and c_i. See Figure A2.2.11. It is straightforward to see that $T_i' = q_i(T_i)$ is a disk in Q_i. Let $f: \partial T_1 \to \partial T_2$ be the homeomorphism such that $f(a_1) = a_2$, $f(b_1) = b_2$, $f(c_1) = c_2$ and f is an affine linear map of each edge of the triangle T_1. Let $r: \partial T_2 \to \partial T_2$ be the homeomorphism such that $r(a_2) = a_2$, $r(b_2) = c_2$, $r(c_2) = b_2$ and r is an affine linear map of each edge of the triangle T_2. Observe that the map $f' = q_2|\partial T_2 \circ f \circ (q_1|\partial T_1)^{-1}$ is a homeomorphism $\partial T_1' \to \partial T_2'$, and that $r' = q_2|\partial T_2 \circ r \circ (q_2|\partial T_2)^{-1}$ is an orientation reversing homeomorphism of $\partial T_2'$ to itself.

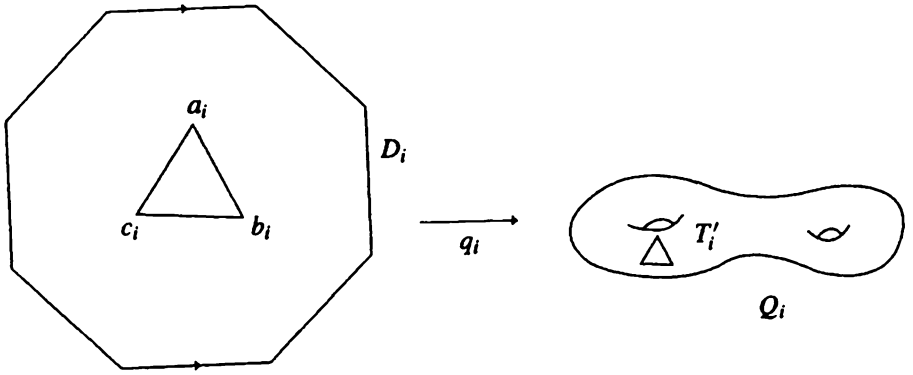


Figure A2.2.11

We now show that $(Q_1 - \text{int } T_1') \cup_{f'} (Q_2 - \text{int } T_2')$ and $(Q_1 - \text{int } T_1') \cup_{r' \circ f'} (Q_2 - \text{int } T_2')$ exist and are surfaces; we start with the first of these attaching spaces, the second being similar. We start by observing that the attaching space $(Q_1 - \text{int } T_1') \cup_{f'} (Q_2 - \text{int } T_2')$, if it exists, would be homeomorphic to the result of attaching the two disks with holes $D_1 - \text{int } T_1$ and $D_2 - \text{int } T_2$ via the homeomorphism $f: \partial T_1 \to \partial T_2$, and then gluing the edges of $(D_1 - \text{int } T_1) \cup_f (D_2 - \text{int } T_2)$ by the gluing schemes S_1 and S_2. See Figure A2.2.12. The problem is that $(D_1 - \text{int } T_1) \cup_f (D_2 - \text{int } T_2)$ is not a disk; we remedy the situation as follows.

For each $i = 1, 2$, make a cut in the $D_i - \text{int } T_i$ as shown in Figure A2.2.13 (i), yielding a disk E_i. Next, we attach the disks E_1 and E_2 by gluing the edge in E_1 with vertices a_1 and b_1 to the edge of E_2 with vertices a_2 and b_2 by an affine linear homeomorphism L of these edges that matches up correspondingly

named vertices. The resulting space, $E_1 \cup_L E_2$, is homeomorphic to a polygonal disk E with one edge for each edge of D_1 and D_2, one edge for each edge of ∂T_1 and ∂T_2 except those that were glued by L, and two edges resulting from each of the cuts used to obtain E_i from D_i. See Figure A2.2.13 (ii). Finally, construct a gluing scheme S for the edges of E by using S_1 and S_2 for the former edges of D_1 and D_2, use f on the former edges of ∂T_1 and ∂T_2, and match up those edges that resulted from cutting the D_i. By Theorem 2.4.3 (i) there is a surface $Q \subset \mathbb{R}^m$ obtained from E and S. We now leave it to the reader to verify that the attaching space $(Q_1 - \operatorname{int} T_1') \cup_{f'} (Q_2 - \operatorname{int} T_2')$ exists, and is homeomorphic to Q. A similar argument shows that $(Q_1 - \operatorname{int} T_1') \cup_{r' \circ f'} (Q_2 - \operatorname{int} T_2')$ exists and is a surface, call it Q_r.

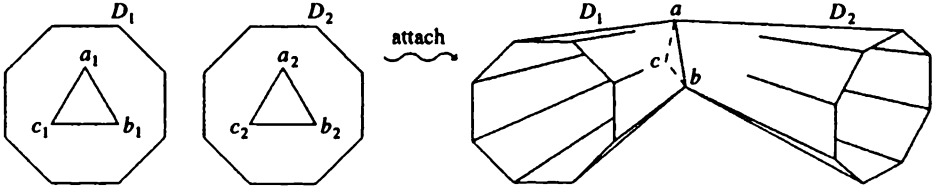


Figure A2.2.12

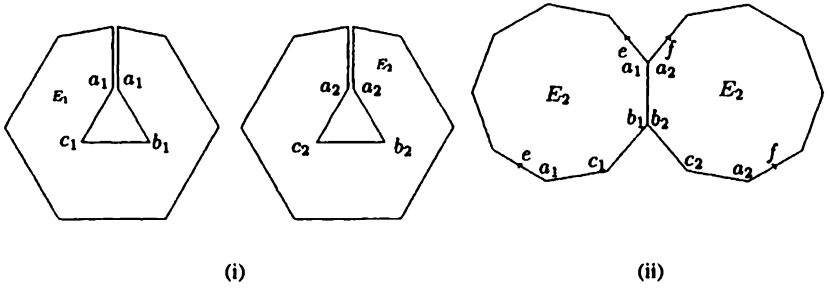


Figure A2.2.13

Now back to the original problem, namely, surfaces Q_1 and Q_2, as well as disks $B_i \subset Q_i$ for $i = 1, 2$ and a homeomorphism $h: \partial B_1 \to \partial B_2$. If we can show that $(Q_1 - \operatorname{int} B_1) \cup_h (Q_2 - \operatorname{int} B_2)$ exists and is homeomorphic to one of Q or Q_r then we would have proved parts (i) and (ii) of this proposition.

By Proposition A2.2.6 there are homeomorphisms $H_i: Q_i \to Q_i$ such that $H_i(B_i) = T_i'$ for $i = 1, 2$. It follows from Exercise 2.2.4 that $H(\partial B_i) = \partial T_i'$, and thus H_i maps $Q_i - \text{int } B_i$ homeomorphically onto $Q_i - \text{int } T_i'$.

Consider the map

$$g = f' \circ H_1|\partial B_1 \circ h^{-1} \circ (H_2|\partial B_2)^{-1}: \partial T_2' \to \partial T_2'.$$

This map is a homeomorphism. Since $\partial T_2'$ is a 1-sphere, it follows that g is either orientation preserving or orientation reversing; we will consider each case separately. First suppose that g is orientation preserving. By Proposition A2.2.7 there is a homeomorphism $G: Q_2 \to Q_2$ such that $G(T_2') = T_2'$ and $G|\partial T_2' = g$. The map $G \circ H_2$ is a homeomorphism of Q_2 to itself such that $G \circ H_2(B_2) = T_2'$. Further, it can be verified that

$$(G \circ H_2)|\partial B_2 \circ h = f' \circ H_1|\partial B_1.$$

Since $(Q_1 - \text{int } T_1') \cup_{f'} (Q_2 - \text{int } T_2')$ exists and is homeomorphic to the surface Q defined above, it now follows using Exercise 1.4.9 that $(Q_1 - \text{int } B_1) \cup_h (Q_2 - \text{int } B_2)$ exists and is homeomorphic to the surface Q.

If, on the other hand, the homeomorphism g is orientation reversing, then the map $r' \circ f'$ is an orientation preserving homeomorphism (because r' is orientation reversing, and making use of Exercise A2.2.4, which applies to all 1-spheres). An argument just like in the previous case shows that $(Q_1 - \text{int } B_1) \cup_h (Q_2 - \text{int } B_2)$ exists and is homeomorphic to the surface Q_r. This completes the proof of parts (i) and (ii) of the proposition.

(iii). From the proof of parts (i) and (ii) it will suffice to prove that the surfaces Q and Q_r are homeomorphic. The proof is similar to parts of the above proof, and details are left to the reader in Exercise A2.2.1. □

Proposition A2.2.8 shows that if $Q_1, Q_2 \subset \mathbb{R}^n$ are surfaces, and if at least one of them is disk-reversible, then $Q_1 \# Q_2$ is well-defined. We now need to show that all compact connected surfaces are disk-reversible. As a first step, we prove the following lemma.

Lemma A2.2.9. *All non-orientable surfaces, as well as S^2 and T^2, are disk-reversible.*

Proof. The disk-reversibility of S^2 and T^2 is shown in Exercise A2.2.10. Let $Q \subset \mathbb{R}^n$ be a non-orientable surface. Thus there is a Möbius strip M contained in Q. Suppose we could show that there is a disk $B \subset M$ and a homeomorphism

$h: M \to M$ such that $h|\partial M$ is the identity map, $h(B) = B$ and $H|\partial B$ is an orientation reversing homeomorphism of ∂B. We could then construct a homeomorphism $H: Q \to Q$ by setting $H|M = h$ and $H|Q - (M - \partial M)$ equal to the identity map, and this homeomorphism would have the desired properties with respect to the disk B.

To construct the homeomorphism h, we observe that if we can find such a homeomorphism on any copy of a Möbius strip, then we could find it on this particular copy, so without loss of generality let M be the standard Möbius strip shown in Figure A2.2.14 (i). For B pick any small round disk in M that does not touch ∂M. We can then deform M as shown in Figure A2.2.14 (ii)–(iv), where the disk B is eventually pushed all the way around M (with enough stretching this can be done without moving anything on ∂M). When B has returned to its original position it will have the orientation of its boundary reversed. Let h be the map that takes each point of M to where it ends up at the end of the deformation. □

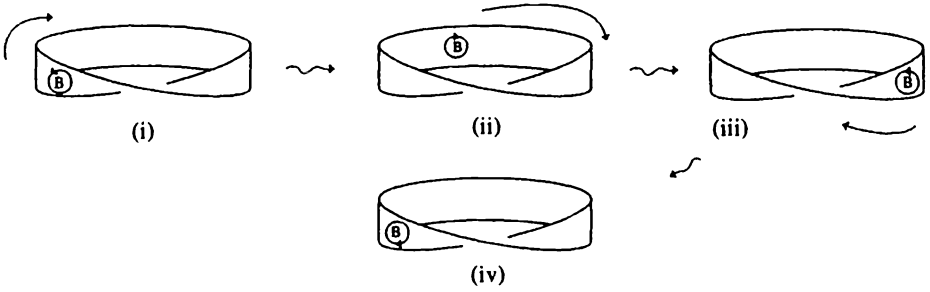


Figure A2.2.14

We can now complete our discussion of connected sum.

Proof of Proposition 2.6.1. The proposition would follow from Proposition A2.2.8 if we knew that every compact connected surface is disk-reversible. We see from Lemma A2.2.9 that all the surfaces referred to in the statement and proof of Theorem 2.6.7 are well defined. We leave it to the reader to show in Exercise A2.2.2 that all the surfaces referred to in the statement of Theorem 2.6.7 are disk-reversible (this follows from Lemma A2.2.9). From Theorem 2.6.7 it now follows that all compact connected surfaces are disk-reversible. (Note that the proof of Theorem 2.6.7 does not make use of connected sum

for any surfaces that are not disk reversible, so there is no circular reasoning here.) □

Exercises

A.2.2.1*. Prove Proposition A2.2.8 (iii).

A.2.2.2*. Show that all the surfaces referred to in the statement of Theorem 2.6.7 are disk-reversible.

A2.2.3*. Let $B \subset \mathbb{R}^2$ be a disk, and let $h: \partial B \to \partial B$ be a homeomorphism. Show that there is a homeomorphism $H: \mathbb{R}^2 \to \mathbb{R}^2$ such that $H|\partial B = h$.

A2.2.4*. Let $h_1, h_2: S^1 \to S^1$ be homeomorphisms. Show that h_1 is orientation preserving iff $(h_1)^{-1}$ is orientation preserving. Show that $h_2 \circ h_1$ is orientation preserving iff h_1 and h_2 are both orientation preserving or both orientation reversing.

A2.2.5*. Let $C \subset \mathbb{R}^n$ be a 1-sphere, let $h: C \to C$ be a homeomorphism and let $f_1, f_2: S^1 \to C$ be homeomorphisms. Show that the map $(f_1)^{-1} \circ h \circ f_1$ is an orientation preserving homeomorphism of S^1 to itself iff $(f_2)^{-1} \circ h \circ f_2$ is an orientation preserving homeomorphism of S^1 to itself.

A2.2.6*. Let $h: S^1 \to S^1$ be a homeomorphism. A pair of points in S^1 are **antipodal** if they are at opposite ends of a diameter of S^1. Show that there is a pair of antipodal points in $x, y \in S^1$ such that $h(x)$ and $h(y)$ are antipodal.

A2.2.7*. This exercise proves the one-dimensional version of what is known as the Alexander trick; this result, usually phrased in terms of isotopies, holds in all dimensions, where the appropriate closed ball replaces the closed interval $[-1, 1]$. Let $f: [-1, 1] \to [-1, 1]$ be a homeomorphism fixing the endpoints of the interval. Show that there is a homeomorphism $F: [-1, 1] \times [0, 1] \to [-1, 1] \times [0, 1]$ such that $F|[-1, 1] \times \{0\} = f$ and F is the identity map on the rest of the boundary of the rectangle $[-1, 1] \times [0, 1]$.

A2.2.8*. Prove Proposition A2.2.7.

A2.2.9*. Let $Q \subset \mathbb{R}^n$ be a disk-reversible surface. Show that for any disk $B \subset Q$ (not necessarily the one given in the definition of disk-reversibility)

there is a homeomorphism $H: Q \to Q$ such that $H(B) = B$ and $H|\partial B$ is an orientation reversing homeomorphism of ∂B.

A2.2.10*. Show that S^2 and T^2 are disk-reversible.

A2.2.11*. Let $B \subset S^2$ be a disk. Show that $S^2 - \text{int } B$ is a disk with ∂B its boundary.

Endnotes

Notes for Section 2.2

The first rigorous proof of Invariance of Domain was given by L. E. J. Brouwer in 1910 and was a major breakthrough in topology.

Notes for Section 2.5

(A) Though intuitively understandable, orientability is considered one of the technically tricky (or annoying) things in topology. We should mention that knowing that a surface is orientable is related to, but is not the same as, choosing an "orientation" for the surface.

(B) Another type of extrinsic property of surfaces in $\mathbb{R}^n$, which we will not be making use of but which the reader might wish to look up, is the issue of wildness vs. tameness. By contrast with Invariance of Domain (Theorem 2.2.1), which holds in all dimensions, the exact analog of the Schönflies Theorem (Theorem 2.2.4) in higher dimensions (concerning homeomorphic copies of the n-sphere in $\mathbb{R}^{n+1}$) does not hold. See [MO, p. 72] and [BI, chapter IV]. The history of this question is curious. Early counterexamples to the conjectured three-dimensional Schönflies Theorem were the "Antoine sphere" (first constructed in [AN2]) and the "Alexander horned sphere" (first constructed in [AL2]); it seems that the Antoine sphere was rediscovered by Alexander in [AL3], who made use of the "Antoine necklace" (discussed in the brief [AN1]), but who did not refer to the lengthy [AN2]. It is also interesting to note that Antoine, who discovered some very geometric examples, was blind.

The Schönflies Theorem does hold in higher dimensions if additional hypotheses are added; see [BN]. The higher dimensional Jordan Curve Theorem,

which is weaker than the Schönflies Theorem, is true in all dimensions; see [MU3] for a proof using algebraic topology.

Notes for Section 2.6

In contrast to the very clean statement of the classification theorem for compact connected surfaces (Theorem 2.6.7), it has been proved by [MK] that there can be no algorithm for the analogous type of classification for four-dimensional manifolds. Also, non-compact surfaces can be much more complicated than compact ones; see [RI].

Notes for Section A2.1

In the more elementary books that deal with surfaces, the gluing process is simply defined intuitively (with no mention of quotient spaces), and the fact that the result of gluing actually yields a surface in some Euclidean space is left unproved. The more advanced texts skirt the problem entirely by defining abstract surfaces, which are surfaces that do not necessarily sit inside of any Euclidean space; to formalize such a definition one needs the concept of a topological space, not developed in this text. It is not hard to show that the result of gluing the edges of a polygonal disk is an abstract surface, though of course it is then necessary to show that every abstract surface is homeomorphic to a surface sitting in some Euclidean space — for if not we would be faced with two different categories of surfaces: those in Euclidean spaces and those not. Our approach, in which we stay within Euclidean space but nonetheless provide a rigorous proof of Theorem 2.4.3 (i), is essentially a conflation of the two stages of the method used in more advanced books; we roughly follow [MU2, §4-5].

CHAPTER III

Simplicial Surfaces

3.1 Introduction

Topological surfaces can sit in Euclidean space very wildly, and as such can be difficult to work with. In order to develop the tools necessary for our proof of the classification of surfaces, as well as for other results, we turn our attention to simplicial surfaces, which are surfaces built out of triangles, and which are much easier to work with than arbitrary surfaces. Examples of simplicial surfaces include the surface of a tetrahedron (a pyramid with a triangular base) or an octahedron. See Figure 3.1.1. Simplicial surfaces have two advantages: They cannot sit wildly in Euclidean space, and they have things we can count (for example, the number of vertices) and measure (for example, angles).

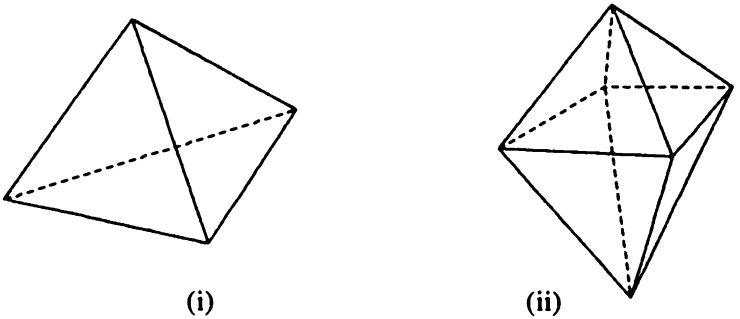


Figure 3.1.1

For reasons that will become clear later in this chapter, surfaces made out of triangles are easier to work with than general polyhedra, and we will therefore be focusing on simplicial surfaces; polyhedra will be mentioned only in passing and in the exercises. Using triangles is no real restriction, however, since the faces of any polyhedral surface can always be cut up into triangles.

3.2 Simplices

We start by examining triangles, edges and vertices, which are the building blocks for simplicial surfaces. These three types of objects are all convex, and we begin with a brief discussion of convexity; see [VA] or [BE] for more details. The reader should refer to the Appendix for some concepts from affine linear algebra that we will be using.

Intuitively, a convex set is one that has no "indentations." See Figure 3.2.1. The following definition, which conveys the same intuitive concept as having no indentations, is much more technically useful.

convex not convex

Figure 3.2.1

Definition. If $v, w \in \mathbb{R}^n$ are two points, the **line segment** from p to q is the set of points

$$\overline{vw} = \{x \in \mathbb{R}^n \mid x = tv + (1-t)w \text{ for } 0 \leq t \leq 1\}.$$

A subset $X \subset \mathbb{R}^n$ is **convex** if for every pair of points $v, w \in X$, the line segment $\overline{vw}$ is entirely contained in the set X. $\Diamond$

convex not convex

Figure 3.2.2

Example 3.2.1. The simplest examples of convex sets in $\mathbb{R}^n$ are $\mathbb{R}^n$ itself and any single point in $\mathbb{R}^n$. A more interesting example is that any line segment in $\mathbb{R}^n$ is convex. Let $\overline{vw} \subset \mathbb{R}^n$ be a line segment, and let $x, y \in \overline{vw}$ be two points. We need to show that $\overline{xy} \subset \overline{vw}$. By definition we can write $x = rv + (1-r)w$ and $y = sv + (1-s)w$ for some numbers $r, s \in [0, 1]$. Any element in $\overline{xy}$ has the form $tx + (1-t)y$ for some $t \in [0, 1]$. We now compute

$$\begin{aligned} tx + (1-t)y &= t[rv + (1-r)w] + (1-t)[sv + (1-s)w] \\ &= (s + rt - rs)v + [1 - (s + rt - rs)]w. \end{aligned}$$

To prove the desired result it suffices to show that $0 \leq s + rt - rs \leq 1$, and this verification is left to the reader. ◊

Since any two points in a convex set can be joined by a line segment, it follows immediately that all convex sets in $\mathbb{R}^n$ are path connected (and hence connected by Proposition 1.5.7). Convexity is a geometric property and is not preserved by arbitrary continuous maps. Affine linear maps do preserve convexity.

Lemma 3.2.2. *Let $F: \mathbb{R}^n \to \mathbb{R}^m$ be an affine linear map, and let $C \subset \mathbb{R}^n$ be a set. If C is convex then so is $F(C)$.*

Proof. Exercise 3.2.1. □

Any subset $X \subset \mathbb{R}^n$, not necessarily convex itself, is contained in some convex set (for example all of $\mathbb{R}^n$). Is X always contained in a smallest convex set? The following definition (which makes use of Exercise 3.2.2) and lemma show that the answer is yes.

Definition. Let $X \subset \mathbb{R}^n$ be any set. The **convex hull** of X, denoted conv X, is defined by

$$\operatorname{conv} X = \bigcap \{C \subset \mathbb{R}^n \mid X \subset C \text{ and } C \text{ is convex}\}. \quad \Diamond$$

Lemma 3.2.3. *For any set $X \subset \mathbb{R}^n$, the set conv X is convex, and if $S \subset \mathbb{R}^n$ is any convex set containing X, then conv $X \subset S$.*

Proof. Exercise 3.2.3. □

The convex hull of two distinct points is a line segment. The convex hull of three non-collinear points in $\mathbb{R}^2$ is a triangle. If three distinct points in $\mathbb{R}^2$ are

collinear then their convex hull is a line segment, though this is an inefficient way to obtain a line segment as a convex hull. The following definition generalizes the notion of line segment and triangle to arbitrary dimensions.

Definition. Let $a_0, \ldots, a_k \in \mathbb{R}^n$ be affinely independent points, where k is a non-negative integer. The **simplex** spanned by the points $a_0, \ldots, a_k$ is the convex hull of these points, and is denoted $\langle a_0, \ldots, a_k \rangle$; the points $a_0, \ldots, a_k$ are called the **vertices** of the simplex. $\Diamond$

It is straightforward to see that a simplex with one vertex is a single point, with two vertices is a line segment, with three vertices is a triangle and with four vertices is a solid tetrahedron (though not necessarily a regular tetrahedron). A useful characterization of the points in a simplex is given in the following lemma.

Lemma 3.2.4. *Let k be a non-negative integer and let $a_0, \ldots, a_k \in \mathbb{R}^n$ be affinely independent points. Then*

$$\langle a_0, \ldots, a_k \rangle = \{x \in \mathbb{R}^n \mid x = \sum_{i=0}^{k} t_i a_i \text{ for some numbers } t_0, \ldots, t_k \in \mathbb{R} \text{ such that } \sum_{i=0}^{k} t_i = 1 \text{ and } t_i \geq 0 \text{ for all } i\}; \tag{3.2.1}$$

for each point $x = \sum_{i=0}^{k} t_i a_i \in \langle a_0, \ldots, a_k \rangle$ the coefficients t_i are unique.

Proof. The uniqueness of the coefficients t_i is left to the reader in Exercise 3.2.5; we prove the first part of the lemma by induction on k. If $k = 0$ then both sides of Equation 3.2.1 are the single element set $\{a_0\}$, and the result holds. Now assume that the result is true for $k - 1$, and we will deduce the result for k. For convenience let D denote the right hand side of Equation 3.2.1. To prove the lemma we need to show three facts about D: (1) It contains the points $a_0, \ldots, a_k$, (2) it is contained in any convex set containing these points and (3) it is convex. Fact (1) is easy, since $a_i = 0a_1 + \cdots 0a_{i-1} + 1a_i + 0a_{i+1} + \cdots 0a_k$.

To show fact (2), let C be a convex set in $\mathbb{R}^n$ containing the points $a_0, \ldots, a_k$, and let $x = \sum_{i=0}^{k} t_i a_i$ be an element of D. If $t_k = 1$ then $x = a_k$, so $x \in C$ by hypothesis. Now assume $t_k \neq 1$. We can write

$$x = \sum_{i=0}^{k} t_i a_i = (1 - t_k)\{\sum_{i=0}^{k-1} \frac{t_i}{1 - t_k} a_i\} + t_k a_k.$$

Thus x is contained in the line segment from $y = \sum_{i=0}^{k-1} \frac{t_i}{1-t_k} a_i$ to a_k. If we could show that $y \in C$, then it would follow from the convexity of C that $x \in C$. Observe that $\sum_{i=0}^{k-1} \frac{t_i}{1-t_k} = 1$ and $\frac{t_i}{1-t_k} \geq 0$ for all $i \in \{0, \ldots, k-1\}$. By applying the inductive hypothesis to the affinely independent points $a_0, \ldots, a_{k-1}$ it is seen that $y \in \langle a_0, \ldots, a_{k-1} \rangle$. Since C is a convex set containing the points $a_0, \ldots, a_{k-1}$, we know that $\langle a_0, \ldots, a_{k-1} \rangle \subset C$. Hence $y \in C$.

To show fact (3) let $v = \sum_{i=0}^{k} t_i a_i$ and $w = \sum_{i=0}^{k} s_i a_i$ be points in D. Let $z \in \overline{vw}$ be a point, so that $z = rv + (1-r)w$ for some $r \in [0, 1]$. Hence

$$z = r \sum_{i=0}^{k} t_i a_i + (1-r) \sum_{i=0}^{k} s_i a_i = \sum_{i=0}^{k} (rt_i + s_i - rs_i) a_i.$$

A straightforward calculation shows that $\sum_{i=0}^{k} (rt_i + s_i - rs_i) = 1$. Further, note that $rt_i + s_i - rs_i = rt_i + s_i(1-r) \geq 0$. Hence $z \in D$, which completes the proof of fact (3). □

Observe that the coefficients t_i in Equation 3.2.1 must satisfy $0 \leq t_i \leq 1$.

We have defined a simplex by specifying its vertices; could the same simplex be defined by some other set of vertices? The following lemma answers this question.

Lemma 3.2.5. *Let $\{a_0, \ldots, a_k\}$ and $\{b_0, \ldots, b_p\}$ be two sets of affinely independent points in $\mathbb{R}^n$. If $\langle a_0, \ldots, a_k \rangle = \langle b_0, \ldots, b_p \rangle$, then $\{a_0, \ldots, a_k\} = \{b_0, \ldots, b_p\}$.*

Proof. If one of k or p is zero then the result is trivial, so assume $k, p > 0$. Let $r \in \{0, \ldots, k\}$ be a number. Since $a_r \in \langle b_0, \ldots, b_p \rangle$, it follows from Lemma 3.2.4 that $a_r = \sum_{i=0}^{p} t_i b_i$ for some $t_0, \ldots, t_p \in \mathbb{R}$ such that $\sum_{i=0}^{p} t_i = 1$ and $0 \leq t_i$ for all $i \in \{0, \ldots, p\}$. Since $b_i \in \langle a_0, \ldots, a_k \rangle$ for all i, we have $b_i = \sum_{j=0}^{k} s_{ij} a_j$ for some $s_{i1}, \ldots, s_{ik} \in \mathbb{R}$ such that $\sum_{j=0}^{k} s_{ij} = 1$ and $0 \leq s_{ij}$ for all i and j. Hence

$$a_r = \sum_{i=0}^{p} t_i \sum_{j=0}^{k} s_{ij} a_j = \sum_{j=0}^{k} (\sum_{i=0}^{p} t_i s_{ij}) a_j.$$

By the uniqueness of the coefficients in Lemma 3.2.4 it follows that $\sum_{i=0}^{p} t_i s_{ir} = 1$ and $\sum_{i=0}^{p} t_i s_{ij} = 0$ for $j \neq r$. Since $\sum_{i=0}^{p} t_i = 1$ and all numbers involved are in $[0, 1]$, it follows that whenever $t_i > 0$, then $s_{ir} = 1$ and $s_{ij} = 0$ for $j \neq r$. Thus, whenever $t_i > 0$ we deduce that $b_i = \sum_{j=0}^{k} s_{ij} a_j = a_r$. Since the b_i are

affinely independent they must be distinct, and so only one of the t_i is non-zero. Hence a_r is precisely one of the b_i. Thus $\{a_0, \ldots, a_k\} \subset \{b_0, \ldots, b_p\}$. A similar argument shows the reverse inclusion. □

Since we now know that a simplex has a unique set of vertices, the following definition can be made safely.

Definition. Let $\sigma = \langle a_0, \ldots, a_k \rangle$ be a simplex. The **dimension** of σ is k, and σ is called a **k-simplex**. ◊

A line segment in $\mathbb{R}^3$ is contained in a unique straight line in $\mathbb{R}^3$, and similarly a triangle in $\mathbb{R}^3$ is contained in a unique plane in $\mathbb{R}^3$. The following lemma generalizes this result to all dimensions.

Lemma 3.2.6. *Let $\eta = \langle a_0, \ldots, a_k \rangle$ be a k-simplex in $\mathbb{R}^n$. Then aspan$\{a_0, \ldots, a_k\}$ contains η, and it is the only k-plane in $\mathbb{R}^n$ containing η.*

Proof. Exercise 3.2.6. □

The following definition generalizes the observation that the boundary of a triangle (that is a 2-simplex) consists of edges (that is 1-simplices) and vertices (that is 0-simplices), and these edges and vertices are spanned by subsets of the set of vertices of the triangle.

Definition. Let $\sigma = \langle a_0, \ldots, a_k \rangle$ be a k-simplex in $\mathbb{R}^n$. A **face** of σ is a simplex spanned by a non-empty subset of $\{a_0, \ldots, a_i\}$; if this subset is proper the face is called a **proper face**. A face of σ that is a k-simplex is called a **k-face**. The **combinatorial boundary** of σ, denoted $\mathrm{Bd}\,\sigma$, is the union of all proper faces of σ. The **combinatorial interior** of σ, denoted $\mathrm{Int}\,\sigma$, is defined to be $\sigma - \mathrm{Bd}\,\sigma$. ◊

The term "face" does not necessarily mean "proper face." A 0-simplex has no proper faces. A 1-simplex $\langle a, b \rangle$ has two proper faces: the 0-simplices $\langle a \rangle$ and $\langle b \rangle$. A 2-simplex $\langle a, b, c \rangle$ has six proper faces: the 0-simplices $\langle a \rangle$, $\langle b \rangle$ and $\langle c \rangle$, and the 1-simplices $\langle a, b \rangle$, $\langle a, c \rangle$ and $\langle b, c \rangle$. See Figure 3.2.3. Observe also that a face of a face is a face.

Though the term "combinatorial boundary" and "combinatorial interior" used here are reminiscent of the terms "boundary" and "interior" used for disks and arcs in Section 2.2, the definition of the former terms is quite different in nature than that of the latter terms; hence different symbols are used. As shown in part (iii) of the following lemma, it turns out that these two types of

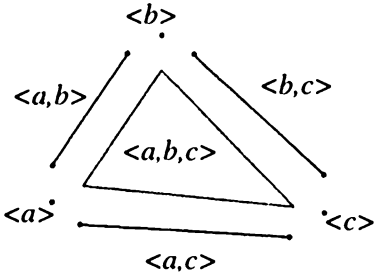


Figure 3.2.3

boundaries and interiors coincide for simplices. Since we have only defined "boundary" and "interior" for disks and arcs, we can only state this part of the lemma for dimensions 1 and 2, though in fact it holds in all dimensions. For each integer $k \geq 1$ we define

$$\begin{aligned} D^k &= \{x \in \mathbb{R}^k \mid \|x\| \leq 1\} \\ S^{k-1} &= \{x \in \mathbb{R}^k \mid \|x\| < 1\}, \end{aligned} \tag{3.2.2}$$

which are the closed unit disk and unit sphere in $\mathbb{R}^k$ respectively.

Lemma 3.2.7. *Let $\sigma = \langle a_0, \ldots, a_k \rangle$ be a k-simplex in $\mathbb{R}^n$.*

(i) *We have*

$$\begin{aligned} Bd\,\sigma = \{x \in \mathbb{R}^n \mid x = \sum_{i=0}^{k} t_i a_i \text{ for some numbers } t_0, \ldots, t_k \in \mathbb{R} \\ \text{such that } \sum_{i=0}^{k} t_i = 1,\ t_i \geq 0 \text{ for all } i \text{ and } t_j = 0 \text{ for some } j\}; \end{aligned} \tag{3.2.3}$$

$$\begin{aligned} Int\,\sigma = \{x \in \mathbb{R}^n \mid x = \sum_{i=0}^{k} t_i a_i \text{ for some numbers } t_0, \ldots, t_k \in \mathbb{R} \\ \text{such that } \sum_{i=0}^{k} t_i = 1 \text{ and } t_i > 0 \text{ for all } i\}. \end{aligned} \tag{3.2.4}$$

(ii) *There is a homeomorphism h: $D^k \to \sigma$ such that $h(S^{k-1}) = Bd\,\sigma$.*

(iii) *If σ is a 1-simplex it is an arc, if it is a 2-simplex it is a disk, and in both cases $Bd\,\sigma = \partial\sigma$ and $Int\,\sigma = int\,\sigma$.*

(iv) *Both σ and $Bd\,\sigma$ are compact and path connected.*

Proof. (i). Suppose that $x \in \mathrm{Bd}\,\sigma$. Thus x is in a proper face of σ. Since any proper face of σ is contained in a $(k-1)$-face of σ, there is some $j \in$

$\{0, \ldots, k\}$ such that $x \in \langle a_0, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k \rangle$. It now follows from Lemma 3.2.4 applied to this $(k-1)$-face that $x = \sum_{i \neq j} s_i a_i$ for some numbers $s_0, \ldots, s_{j-1}, s_{j+1}, \ldots, s_k \in \mathbb{R}$ such that $\sum_{i \neq j} s_i = 1$ and $s_i \geq 0$ for all $i \in \{0, \ldots, j-1, j+1, \ldots, k\}$. It follows from the uniqueness property of the coefficients in Lemma 3.2.4 applied to x and the simplex σ that if we express $x = \sum_{i=0}^{k} t_i a_i$ for some numbers $t_0, \ldots, t_k \in \mathbb{R}$ such that $\sum_{i=0}^{k} t_i = 1$ and $t_i \geq 0$ for all i, then $t_i = s_i$ for $i \neq j$ and $t_j = 0$. Conversely, suppose that $x = \sum_{i=0}^{k} t_i a_i$ for some numbers $t_0, \ldots, t_k \in \mathbb{R}$ such that $\sum_{i=0}^{k} t_i = 1$, $t_i \geq 0$ for all i and $t_j = 0$ for some j. Then once again using Lemma 3.2.4 it follows that $x \in \langle a_0, \ldots, a_{j-1}, a_{j+1}, \ldots, a_k \rangle$, and this $(k-1)$-simplex is contained in $\operatorname{Bd} \sigma$. Equations 3.2.3 and 3.2.4 now follow

(iv). Let b_0 be the origin in $\mathbb{R}^k$, and let

$$b_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \ldots, b_k = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 1 \end{pmatrix}$$

be the standard basis vectors in $\mathbb{R}^k$. It is straightforward to see that $\{b_0, \ldots, b_k\}$ is an affinely independent set, so that $\tau = \langle b_0, \ldots, b_k \rangle$ is a k-simplex in $\mathbb{R}^k$. It can be verified, using Lemma 3.2.4, that

$$\tau = \{ \begin{pmatrix} x_1 \\ \vdots \\ x_k \end{pmatrix} \in \mathbb{R}^k \mid x_i \geq 0 \text{ for all } i \text{ and } \sum_{i=1}^{k} x_i \leq 1 \}.$$

See Figure 3.2.4 for the case $k = 2$. From this simple description of τ it follows that both τ and $\operatorname{Bd} \tau$ are compact (since they are closed and bounded) and path connected. Using Exercises 3.2.8 and 1.6.3 it follows that any k-simplex and its combinatorial boundary are both compact and path connected.

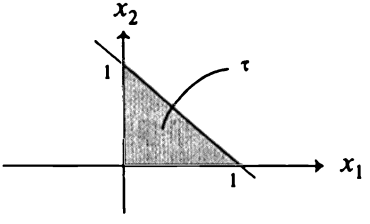


Figure 3.2.4

(ii). It follows from Exercises 3.2.7 and 3.2.8 that it would suffice to prove part (ii) for any k-simplex of our choice, rather than the given k-simplex σ. Choose some k-simplex η in $\mathbb{R}^k$ which contains the origin in its combinatorial interior; we will construct a homeomorphism $h: D^k \to \eta$ with the desired property. We start by defining a map $g: \operatorname{Bd}\eta \to S^{k-1}$ by setting $g(x) = \frac{x}{\|x\|}$, where this definition makes sense because $O_k \notin \operatorname{Bd}\eta$. It is seen that each ray in $\mathbb{R}^k$ starting at O_k intersects $\operatorname{Bd}\eta$ in precisely one point (this result is evidently true with respect to the simplex τ mentioned above and any point in $\operatorname{Int}\tau$, and using Exercise 3.2.7 and the fact that an injective affine linear take straight lines to straight lines it follows that this property holds for any k-simplex in $\mathbb{R}^k$). We thus see that the map g is bijective, and it is not hard to verify that g is continuous. By the compactness of $\operatorname{Bd}\eta$ and Proposition 1.6.14 (iii) it follows that g is a homeomorphism, so g^{-1} is a continuous map. The map h is now defined by setting $h|S^{k-1} = g^{-1}$, setting $h(O_k) = O_k$, and then extending h linearly on each radial line segment from O_k to a point in S^{k-1}. It can be verified that h is bijective. To show that h is continuous requires a bit more effort, making use of the ϵ-δ technique and some of the standard properties of continuous maps as found in any real analysis text; we omit the details. By the compactness of η and Proposition 1.6.14 (iii) it follows that h is a homeomorphism.

(iii). This follows from part (ii). □

Exercises

3.2.1*. Prove Lemma 3.2.2.

3.2.2*. Show that the intersection of convex sets is convex (there could be finitely or infinitely many sets).

3.2.3*. Prove Lemma 3.2.3.

3.2.4*. Show that any open ball in $\mathbb{R}^n$ is convex.

3.2.5*. Prove the uniqueness of the coefficients t_i in Lemma 3.2.4.

3.2.6*. Prove Lemma 3.2.6.

3.2.7*. Let η be a k-simplex in $\mathbb{R}^n$, and let $F: \mathbb{R}^n \to \mathbb{R}^m$ be an injective affine linear map. Show that $F(\eta)$ is a k-simplex, that F maps η homeomorphically onto $F(\eta)$ and that $F(\operatorname{Bd}\eta) = \operatorname{Bd}F(\eta)$ and $F(\operatorname{Int}\eta) = \operatorname{Int}F(\eta)$.

3.2.8*. Show that any two k-simplices are homeomorphic by an affine linear homeomorphism. More specifically, let $\langle x_0, \ldots, x_k \rangle$ and $\langle y_0, \ldots, y_k \rangle$ be k-simplices in $\mathbb{R}^n$ and $\mathbb{R}^m$ respectively, and let $F\colon \operatorname{aspan}\{x_0, \ldots, x_k\} \to \mathbb{R}^m$ be the unique affine linear map such that $F(x_i) = y_i$ for all $i \in \{1, \ldots, k\}$ (using Lemma A.7); show that F maps $\langle x_0, \ldots, x_k \rangle$ homeomorphically to $\langle y_0, \ldots, y_k \rangle$.

3.2.9*. Let η be a k-simplex in $\mathbb{R}^n$. Show that the intersection of any two faces of η is either a face of η or the empty set.

3.3 Simplicial Complexes

Simplices are used as building blocks for simplicial surfaces (and other objects), which are constructed by gluing simplices together along their faces. It is easiest if we glue the simplices together either edge-to-edge or corner-to-corner. See Figure 3.3.1. In order to keep track of what is used to build our objects, it is convenient to include the faces of each simplex used. The following definition gives the most general type of object we will construct out of simplices.

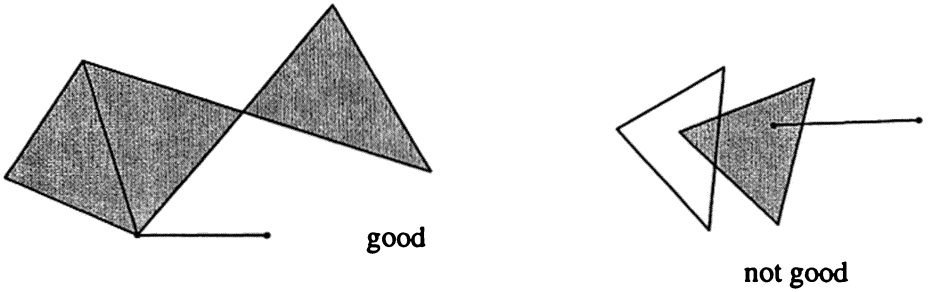


Figure 3.3.1

Definition. A **simplicial complex** K in $\mathbb{R}^n$ is a finite collection of simplices in $\mathbb{R}^n$ such that:

(a) if a simplex is in K, then all its faces are in K;

(b) if $\sigma, \tau \in K$ are simplices such that $\sigma \cap \tau \neq \emptyset$, then $\sigma \cap \tau$ is a face of each of σ and τ.

The **dimension** of a simplicial complex is defined to be the dimension of the highest-dimensional simplex that is in the simplicial complex. An i-dimensional simplicial complex will be referred to as an **i-complex**. ◊

A simplicial complex is not a single subset of $\mathbb{R}^n$, but rather is a collection of simplices; hence we do not write "$K \subset \mathbb{R}^n$" when we are referring to a simplicial complex K in $\mathbb{R}^n$. Although we have defined simplicial complexes to be finite, since that will suffice for our purposes, it is possible to define infinite simplicial complexes if certain local finiteness conditions are imposed.

Example 3.3.1. The simplicial complex corresponding to the tetrahedron (which we will always think of as a surface rather than a solid) is composed of four 2-simplices, six 1-simplices and four 0-simplices. Throughout this section the term "tetrahedron" will refer to this 2-complex. A non-example of a simplicial complex is a single triangle in $\mathbb{R}^n$. Although a single triangle is a 2-simplex, any simplicial complex must contain all the faces of each of its simplices, which in the case of a 2-simplex consist of three 1-simplices and three 0-simplices. ◊

We now make a number of useful technical definitions involving simplicial complexes.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$. For each non-negative integer i less than or equal to the dimension of K, we define $K^{((i))}$ to be the collection of all i-simplices in K. (This is slightly different from the usual notion of "i-skeleton" found in most texts.) If σ is a simplex in K, the **star** and **link** of σ in K, denoted $\text{star}(\sigma, K)$ and $\text{link}(\sigma, K)$ respectively, are defined to be

$$\text{star}(\sigma, K) = \{\eta \in K \mid \eta \text{ is a face of a simplex of } K \text{ which has } \sigma \text{ as a face}\}$$

and

$$\text{link}(\sigma, K) = \{\eta \in \text{star}(\sigma, K) \mid \eta \cap \sigma = \emptyset\}.$$

A subcollection L of K is a **subcomplex** of K if it is a simplicial complex itself. ◊

To verify that a subcollection L of a simplicial complex K is a subcomplex it suffices to verify that condition (a) of the definition of simplicial complexes holds, since condition (b) of the definition holds automatically.

Example 3.3.2. In the simplicial complex K in Figure 3.3.2 we see that $\text{star}(v, K)$ consists of the two 2-simplices $\langle a, b, v\rangle$ and $\langle e, f, v\rangle$ together with

all the faces of these 2-simplices; we see that link(v, K) consists of the two 1-simplices $\langle a, b\rangle$ and $\langle e, f\rangle$ together with all the faces of these 1-simplices. Both star(v, K) and link(v, K) are subcomplexes of K; this fact is proved in general in Exercise 3.3.2. ◊

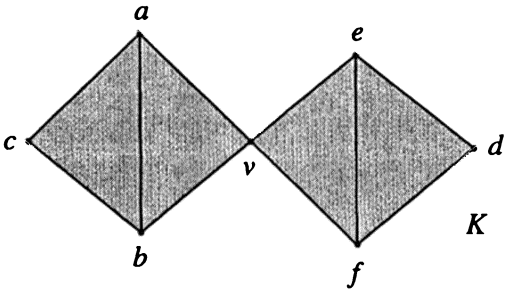


Figure 3.3.2

We need to define maps between simplicial complexes that preserve the relations between simplices and their faces. For technical ease it will suffice to use maps that take the 0-simplices of one simplicial complex to the 0-simplices of another.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$, and let L be a simplicial complex in $\mathbb{R}^m$. A map $f: K^{((0))} \to L^{((0))}$ is a **simplicial map** if whenever $\langle a_0, \ldots, a_i\rangle$ is a simplex in K, then $\langle f(a_0), \ldots, f(a_i)\rangle$ is a simplex in L. A simplicial map is a **simplicial isomorphism** if it is a bijective map on the set of vertices, and if its inverse is also a simplicial map. If there is a simplicial isomorphism from K to L then we say that K and L are **simplicially isomorphic**. ◊

Example 3.3.3. (1) Let K be a tetrahedron. Because any collection of two or three vertices in K are the vertices of a simplex in K, it follows that any map $f: K^{((0))} \to K^{((0))}$ defines a simplicial map $K \to K$.

(2) Let K and L be the 1-complexes shown in Figure 3.3.3. The map $g: K^{((0))} \to L^{((0))}$ defined by $g(a) = a'$, $g(b) = b'$, $g(c) = c'$ and $g(d) = d'$ is a simplicial map, and it is bijective. However, the map $g^{-1}: L^{((0))} \to K^{((0))}$ is not a simplicial map since $\langle a', c'\rangle$ is a simplex of L, and yet $\langle g^{-1}(a'), g^{-1}(c')\rangle = \langle a, c\rangle$ is not a simplex of K. Hence the separate requirements that the map and its inverse be simplicial in the definition of a simplicial isomorphism are both necessary. ◊

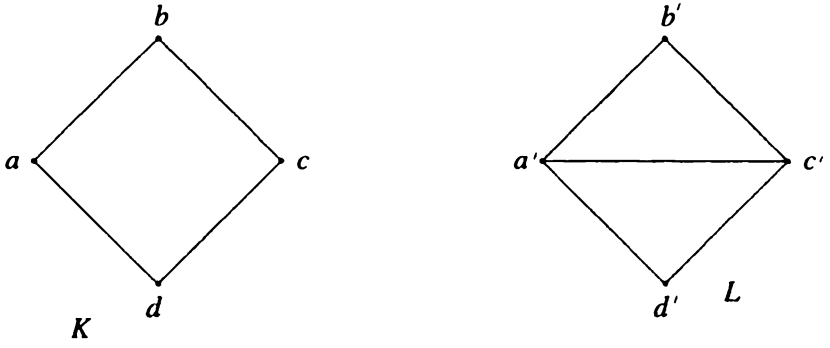


Figure 3.3.3

Consider a tetrahedron. We are presently thinking of it as a 2-complex consisting of four 0-simplices, six 1-simplices and four 2-simplices; from the point of view of Chapter 2, however, it can be thought of as a surface (homeomorphic to S^2) sitting in some Euclidean space. More generally, it will be useful to take simplicial complexes and "forget" their simplicial structures.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$. The **underlying space** of K, denoted $|K|$, is the subset of $\mathbb{R}^n$ that is the union of all the simplices in K. $\Diamond$

A simple consequence of the above definition is the following lemma.

Lemma 3.3.4. *Let K be a simplicial complex in $\mathbb{R}^n$. For each point $x \in |K|$ there exists a unique simplex η of K such that $x \in Int\, \eta$.*

Proof. Exercise 3.3.4. $\square$

We can rephrase our previous remarks about the tetrahedron by saying that the underlying space of a tetrahedron is homeomorphic to S^2. There are, however, other simplicial complexes with underlying spaces homeomorphic to S^2 (such as the octahedron), and in general we will need to be able to relate the various simplicial complexes that have the same underlying spaces up to homeomorphism. We start with the following definition.

Definition. Let K and K' be simplicial complexes in $\mathbb{R}^n$. The simplicial complex K' **subdivides** K if $|K'| = |K|$ and if every simplex of K' is a subset (not necessarily proper) of a simplex of K. $\Diamond$

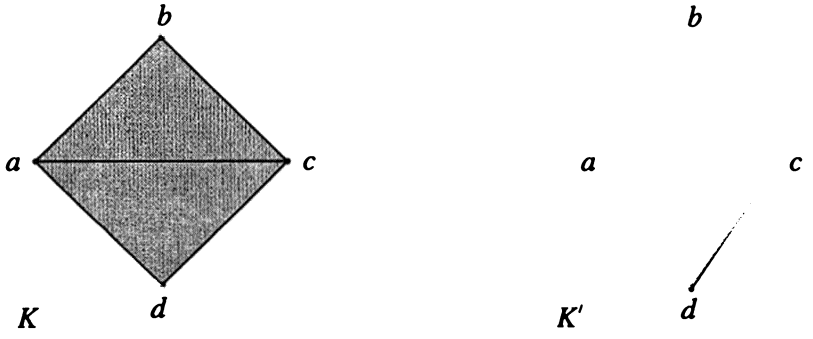


Figure 3.3.4

See Figure 3.3.4 for an example of a simplicial complex and a subdivision of it.

Given simplicial complexes K in $\mathbb{R}^n$ and L in $\mathbb{R}^m$, we have two ways of mapping one to the other, namely by a simplicial map (which is a map from $K^{((0))}$ to $L^{((0))}$), or by a map of underlying spaces (which is a map from $|K| \subset \mathbb{R}^n$ to $|L| \subset \mathbb{R}^m$). Are these two types of maps related? We show one type of relation. First, from the discussion of affine linear maps in the Appendix it follows that if $\langle a_0, \ldots, a_i \rangle$ is a simplex in $\mathbb{R}^n$, then any map $\{a_0, \ldots, a_i\} \to \mathbb{R}^m$ defined on the vertices of the simplex can be extended uniquely to an affine linear map $\langle a_0, \ldots, a_i \rangle \to \mathbb{R}^m$. (This ability to extend maps affine linearly is why simplices are more convenient than arbitrary polygons.)

Definition. Let K be a simplicial complex in $\mathbb{R}^n$, let L be a simplicial complex in $\mathbb{R}^m$, and let $f: K^{((0))} \to L^{((0))}$ be a simplicial map. The **induced map** of the underlying spaces of these complexes is the map $|f|: |K| \to |L|$ defined by extending f affine linearly over each simplex. $\Diamond$

The following lemma shows that induced maps work as expected.

Lemma 3.3.5. *Let K be a simplicial complex in $\mathbb{R}^n$, let L be a simplicial complex in $\mathbb{R}^m$.*

(i) *If $f: K^{((0))} \to L^{((0))}$ is a simplicial map, then the induced map $|f|: |K| \to |L|$ is continuous.*

(ii) *If K and L are simplicially isomorphic, then $|K| \approx |L|$.*

(iii) *If K and L have simplicially isomorphic subdivisions, then $|K| \approx |L|$.*

Proof. We will prove part (ii), leaving the rest to the reader in Exercise 3.3.5. Let $f: K^{((0))} \to L^{((0))}$ be a simplicial isomorphism; by definition the map $f^{-1}: L^{((0))} \to K^{((0))}$ is a simplicial map. Using part (i) of this lemma we deduce that $|f|$ and $|f^{-1}|$ are continuous maps. It will therefore suffice to show that $|f|$ is bijective and that $|f|^{-1} = |f^{-1}|$, the latter fact implying that $|f|^{-1}$ is continuous. To show that $|f|$ is injective, let $x, y \in |K|$ be any two points such that $x \neq y$. By Exercise 3.3.6 there exist unique simplices σ and τ of K such that $x \in \operatorname{Int}\sigma$ and $y \in \operatorname{Int}\tau$. The injectivity of f, the definition of a simplicial map and Lemma A.7 together imply that $|f|$ is an injective affine linear map on each simplex of K. By Exercise 3.2.7 we see that $|f|(x) \in \operatorname{Int}|f|(\sigma)$ and $|f|(y) \in \operatorname{Int}|f|(\tau)$. We now have two cases to consider, namely either $\sigma = \tau$ or not. In the former case we deduce that $|f|(x) \neq |f|(y)$, since $|f|$ is injective on $\sigma = \tau$. If $\sigma \neq \tau$ then the injectivity of f implies that $|f|(\sigma) \neq |f|(\tau)$. Using Exercise 3.3.6 again it follows that $\operatorname{Int}|f|(\sigma)$ and $\operatorname{Int}|f|(\tau)$ are disjoint, and thus $|f|(x) \neq |f|(y)$. Hence $|f|$ is injective. The surjectivity of $|f|$ follows straightforwardly from the surjectivity of f and the fact that f^{-1} is a simplicial map; details are left to the reader. Hence $|f|$ is bijective. Finally, to see that $|f|^{-1} = |f^{-1}|$, we observe that these two maps certainly agree on the vertices of L, and they agree on each simplex of L because the inverse of a bijective affine linear map is affine linear (Lemma A.6), and an affine linear map on a simplex is uniquely determined by what it does to the vertices of the simplex (Lemma A.7). □

From continuous maps we turn to other topological constructions. Since the underlying space of a simplicial complex is a subset of Euclidean space, we can apply concepts such as compactness and connectivity to simplicial complexes by examining whether these properties hold for the underlying spaces of simplicial complexes. Because we are only considering simplicial complexes with finitely many simplices it follows immediately from Lemmas 1.6.2 and 3.2.7 (iv) that all simplicial complexes have compact underlying spaces. Not all simplicial complexes have connected underlying spaces. However, we can characterize the connectedness of the underlying space of a simplicial complex in terms of the simplicial complex itself.

Lemma 3.3.6. *Let K be a simplicial complex in $\mathbb{R}^n$. The following are equivalent:*

(1) *$|K|$ is path connected.*
(2) *$|K|$ is connected.*

(3) *For any two simplices σ and τ of K there is a collection of simplices*

$$\tau = \eta_1, \eta_2, \ldots, \eta_p = \sigma$$

of K such that $\eta_i \cap \eta_{i+1} \neq \emptyset$ for all $i \in \{1, \ldots, p-1\}$.

Proof. Exercise 3.3.6. □

We need to find simplicial complexes whose underlying spaces are familiar objects such as T^2 and P^2. For T^2 this is easy, since it sits in $\mathbb{R}^3$; see Figure 3.3.5. Surfaces that do no sit in $\mathbb{R}^3$, such as P^2, are harder to work with. We will eventually solve this problem using the following construction, which is a simplicial analog of quotient maps and identification spaces (discussed in Section 1.4).

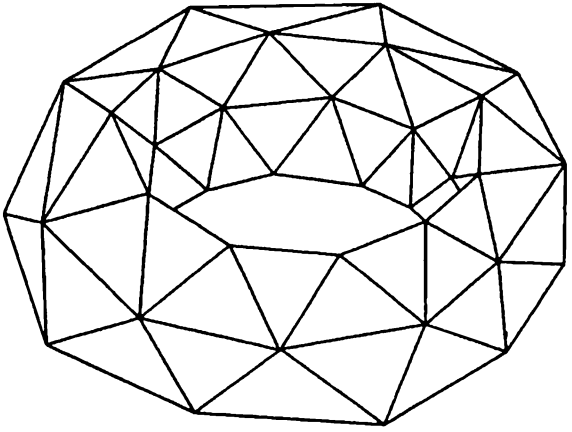
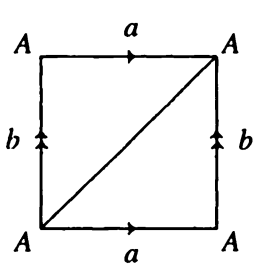
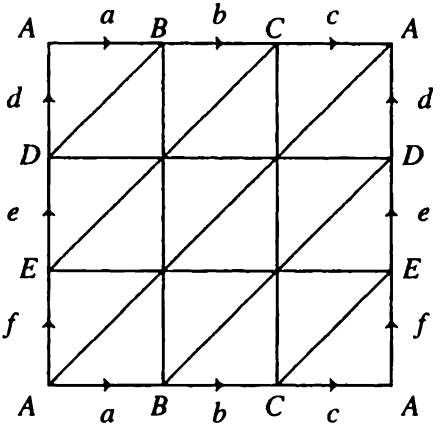


Figure 3.3.5

Consider the way in which T^2 is formed out of gluing the edges of a square, as described in Section 2.4. If we want to obtain a simplicial complex with an underlying space that is a torus, it would be tempting to break up the square shown in Figures 2.4.2 and 2.4.3 into two 2-simplices, as shown in Figure 3.3.6 (i). Unfortunately, when the edges of this square are identified as prescribed by the gluing scheme, all three vertices of both triangles are identified to a single point; since a 2-simplex must have three distinct vertices, we have not produced a simplicial complex by this process of breaking up the original square and then gluing. However, if we break up the original square into 2-simplices a bit more



(i)



(ii)

Figure 3.3.6

judiciously then we do not run into the same problems; see for example Figure 3.3.6 (ii). In general, we use the following definition.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$ and let L be a simplicial complex in $\mathbb{R}^m$. A simplicial map $f\colon K^{((0))} \to L^{((0))}$ is a **simplicial quotient map** if the following two conditions hold:

(1) For every simplex $\langle b_0, \ldots, b_p\rangle$ of L there is a simplex $\langle a_0, \ldots, a_p\rangle$ of K such that $f(a_i) = b_i$ for all $i \in \{0, \ldots, p\}$;

(2) if $a, b \in K^{((0))}$ are both vertices of a common simplex of K, then $f(a) \neq f(b)$. $\Diamond$

Example 3.3.7. (1) Let K and L be the 2-complexes shown in Figure 3.3.7. The map $f\colon K^{((0))} \to L^{((0))}$ defined by $f(a) = f(d) = z$, $f(b) = x$ and $f(c) = y$ is a simplicial quotient map. The map $g\colon K^{((0))} \to L^{((0))}$ defined by $g(a) = z$, $g(b) = x$ and $g(c) = g(d) = y$ is not a simplicial quotient map, since condition (1) of the definition is not satisfied.

(2) Any bijective simplicial map is a simplicial quotient map. $\Diamond$

Just as simplicial maps induce continuous maps of the underlying spaces, the following lemma shows that simplicial quotient maps induce quotient maps of the underlying spaces. For the second part of the lemma recall that if K is a simplicial complex in $\mathbb{R}^n$, then for any point $x \in |K|$ there is a unique simplex $\eta = \langle a_0, \ldots, a_k\rangle$ of K such that $x \in \operatorname{Int}\eta$ (using Lemma 3.3.4), and

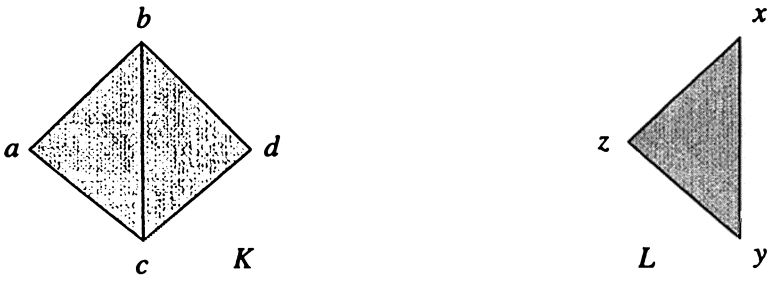


Figure 3.3.7

there are unique numbers $t_0, \ldots, t_k \in \mathbb{R}$ such that $\sum_{i=0}^{k} t_i = 1$, $t_i > 0$ for all $i \in \{0, \ldots, k\}$ and $x = \sum_{i=0}^{k} t_i a_i$ (using Lemmas 3.2.4 and 3.2.7).

Lemma 3.3.8. *Let K be a simplicial complex in $\mathbb{R}^n$, let L be a simplicial complex in $\mathbb{R}^m$, and let $f\colon K^{((0))} \to L^{((0))}$ be a simplicial quotient map.*

(i) *The induced map $|f|\colon |K| \to |L|$ is a quotient map.*

(ii) *Let $x \in |L|$ be a point, let $\eta = \langle a_0, \ldots, a_k \rangle$ be the unique simplex of L such that $x \in \operatorname{Int} \eta$, and let $t_0, \ldots, t_k \in \mathbb{R}$ be the unique numbers such that $\sum_{i=0}^{k} t_i = 1$, $t_i > 0$ for all $i \in \{0, \ldots, k\}$ and $x = \sum_{i=0}^{k} t_i a_i$. Then $|f|^{-1}(x)$ consists of all points $y \in |K|$ such that $y = \sum_{i=0}^{k} t_i b_i$, where $\langle b_0, \ldots, b_k \rangle$ is a simplex of K such that $f(b_i) = a_i$ for all $i \in \{0, \ldots, k\}$.*

Proof. (i). It is seen from condition (1) in the definition of simplicial quotient maps and Exercise 3.2.8 that the map $|f|$ is surjective. Using Lemma 3.3.5 (i) it follows that $|f|$ is continuous. As remarked above $|K|$ is compact, and hence $|f|$ is a quotient map by Proposition 1.6.14 (ii).

(ii). It is seen using condition (2) of the definition of simplicial quotient maps and Exercises 3.2.7 and 3.2.8 that $|f|$ is injective on each simplex of K, and that F maps k-simplices to k-simplices, taking interiors of simplices to interiors of simplices and boundaries to boundaries. If x and η are as in the statement of part (ii) of this lemma, then $|f|^{-1}(x)$ consists of points in the interiors of the k-simplices that are mapped onto η by $|f|$. The desired result now follows from the definition of affine linear maps. □

Recall the notion of partitions and identification spaces discussed in Section 1.4. We now define the simplicial analog of partitions, although we restrict the

type of partitions used in order to give rise to simplicial quotient maps. This definition corresponds to the simple idea that if we glue the 0-simplices by some scheme, then corresponding higher-dimensional simplices become glued as a result.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$. An **admissible partition** of $K^{((0))}$ is a collection $\mathcal{V} = \{A_i\}_{i \in I}$ of disjoint subsets of $K^{((0))}$ such that $\bigcup_{i \in I} A_i = K^{((0))}$, and such that no two vertices of the same simplex of K are in same set A_i. If $\mathcal{V}$ is an admissible partition of $K^{((0))}$, the **induced partition** of $|K|$ is the unique partition $\mathcal{P}(\mathcal{V})$ of $|K|$ such that two points $x, y \in |K|$ are in the same member of the partition iff $x \in \text{Int}\langle a_0, \ldots, a_k\rangle$ and $y \in \text{Int}\langle b_0, \ldots, b_k\rangle$ for k-simplices $\langle a_0, \ldots, a_k\rangle$ and $\langle b_0, \ldots, b_k\rangle$ of K, such that for all $i \in \{0, \ldots, k\}$ the 0-simplices a_i and b_i are both in the same member of the partition $\mathcal{V}$, and $F(x) = y$ where $F\colon \langle a_0, \ldots, a_k\rangle \to \langle b_0, \ldots, b_k\rangle$ is the unique affine linear map such that $F(a_i) = b_i$ for all i. $\Diamond$

We leave it to the reader to verify that for a given admissible partition of $K^{((0))}$ as in the above definition, there exists an induced partition of $|K|$, and that this induced partition is unique.

Example 3.3.9. Consider the 2-complex K shown in Figure 3.3.6 (ii). The underlying space $|K|$ of this 2-complex is a polygonal disk, and this disk has a gluing scheme as indicated in the figure. If two 0-simplices in the boundary of $|K|$ are glued together by the gluing scheme they are labeled with the same letter. The subset of $\mathbb{R}^n$ obtained from $|K|$ and the given gluing scheme is a torus. Now let $\mathcal{V}$ be the partition of $K^{((0))}$ consisting of all pairs of similarly labeled 0-simplices in the boundary of $|K|$, and single-element sets containing each 0-simplex in the interior of $|K|$. It is seen that $\mathcal{V}$ is an admissible partition of $K^{((0))}$. Further, the induced partition $\mathcal{P}(\mathcal{V})$ of $|K|$ is seen to be the same as the partition of $|K|$ induced by the gluing scheme, as described in Section 2.4. Hence the identification space of $|K|$ and $\mathcal{P}(\mathcal{V})$ is a torus. $\Diamond$

The above example suggests the following result.

Lemma 3.3.10. *Let K be a simplicial complex in $\mathbb{R}^n$ and let $\mathcal{V}$ be an admissible partition of $K^{((0))}$.*

(i) *There is a simplicial complex K' in $\mathbb{R}^m$ for some m and a simplicial quotient map $f\colon K^{((0))} \to K'^{((0))}$ such that $\{f^{-1}(v) \mid v \in K'^{((0))}\} = \mathcal{V}$.*

(ii) *If $\mathcal{P}(\mathcal{V})$ is the induced partition of $|K|$, then $\mathcal{P}(\mathcal{V}) = \{|f|^{-1}(x) \mid x \in |K'|\}$.*

(iii) *The identification space of* $|K|$ *and* $\mathcal{P}(\mathcal{V})$ *is homeomorphic to* $|K'|$.

Proof. (i). Suppose that $\mathcal{V} = \{A_1, \ldots, A_m\}$, where the A_i are disjoint subsets of $K^{((0))}$. Let $e_1, \ldots, e_m \in \mathbb{R}^m$ denote the standard basis vectors. It is straightforward to see that $e_1, \ldots, e_m$ are affinely independent. The simplicial complex K' we are constructing will consist of some of the faces of the simplex $\Delta = \langle e_1, \ldots, e_m \rangle$. More specifically, let K' consist of all faces of Δ of the form $\langle e_{i_1}, \ldots, e_{i_p} \rangle$, where $i_1, \ldots, i_p \in \{1, \ldots, m\}$ are numbers for which there exists a simplex $\langle b_{i_1}, \ldots, b_{i_p} \rangle$ in K with $b_{i_j} \in A_{i_j}$; all simplices of K are of this form, by the definition of admissible partitions of $K^{((0))}$. Also, it is straightforward to verify that if a face of Δ is in K, then any face of this face is also in K', which implies that condition (a) in the definition of simplicial complexes holds for K'. Condition (b) in the definition of simplicial complexes holds for K' because of Exercise 3.2.9; and hence K' is a simplicial complex.

A map $f\colon K^{((0))} \to K'^{((0))}$ is defined by setting $f(v) = e_i$ if $v \in A_i$. That f has the desired properties now follows straightforwardly from the construction of K' and f, and the fact that $\mathcal{V}$ is an admissible partition of $K^{((0))}$.

(ii). This follows from the construction of K', Lemma 3.3.8 (ii), properties of affine linear maps and the definition of the induced partition $\mathcal{P}(\mathcal{V})$. Details are left to the reader.

(iii). The map $|f|$ is a quotient map by Lemma 3.3.8 (i), and the result now follows from part (ii) of the present lemma and the definition of identification spaces in Section 1.4. □

Example 3.3.11. We continue Example 3.3.9. Let K' be a simplicial complex in some $\mathbb{R}^m$ as guaranteed by the above lemma. It follows from part (iii) of the lemma that $|K'|$ is homeomorphic to the identification space of $|K|$ and $\mathcal{P}(\mathcal{V})$, and this latter space is a torus as mentioned in Example 3.3.9. Thus we have constructed a simplicial complex with underlying space a torus. We will make use of this construction in Section 3.5. ◊

Exercises

3.3.1. What are the star and link of the vertices v and w in the simplicial complex shown in Figure 3.3.8?

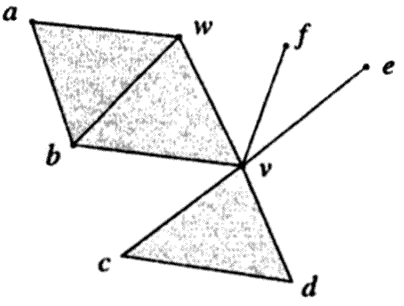


Figure 3.3.8

3.3.2*. Let K be a simplicial complex in $\mathbb{R}^n$, and let σ be a simplex in K. Show that $\text{star}(\sigma, K)$ and $\text{link}(\sigma, K)$ are subcomplexes of K.

3.3.3. Let K and L be the simplicial complexes shown in Figure 3.3.9. Let $f: K^{((0))} \to L^{((0))}$ be given by $f(a) = v$, $f(b) = w$, $f(c) = x$ and $f(d) = y$. Is this a simplicial map? Is this a simplicial quotient map?

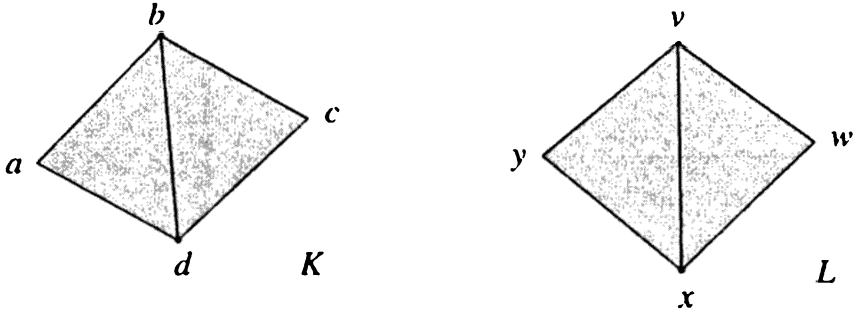


Figure 3.3.9.

3.3.4*. Prove Lemma 3.3.4.

3.3.5*. Prove Lemma 3.3.5 parts (i) and (iii).

3.3.6. Prove Lemma 3.3.6.

3.3.7*. Let K be a simplicial complex in $\mathbb{R}^n$, and let σ be a simplex of K. Show that $|\text{star}(\sigma, K)|$ is path connected, and that $|\text{link}(\sigma, K)|$ is path connected iff $|\text{star}(\sigma, K)| - \sigma$ is path connected.

3.3.8*. Let K be a simplicial complex in $\mathbb{R}^n$ and let L be a subdivision of K. Show that for each simplex η of L there is a unique simplex σ of K such that $\eta \cap \operatorname{Int}\sigma \neq \emptyset$; further, show that $\eta \subset \sigma$ for this unique simplex σ. Is the statement true if the condition $\eta \cap \operatorname{Int}\sigma \neq \emptyset$ is replaced with the condition $\eta \cap \sigma \neq \emptyset$?

The following exercises discuss a slightly more general type of object than simplicial complexes. To simplify matters we restrict our attention to the two-dimensional case.

3.3.9. Recall the definition of a polygonal disk in Section 2.4; such a disk need not be convex. If D is a polygonal disk, show that there is a simplicial complex K such that $|K| = D$ and the 0-simplices of K are precisely the vertices of D (we might say that the complex K is a simplicial subdivision of D with no new vertices).

3.3.10. A two-dimensional **cell complex** C in $\mathbb{R}^n$ is a finite collection of polygonal disks in $\mathbb{R}^n$ that satisfy the same two conditions as in the definition of simplicial complexes. Define the notions of star, link, subcomplex, subdivision and underlying space of two-dimensional cell complexes analogously to the case for simplicial complexes. Show that any convex cell complex in $\mathbb{R}^n$ has a subdivision that is simplicial complex (called a **simplicial subdivision**); moreover, a simplicial subdivision can always be found that has no new vertices.

3.4 Simplicial Surfaces

We would like to look at all simplicial complexes with underlying spaces that are topological surfaces (for example, the octahedron). Which properties of the simplicial complex structure of the octahedron, shown in Figure 3.4.1. (i), distinguish it from the 2-complex in Figure 3.4.1 (ii), the underlying space of which is clearly not a surface? By Lemma 3.2.7 (ii) the points in the interiors of all 2-simplices in any 2-complex have neighborhoods that are homeomorphic to open disks in $\mathbb{R}^2$, so they present no problem. What makes the neighborhood of each point in the interior of a 1-simplex work out correctly in the octahedron is that each 1-simplex in the octahedron is a face of precisely two 2-simplices, which is not the case in Figure 3.4.1 (ii). Further, the link of each 0-simplex of the octahedron is a 1-sphere, which once again does not hold in Figure 3.4.1

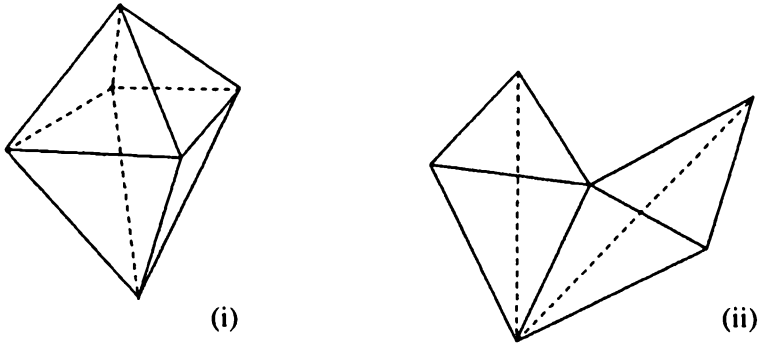


Figure 3.4.1.

(ii). These two observations just do the trick, as seen in the following theorem. The proof of this theorem makes use of Invariance of Domain (Theorem 2.2.1).

Theorem 3.4.1. *Let K be a simplicial complex in $\mathbb{R}^n$. Then $|K|$ is a topological surface iff K is a 2-complex such that each 1-simplex of K is the face of precisely two 2-simplices, and the underlying space of the link of each 0-simplex of K is a 1-sphere.*

Proof. If K is a 2-complex such that each 1-simplex of K is contained in precisely two 2-simplices, and the link of each 0-simplex is a 1-sphere, then it is straightforward to verify that each point in $|K|$ has a neighborhood homeomorphic to int D^2, and we leave the details to the reader. The difficult part of the proof is the other direction; we follow the treatment in [MO, Chapter 23]. Suppose from now on that $|K|$ is a topological surface.

We start by showing that K is a 2-complex. Let m be the dimension of K, where m is a non-negative integer. There is some m-simplex τ in K; let $x \in \text{Int } \tau$ be a point. The maximal dimensionality of τ implies that any small enough open ball in $\mathbb{R}^n$ centered at x intersects no simplex of K other than τ. Hence it follows from Lemma 3.2.7 (ii) that the point x has an open neighborhood in $|K|$ that is homeomorphic to $\mathbb{R}^m$. Since $|K|$ is a topological surface x has an open neighborhood $A \subset |K|$ homeomorphic to int D^2, and hence to $\mathbb{R}^2$. By Exercise 2.3.3 we can choose A as small as desired, and in particular we can choose it to be contained in the open neighborhood of x that is homeomorphic to $\mathbb{R}^m$. Hence, up to homeomorphism, there is a set homeomorphic to $\mathbb{R}^2$ sitting as an open subset of $\mathbb{R}^m$. By Exercise 2.2.10 it must be the case that $m = 2$.

We next show that each 1-simplex of K is the face of precisely two 2-simplices of K. Let η be a 1-simplex of K; for our entire discussion of η fix a point $x \in \operatorname{Int}\eta$, and let $U \subset |K|$ be an open subset containing x, and which is homeomorphic to int D^2, and hence to $\mathbb{R}^2$; we can choose U to be as small as desired. Note that if $\operatorname{Int}\eta$ intersects a 2-simplex σ of K then η is a face of σ.

Suppose that η is not the face of any 2-simplex. Then if we choose the set U small enough, it will be contained in η, and hence it will be contained in a straight line in $\mathbb{R}^n$. Let Π be a plane in $\mathbb{R}^n$ containing η (it does not matter which plane). By Exercise 1.2.18 (i) the set U is not open in Π (it makes no difference that we are in an arbitrary plane in $\mathbb{R}^n$ rather than in $\mathbb{R}^2$). On the other hand, U is homeomorphic to $\mathbb{R}^2$ and is contained in the plane Π (itself homeomorphic to $\mathbb{R}^2$). By Theorem 2.2.1 we see that U must be open in Π, a contradiction. Hence η must be the face of at least one 2-simplex of K.

Now suppose η is the face of precisely one 2-simplex σ. Using Lemma 3.2.6 let Π be the unique plane (that is a 2-plane) in $\mathbb{R}^n$ containing σ (and hence η). If we choose the set U small enough it will be contained in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma$, and hence it will be contained in a closed half-plane in Π, where the boundary of the half-plane is the unique line containing η. Observe that U intersects the boundary of the half-plane, namely in the point x. By Exercise 1.2.18 (ii) the set U is not open in Π. On the other hand, as in the previous case, since U is homeomorphic to $\mathbb{R}^2$ it follows from Theorem 2.2.1 that U must be open in Π, a contradiction. Hence η must be the face of at least two 2-simplices of K.

Next suppose η is contained in more than two 2-simplices; let $\sigma_1, \ldots, \sigma_p$ be the 2-simplices of K that have η as a face, where $p \geq 3$. Choose U small enough so that it is entirely contained in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma_1 \cup \cdots \cup \operatorname{Int}\sigma_p$. It is straightforward to see that the set $\sigma_1 \cup \eta \cup \sigma_2$ is a disk, and that $\operatorname{Int}\sigma_1 \cup \operatorname{Int}\eta \cup \operatorname{Int}\sigma_2$ is homeomorphic to int D^2, and hence to $\mathbb{R}^2$. We can thus find a small open subset V of x in $\operatorname{Int}\sigma_1 \cup \operatorname{Int}\eta \cup \operatorname{Int}\sigma_2$ such that V is homeomorphic to $\mathbb{R}^2$ and $V \subset U$. Now, on the one hand, Exercise 1.2.18 (iii) implies that V is not open in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma_1 \cup \cdots \cup \operatorname{Int}\sigma_p$. On the other hand $U \approx \mathbb{R}^2 \approx V$, and U is open in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma_1 \cup \cdots \cup \operatorname{Int}\sigma_p$; it follows from Exercise 2.2.1 that V is open in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma_1 \cup \cdots \cup \operatorname{Int}\sigma_p$, a contradiction. We conclude that η is contained in precisely two 2-simplices.

Now let w be a 0-simplex of K; we need to show that $|\operatorname{link}(w, K)|$ is a 1-sphere. The subcomplex $\operatorname{link}(w, K)$ consists of a finite number of 0-simplices and 1-simplices of K. Each 0-simplex in $\operatorname{link}(w, K)$ is the face of a unique 1-simplex in $\operatorname{star}(w, K)$, and each 1-simplex in $\operatorname{link}(w, K)$ is the face of a unique

2-simplex in star(w, K); moreover, two 1-simplices in link(w, K) intersect (in a common endpoint) iff the 2-simplices in star(w, K) of which they are faces intersect in a common 1-simplex. Since we just saw that every 1-simplex of K is contained in precisely two 2-simplices, it follows that every 0-simplex in link(w, K) is contained in precisely two 1-simplices in link(w, K). It is therefore not hard to see that $|\text{link}(w, K)|$ must be the union of disjoint polygonal 1-spheres.

To prove the desired result it therefore suffices to show that $|\text{link}(w, K)|$ is path connected. Suppose otherwise; it would follow from Exercise 3.3.7 that $|\text{star}(w, K)| - \{w\}$ would not be path connected. By Exercise 2.3.3 we can find an open neighborhood $V \subset |K|$ of w that is entirely contained in $|\text{star}(w, K)|$ and is homeomorphic to int D^2. By Exercise 1.5.11 the set $V - \{w\}$ is not path connected. It would follow that we have found a set homeomorphic to int D^2 that becomes non-path connected when a single point is removed, a contradiction to Exercise 1.5.7. Thus $|\text{link}(w, K)|$ is path connected. □

Using the above theorem we can make the following definition, which makes no reference to the underlying space of K.

Definition. A 2-complex K is called a **simplicial surface** if K is a 2-complex such that each 1-simplex of K is the face of precisely two 2-simplices, and the underlying space of the link of each 0-simplex of K is a 1-sphere. The underlying space of a simplicial surface is called the **underlying surface** of the simplicial surface. ◊

Example 3.4.2. A tetrahedron is a simplicial surface, since it is a 2-complex, each 1-simplex is the face of precisely two 2-simplices, and the underlying space of the link of each 0-simplex is a triangle. On the other hand, a single 2-simplex together with its faces is not a simplicial surface, since the underlying space of the link of each 0-simplex is a line segment. ◊

The definition of simplicial surfaces can be made more elegant as follows. Recall the definitions of S^k in Equation 3.2.2; note that S^0 consists of two points in $\mathbb{R}$, namely ± 1. For convenience we could define S^{-1} to be the empty set. A simplicial surface is then seen to be a simplicial complex in which the underlying space of the link of every i-simplex is homeomorphic to S^{1-i} for $i = 0, 1, 2$.

More importantly, though the criteria in Theorem 3.4.1 are quite natural, there is in fact a redundancy in the criteria. The following lemma will make it

slightly easier to verify whether a given 2-complex is a simplicial surface.

Lemma 3.4.3. *Let K be a 2-complex such that the underlying space of the link of each 0-simplex is a 1-sphere. Then K is a simplicial surface.*

Proof. Exercise 3.4.1. □

All the properties of topological surfaces discussed in Section 2.5 can be applied to simplicial surfaces by applying them to the underlying surfaces. Thus, we say that a simplicial surface is compact, connected, etc., if the underlying topological surface is compact, connected, etc. Since, in fact, all simplicial complexes have compact underlying spaces, all simplicial surfaces are compact.

Returning to the relation between topological surfaces and simplicial complexes, we have now shown that the underlying space of a simplicial surface (and of no other type of simplicial complex) is a topological surface (Theorem 3.4.1), and if any two simplicial surfaces have simplicially isomorphic subdivisions then their underlying surfaces are homeomorphic (Lemma 3.3.5). What about the other way around: Is every topological surface a simplicial surface as well? To answer this question we must state it with a bit more care. Even a simple surface such as S^2 is not a finite simplicial complex as it is sitting in $\mathbb{R}^3$. However, certainly S^2 is homeomorphic to the underlying space of a number of simplicial surfaces, such as the octahedron. Rather amazingly, this same result holds for any topological surface. Moreover, given a topological surface, there is essentially only one way to find the requisite simplicial surfaces. To state this fact precisely we need the following definition.

Definition. Let $Q \subset \mathbb{R}^n$ be a topological surface. A simplicial complex K **triangulates** Q if there is a homeomorphism $t\colon |K| \to Q$; we say that Q is **triangulated** by K; the simplicial complex K together with the homeomorphism t are called a **triangulation** of Q. ◊

Example 3.4.4. The topological surface S^2 is triangulated by the tetrahedron via the radial projection map from the underlying space of a small tetrahedron with center of gravity at the origin to S^2. See Figure 3.4.2. ◊

If a simplicial complex K triangulates a topological surface, then, by Theorem 3.4.1, we know that K must be a simplicial surface. We now state the following result without proof; see [MO, §8] for details.

Theorem 3.4.5.

(i) *Any compact topological surface in $\mathbb{R}^n$ can be triangulated.*

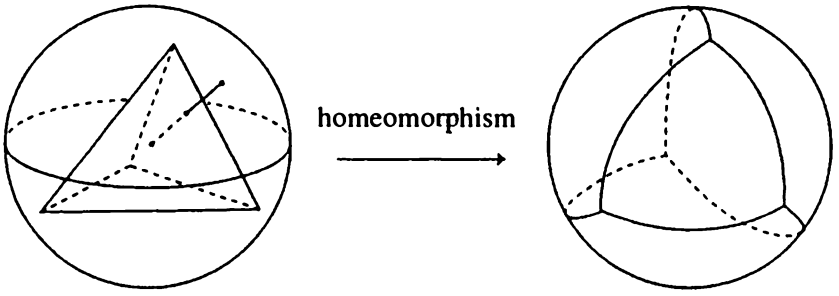


Figure 3.4.2

(ii) *If a topological surface is triangulated by two simplicial complexes K_1 and K_2, then K_1 and K_2 have simplicially isomorphic subdivisions.*

The analog of neither part of the theorem is true in higher dimensions. The counterexamples are quite sophisticated, and they were discovered only relatively recently. See [K-S]. Also, we need to assume compactness since our simplicial complexes are finite by definition, though it is possible to deal with non-compact surfaces as well.

As an application of Theorem 3.4.5 we prove Theorem 2.4.3 (ii), which has been left hanging up till now.

Proof of Theorem 2.4.3 (ii). By Theorem 3.4.5 we know that Q is homeomorphic to $|K|$ for some simplicial surface K. It therefore suffices to prove that for every simplicial surface K there is a polygonal disk D and a gluing scheme S for the edges of D such that $|K|$ is obtained from gluing the edges of D by the scheme S. We prove the theorem backward. First, suppose that $|K|$ can be obtained by gluing pairs of edges of a finite number of disjoint polygonal disks; we will prove by induction on the number of polygonal disks that $|K|$ can in fact be obtained from a single polygonal disk.

Let n be the number of polygonal disks used. If $n = 1$ then there is nothing to prove. Next suppose that $n > 1$, and that the claim holds whenever fewer than n disks are used. Now, observe that if the edges of each polygonal disk are only glued to edges of the same disk, then the net result of gluing all the edges of all the disks will be an object that has as many pieces as there are polygonal disks, namely n. Since we are assuming $n > 1$, we would have contradicted the fact that K is connected. Hence it could not have been the case that the edges of each polygonal disk are only glued to edges of the same disk. We can

therefore find two polygonal disks D_1 and D_2 such that D_1 has an edge with the same label as an edge of D_2. So, glue D_1 and D_2 along this commonly label edge. We note two facts. First, when two polygonal disks are glued together along a single edge of each the result is a polygonal disk. Second, if we take all n disks and glue the appropriate edges all at once, or if we glue the edges one pair at a time, we obtain the same object. It is now easy to see that $|K|$ can be obtained from $n-1$ polygonal disks, using the result of gluing D_1 and D_2 together, and the other $n-2$ disks that were originally used for $|K|$. By the inductive hypothesis, $|K|$ can be obtained from a single polygonal disk.

Finally, we need to show that $|K|$ can be obtained by gluing the edges of some finite number of polygonal disks. Well, $|K|$ is obtained by gluing the 2-simplices of K along their faces, and 2-simplices are polygonal disks. □

Exercises

3.4.1*. Prove Lemma 3.4.3.

3.4.2*. Find a simplicial complex that triangulates the torus with the smallest number of 2-simplices you can get away with. What about with the smallest number of 0-simplices?

The following exercises discuss two-dimensional cell complexes. (See Exercise 3.3.10.)

3.4.3. Define a **polyhedral surface** to be a two-dimensional cell complex in which the underlying space of the link of each 0-simplex is a 1-sphere. Prove that the underlying space of a polyhedral surface is a topological surface.

3.4.4. Let P be a polyhedral surface, and let K_1 and K_2 be simplicial subdivisions of P. Show that K_1 and K_2 have simplicially isomorphic subdivisions.

3.5 The Euler Characteristic

The Euler characteristic is a numerical invariant of compact surfaces that helps distinguish between non-homeomorphic surfaces. Although we are ultimately interested in topological surfaces, for convenience we start with arbitrary 2-complexes. If K is a 2-complex, let $f_0(K)$, $f_1(K)$ and $f_2(K)$ denote the number

of 0-simplices, 1-simplices and 2-simplices of K respectively. (In many texts it is standard to write V for $f_0(K)$, E for $f_1(K)$ and F for $f_2(K)$.)

Example 3.5.1. The tetrahedron, pictured in Figure 3.1.1 (i), has $f_0(K) = 4$, $f_1(K) = 6$ and $f_2(K) = 4$. The octahedron, pictured in Figure 3.4.1 (i), has $f_0(K) = 6$, $f_1(K) = 12$ and $f_2(K) = 8$. $\Diamond$

We wish to associate a single numerical invariant to each 2-complex; a good guess is to use some combination of the numbers $f_0(K)$, $f_1(K)$ and $f_2(K)$. Since we are ultimately concerned with topological surfaces, if two 2-complexes triangulate the same topological surface then we would like the combination of $f_0(K)$, $f_1(K)$ and $f_2(K)$ to be the same for both 2-complexes, even if each of $f_0(K)$, $f_1(K)$ and $f_2(K)$ are different for the two 2-complexes. For example, both the tetrahedron and the octahedron triangulate the 2-sphere. It is apparent that $f_0(K) + f_1(K) + f_2(K)$ is different for these two simplicial surfaces, so this sum is not useful. Before reading on try playing around with $f_0(K)$, $f_1(K)$ and $f_2(K)$ for the tetrahedron, the octahedron and the icosahedron to see if you can come up with some combination of these numbers that yield the same result for all three of these 2-complexes.

Euler hit upon the number $f_0(K) - f_1(K) + f_2(K)$. This number equals 2 for the tetrahedron, the octahedron and the icosahedron. In fact, it will turn out that this alternating sum is always 2 for any 2-complex that triangulates a 2-sphere. By contrast, for the 2-complex in Figure 3.3.5 (which triangulates the torus), the sum $f_0(K) - f_1(K) + f_2(K)$ is zero. We give this sum a name in the following definition, which applies to all 2-complexes, and not just to simplicial surfaces.

Definition. Let K be a 2-complex in $\mathbb{R}^n$. The **Euler characteristic** of K, denoted $\chi(K)$, is the integer

$$\chi(K) = f_0(K) - f_1(K) + f_2(K). \quad \Diamond \tag{3.5.1}$$

There is no comparably simple geometric way to calculate the Euler characteristic of a topological surface. Proceeding indirectly, we could start with any compact topological surface, find a simplicial complex that triangulates it, and then compute the Euler characteristic of the simplicial complex. (Compactness is crucial here, since a non-compact surface would need an infinite simplicial complex to triangulate it.) Since any compact surface can be triangulated by many different simplicial complexes, we ask whether different Euler

characteristics could be obtained from these different simplicial complexes. The following theorem shows, remarkably enough, that the answer is no. The proof of this theorem uses Corollary 3.7.3; the results in Section 3.7, delayed to avoid a digression at this point, do not make use of Theorem 3.5.2 or any subsequent result in sections Section 3.5 and Section 3.6.

Theorem 3.5.2. *Let Q be a compact topological surface, and suppose that K_1 and K_2 are simplicial surfaces which triangulate Q. Then $\chi(K_1) = \chi(K_2)$.*

Proof. By Theorem 3.4.5 we know that K_1 and K_2 have simplicially isomorphic subdivisions. It is straightforward to see that simplicially isomorphic simplicial complexes have equal Euler characteristics; it follows from Corollary 3.7.3 that $\chi(K_1) = \chi(K_2)$. □

Because of Theorem 3.5.2 we can make the following definition.

Definition. Let Q be a compact topological surface. The **Euler characteristic** of Q, denoted $\chi(Q)$, is defined by setting $\chi(Q) = \chi(K)$, where K is any simplicial surface that triangulates Q. ◊

Example 3.5.3. (1) Since $\chi(\text{tetrahedron}) = 2$ it follows that $\chi(S^2) = 2$.

(2) We continue Examples 3.3.9 and 3.3.11, in which a simplicial complex K' that triangulates the torus is constructed via an admissible partition of the vertices of the simplicial complex K, as shown in Figure 3.3.6 (ii), and has underlying that space as a disk. Although we may not be able to visualize K', we can count its simplices. Since none of the 2-simplices of K are glued to each other in the construction of K', we see that $f_2(K') = f_2(K) = 18$. The number of 0-simplices of K' is the number of sets in the partition $\mathcal{V}$ of $K^{((0))}$, and thus $f_0(K') = 9$. The 1-simplices of K not contained in $\operatorname{Bd} K$ are not glued to anything, and the 1-simplices of K contained in $\operatorname{Bd} K$ are glued in pairs. Hence $f_1(K') = f_1(K) - \frac{1}{2} f_1(\operatorname{Bd} K) = 27$. Therefore $\chi(T^2) = \chi(K') = 9 - 27 + 18 = 0$. ◊

The computation of $\chi(T^2)$ in the above example might seem needlessly complicated, since we can construct concrete simplicial complexes in $\mathbb{R}^3$ that triangulate T^2, but the method of the above example can be applied to surfaces such as P^2 and K^2 as well, for which ad hoc constructions in $\mathbb{R}^3$ cannot be used; such a computation is used in Exercise 3.5.2.

Finally, we need to show how connected sum affects the Euler characteristic.

Proposition 3.5.4. *Let Q_1 and Q_2 be compact surfaces in $\mathbb{R}^n$. Then*

$$\chi(Q_1 \# Q_2) = \chi(Q_1) + \chi(Q_2) - 2.$$

Proof. Let K_1 and K_2 be simplicial surfaces in $\mathbb{R}^n$ that triangulate Q_1 and Q_2, respectively. We start by using the method of Lemma 3.3.10 to construct a simplicial surface K that triangulates $Q_1 \# Q_2$. Let $\sigma_i = \langle a_i, b_i, c_i \rangle$ be a 2-simplex of K_i for each $i = 1, 2$. Observe that σ_i is a disk, that $K_i - \{\sigma_i\}$ is a simplicial complex and that $|K_i - \{\sigma_i\}| \approx Q_i - \operatorname{int} \sigma_i$. By moving K_1 if necessary we may assume that $|K_1|$ and $|K_2|$ are disjoint. Let L be the simplicial complex in $\mathbb{R}^n$ that is the union of $K_1 - \{\sigma_1\}$ and $K_2 - \{\sigma_2\}$. Let $\mathcal{V}$ be the partition of $L^{((0))}$ consisting of the three pairs $\{a_1, a_2\}$, $\{b_1, b_2\}$ and $\{c_1, c_2\}$, and single-element sets containing every other 0-simplex of L. It is straightforward to see that $\mathcal{V}$ is an admissible partition, and that if $\mathcal{V}(\mathcal{P})$ is the induced partition of $|L|$ then the identification space of $|L|$ and $\mathcal{V}(\mathcal{P})$ is homeomorphic to $Q_1 \# Q_2$. Now let K be the simplicial complex, the existence of which is guaranteed by Lemma 3.3.10 (i) applied to L and $\mathcal{V}$. It follows from Lemma 3.3.10 that K triangulates $Q_1 \# Q_2$.

We see from the construction of K that

$$\begin{aligned} f_0(K) &= f_0(L) - 3 = f_0(K_1) + f_0(K_2) - 3 \\ f_1(K) &= f_1(L) - 3 = f_1(K_1) + f_1(K_2) - 3 \\ f_2(K) &= f_2(L) = f_2(K_1) + f_2(K_2) - 2. \end{aligned}$$

Hence

$$\begin{aligned} \chi(Q_1 \# Q_2) = \chi(K) &= f_0(K) - f_1(K) + f_2(K) \\ &= (f_0(K_1) + f_0(K_2) - 3) - (f_1(K_1) + f_1(K_2) - 3) \\ &\quad + (f_2(K_1) + f_2(K_2) - 2) \\ &= (f_0(K_1) - f_1(K_1) + f_2(K_1)) + (f_0(K_2) - f_1(K_2) + f_2(K_2)) - 2 \\ &= \chi(K_1) + \chi(K_2) - 2 = \chi(Q_1) + \chi(Q_2) - 2. \quad \square \end{aligned}$$

Exercises

3.5.1. Compute the Euler characteristics for the 2-complexes shown in Figure 3.5.1.

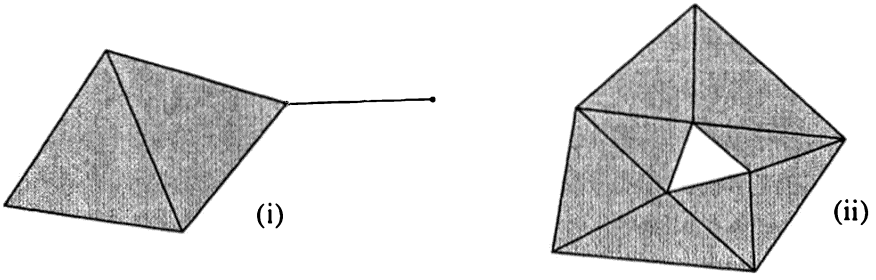


Figure 3.5.1

3.5.2*. Compute $\chi(K^2)$ and $\chi(P^2)$.

3.5.3*. Suppose $Q_1 \subset \mathbb{R}^n$ and $Q_2 \subset \mathbb{R}^m$ are homeomorphic compact surfaces. Show that $\chi(Q_1) = \chi(Q_2)$.

3.5.4*. Prove that no two of the surfaces listed in Theorem 2.6.7 are homeomorphic.

The following exercise discusses the Euler characteristics for polyhedral surfaces (for which Euler characteristics are sometimes easier to compute than simplicial surfaces).

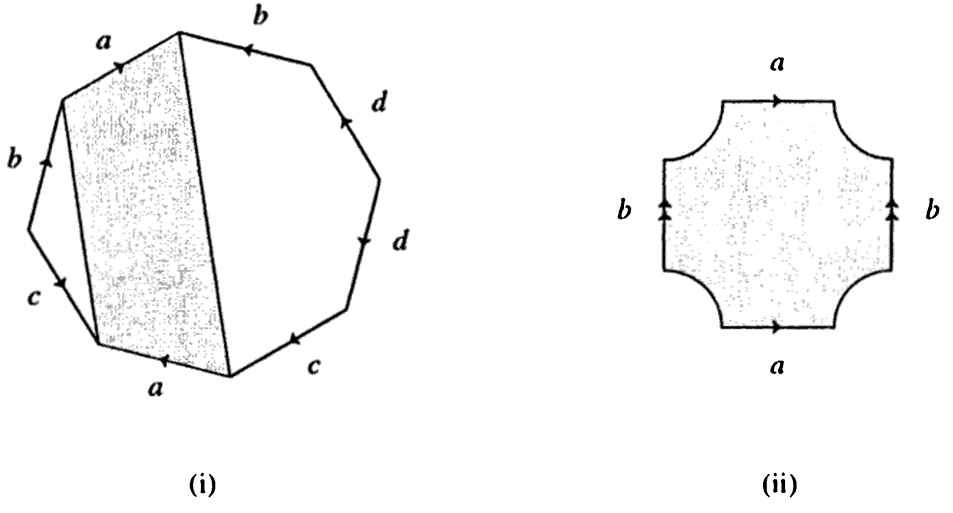
3.5.5. If P is a two-dimensional cell complex we can define $f_0(P)$, $f_1(P)$ and $f_2(P)$ as for simplicial complexes, except that $f_2(P)$ now means the number of 2-cells. Define the Euler characteristic of P by the usual formula $\chi(P) = f_0(P) - f_1(P) + f_2(P)$. Show that $\chi(P) = \chi(|P|)$, where the latter is computed as above by using a triangulation of the compact topological surface $|P|$. Verify that this result works for the surface of a cube.

3.6 Proof of the Classification of Compact Connected Surfaces

We now have all the tools for our proof of the classification of compact connected surfaces, Theorem 2.6.7. For convenience, we use the term "a hole" in an object to mean the result of removing the interior of a disk from the object. Intuitively, we might attempt to prove the classification theorem by looking for a projective plane with a hole (that is, a Möbius strip) or a torus with a hole (called a punctured torus) sitting inside a given surface, cutting it out of the surface, and

continuing this process until nothing is left. We would need to know that this process will terminate eventually, and to do so we might proceed by induction. It is not obvious at first glance on what number we will induct. We will use the clever inductive argument due to [BG], which is a variant of one of the standard proofs (as in [MS1]). By Theorem 2.4.3 (ii) every compact connected surface in Euclidean space is obtained from a polygonal disk and a gluing scheme for the edges of the polygonal disk; the induction will be on the number of edges of such polygonal disks.

Our proof works as follows. Let $Q \subset \mathbb{R}^n$ be a compact connected surface, and suppose Q is obtained from a polygonal disk D and gluing scheme S for the edges of D. If we wish to find a Möbius strip or a punctured torus in Q, how would we recognize it in the disk D? Since a Möbius strip is made from a rectangular strip with its ends glued with a twist, we should look for a strip connecting two edges of D that are matched up by S, and such that the arrows on these edges would cause the strip to be glued with a twist. See Figure 3.6.1 (i). We can similarly look for a punctured torus in Q by looking for something in D that becomes a punctured torus when glued; an unglued version of a punctured torus is shown in Figure 3.6.1 (ii). After locating the appropriate subset of D we will examine what remains after removing the subset, yielding the inductive step.



(i) (ii)

Figure 3.6.1

Proof of Theorem 2.6.7. There are really two parts to the theorem: (1) that every topological surface is homeomorphic to one in the list given in the statement of the theorem, and (2) that all the surfaces in the list are distinct. We will prove part (1) here; part (2) is proved in Exercise 3.5.4. Let $Q \subset \mathbb{R}^n$ be a compact connected surface. Rather than showing part (1) directly, we will prove the apparently weaker statement that Q is homeomorphic to the sphere or to a connected sum of tori and projective planes combined; we leave it to the reader to verify that a connected sum of tori and projective planes is in fact always homeomorphic to either a connected sum of only tori or a connected sum of only projective planes (the trick is to use Lemma 2.6.5).

By Theorem 2.4.3 (ii) there is a polygonal disk D and a gluing scheme S for the edges of D such that Q is obtained from D and S. Let n be the number of sides of the polygonal disk D. We proceed by induction on n, where the statement proved by induction is that every surface obtained from a polygonal disk with n sides is homeomorphic to the sphere or a connected sum of tori and projective planes. As mentioned previously, the number n must be even. For the first step in the proof by induction we look at $n = 2$. There are exactly two cases for what D and S could be, as seen in Figure 3.6.2; in part (i) of the figure the surface is S^2, and in part (ii) the surface is P^2. From now on assume that $n \geq 4$ and that the inductive hypothesis holds for all surfaces obtained from polygonal disks with fewer than n sides.

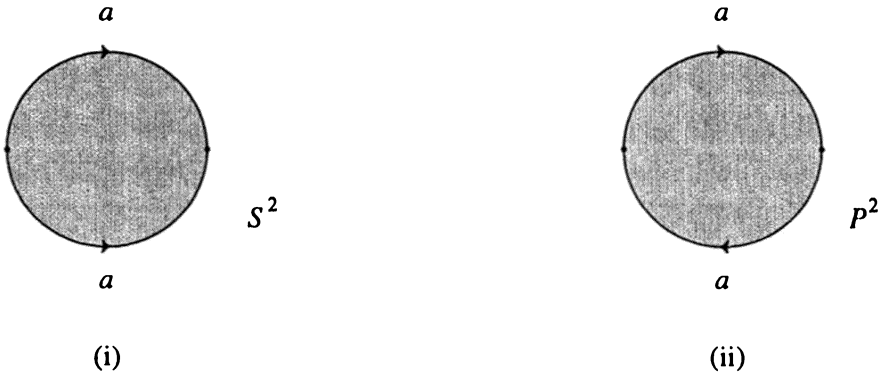


Figure 3.6.2

There are four cases to be considered (some with subcases), though they are all quite similar to one another, and we will only go over some of them

in detail. In each case we proceed in the same fashion, which in outline form consists of cutting the disk D into various pieces, reassembling the pieces into two parts using the gluing scheme S, and then observing that the two resulting parts are surfaces with holes and that the original surface is the connected sum of the two parts (once their holes are plugged up). Before proceeding we need the following straightforward observations.

Observation #1: Suppose we take the disk D and cut it in two along a line as in Figure 3.6.3. We label the two new edges that result from the cut so that gluing the new edges as labeled would undo the cut. See Figure 3.6.3. We thus obtain two polygonal disks, D_1 and D_2, together with a gluing scheme S' for their combined set of edges. The result of gluing the edges of $D_1 \cup D_2$ by the gluing scheme S' will be the same as the result of gluing the original disk D by the original gluing scheme S, namely our surface Q. The same result holds if we make any finite number of cuts in the disk D.

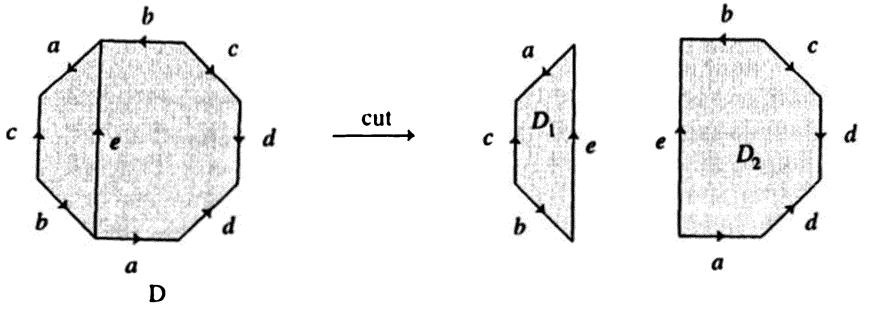


Figure 3.6.3

Observation #2: Suppose we make some cuts as in Observation #1, ending up with a number of disks and a gluing scheme for the edges of all the disks. Instead of performing all the gluing at once, we could first glue some of the pairs of edges or collections of vertices (as mandated by the gluing scheme), and only then glue the rest of the edges and vertices. The order of the gluing does not matter. For example, suppose that after some cutting as in Observation #1 we obtain two disks D_1 and D_2 with gluing scheme as indicated in Figure 3.6.4 (i). Note that the two vertices labeled A in D_1 will be identified when the edges labeled a are glued. We can paste together the two vertices labeled A, to obtain D_1' and D_2 as in Figure 3.6.4 (ii). Although we no longer have disks,

the result of gluing the new disks and disks with holes by the induced gluing scheme will still yield our original surface Q.

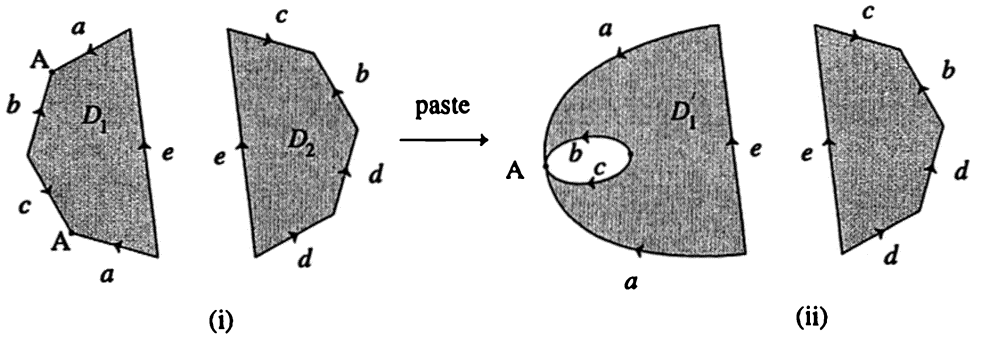


Figure 3.6.4

Observation #3: Suppose we have a disk D_1 with one hole as might arise in Observation #2, and it happens that all the edges of D_1 along the outside boundary of D_1 are glued to one another via the gluing scheme. See Figure 3.6.5 (i). Then the result of gluing all the edges on the outside boundary of D_1, but leaving the edges of the inside boundary unglued, will yield a surface with a hole in it, the same as would be obtained by first filling in the hole in D_1, then gluing as usual to obtain a surface, and then cutting out the hole. See Figure 3.6.5 (ii).

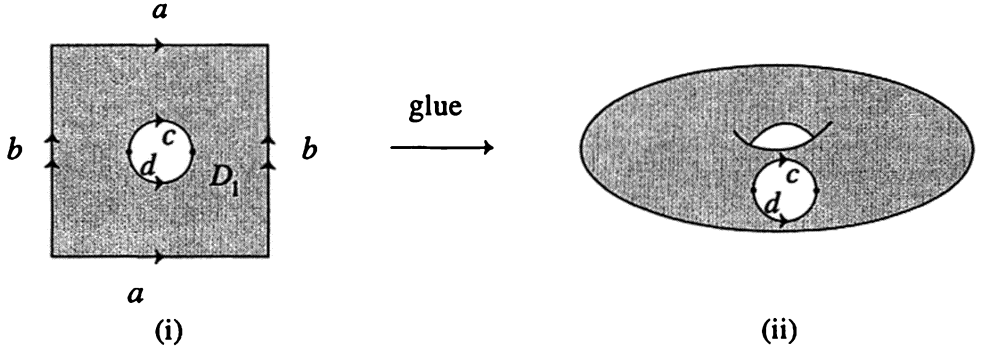


Figure 3.6.5

Our four cases depend upon whether certain phenomena occur with respect to the edges of D and the gluing scheme S. In a gluing scheme each edge is oriented by an arrow; a pair of edges that are glued is said to be **like-oriented** if they both point clockwise or they both point counterclockwise along the boundary of D, as in Figure 3.6.6 (i), and the pair is **unlike-oriented** if one points clockwise and the other points counterclockwise, as in Figure 3.6.6 (ii).

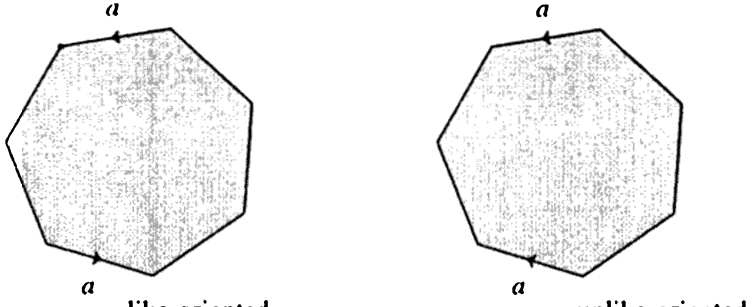


Figure 3.6.6

The following four cases exhaust all possibilities.

Case #1: There is a like-oriented pair of glued edges (the edges might or might not be adjacent).

Case #2: All pairs of edges identified by the gluing scheme are unlike-oriented, and there is a pair of adjacent edges that are glued.

Case #3: All pairs of edges identified by the gluing scheme are unlike-oriented, no pair of adjacent edges are glued, and there is a pair of edges $\{a, a'\}$ such that both members of every other pair of glued edges lie in the same component of $\partial D - \{a, a'\}$.

Case #4: None of the above, so that all pairs of edges identified by the gluing scheme are unlike-oriented, no pair of adjacent edges are glued, and there is no pair of edges $\{a, a'\}$ such that both members of every other pair of glued edges lie in the same component of $\partial D - \{a, a'\}$.

We now consider each case.

Case #1: This case corresponds to finding a Möbius strip in the surface. There are two subcases, depending upon whether the like-oriented pair of glued edges are adjacent or not. See Figure 3.6.7 for the two possibilities.

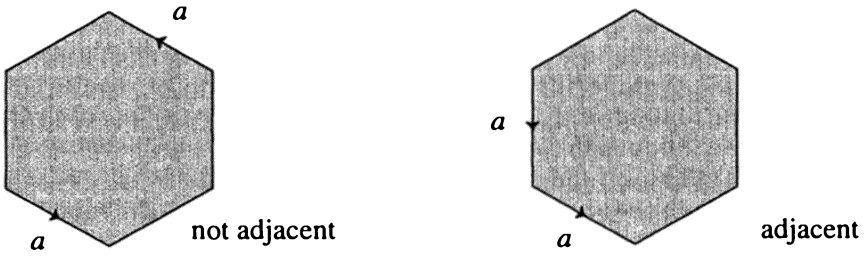


Figure 3.6.7

Subcase (a): The like-oriented edges are not adjacent. The two edges under consideration are labeled a and a', as in Figure 3.6.8 (i). Note that the two points labeled A will be glued to each other when a and a' are glued, as are the two points labeled B. We make two cuts in D, along lines b and c, as in Figure 3.6.8 (i), dividing D into three pieces labeled I, II and W. Take pieces I and II and join them by flipping piece II over, and gluing the points labeled A and B in piece I to the similarly labeled points in piece II. See Figure 3.6.8 (ii). The result of this operation is a disk with a hole D_1 (the boundary of the hole intersects the boundary of the disk, but there is nothing wrong with that). Since D had n edges, it is easy to see that D_1 has $n - 2$ edges. Clearly all the edges of D_1 along the outside boundary are glued to one another via the gluing scheme. By Observation #3 gluing the outside edges of D_1 will yield a surface with a hole. Call this surface with a hole Q_1. By the inductive hypothesis Q_1 is homeomorphic to either a sphere with a hole or a connected sum of tori and projective planes with a hole. The boundary of the hole in Q_1 is a 1-sphere consisting of the two edges labeled b and c, as in Figure 3.6.8 (ii).

Gluing the edges of W labeled a and a' will yield a Möbius strip M. The boundary of M is a 1-sphere consisting of the two edges labeled b and c, as in Figure 3.6.8 (iii). Finally, using Observation #2 we see that the result of attaching Q_1 and M along their boundaries as indicated by the labeling of the edges in their boundaries yields a surface homeomorphic to our original surface Q. However, since Q_1 is homeomorphic to either a sphere with a hole or a connected sum of tori and projective planes with a hole, and since M is

homeomorphic to a projective plane with a hole, attaching Q_1 and M along their boundaries is homeomorphic to the connected sum of either a sphere and a projective plane or a connected sum of tori and projective planes with another projective plane. Since the connected sum of a sphere with a projective plane is a projective plane by Lemma 2.6.2 (iii), it now follows that Q is homeomorphic to a connected sum of tori and projective planes as desired.

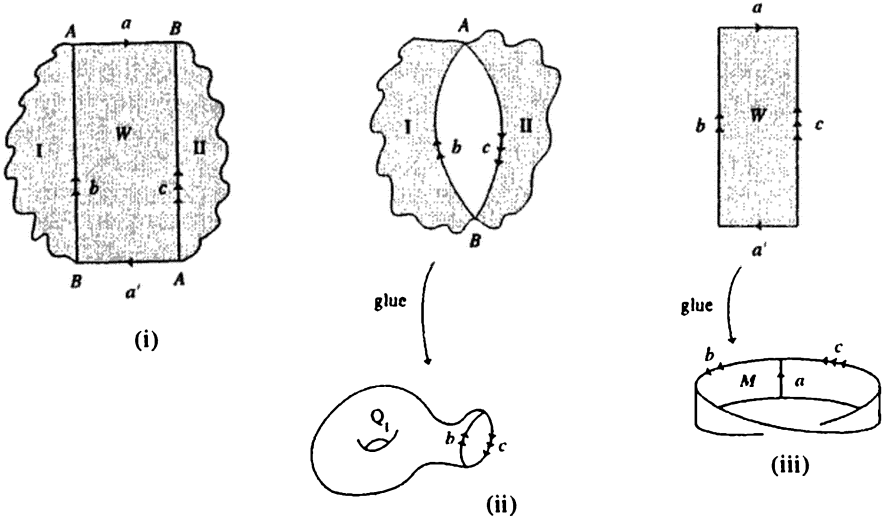


Figure 3.6.8

Subcase (b): The like-oriented edges are adjacent. The strategy here is very similar to subcase (a), and we will not go into detail. The construction is shown in Figure 3.6.9. We make one cut in D along line b, dividing D into two pieces labeled I and W. We glue the two points labeled A in piece I, the result of which is a disk with a hole D_1 with $n-2$ edges in the outside boundary. The piece labeled W turns out to be a Möbius strip just as in the previous case, though this time it is not quite as obvious. The trick is to cut W into two pieces and rearrange, gluing a to a', as in Figure 3.6.9 (iii). The rest of the argument is just as in subcase (a).

Case #2: This case corresponds to cutting a disk out of the surface, which can be done in any surface, but in this case can be done so that the inductive hypothesis is then applicable. The situation is pictured in Figure 3.6.10, and is almost identical to subcase (b) of Case #1, the difference being that the piece labeled

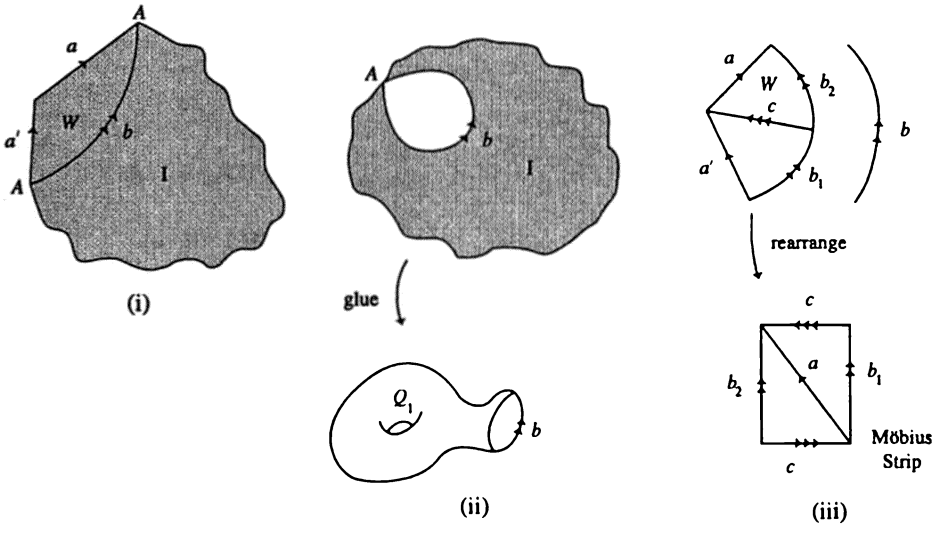


Figure 3.6.9

W yields a disk rather than a Möbius strip. Using Exercise A2.2.11 and Lemma 2.6.2 (iii) it follows that the surface Q is homeomorphic to either a sphere or a connected sum of tori and projective planes.

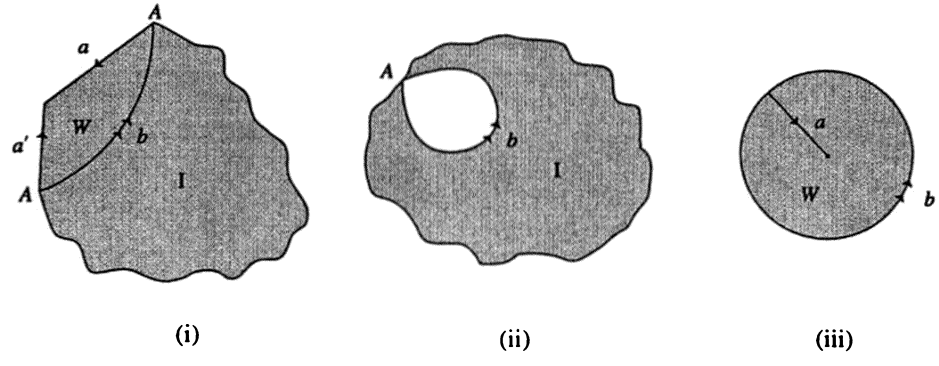


Figure 3.6.10

Case #3: This case corresponds to breaking the surface into the connected sum of two pieces, to each of which the inductive hypothesis applies. The situation is shown in Figure 3.6.11. The assumption of this case is that both members

of every pair of glued edges other than $\{a, a'\}$ lie in the same component of $\partial D - \{a, a'\}$. Since a and a' are not adjacent by assumption, there are at least two edges in each component of $\partial D - \{a, a'\}$ (there cannot be only one in either component since every edge needs another one to be glued to). We make one cut in D along line b as in Figure 3.6.11, dividing D into two pieces labeled I and II. In each of I and II we glue the two points labeled A, the result of which are disks with holes D_1 and D_2, each with at most $n - 2$ edges in the outside boundary. All the edges in the outside boundary of each of these disks with holes are glued under the gluing scheme to other edges in the same disk with a hole. We can thus apply the inductive hypothesis to each of D_1 and D_2; gluing the outside edges of each of D_1 and D_2 will yield surfaces Q_1 and Q_2, each of which is homeomorphic to either a sphere with a hole or a connected sum of tori and projective planes with a hole. As before the result of gluing Q_1 and Q_2 along their boundaries as indicated by the labeling of the edge on each boundary yields a surface homeomorphic to our original surface Q. The surface Q is thus seen to be homeomorphic to either a sphere or a connected sum of tori and projective planes.

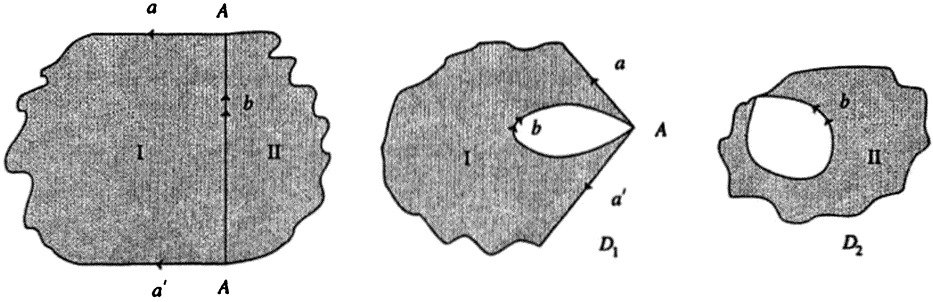


Figure 3.6.11

Case #4: This case corresponds to finding a punctured torus in the surface. Choose any pair of glued edges $\{a, a'\}$, which by hypothesis are unlike-oriented and not adjacent; further, by hypothesis, there must be another pair of glued edges $\{b, b'\}$ such that b is in one component of $\partial D - \{a, a'\}$ and b' is in the other component (if there were no such pair $\{b, b'\}$ then we would be in Case #3). There are six generic possibilities for what can happen, depending upon

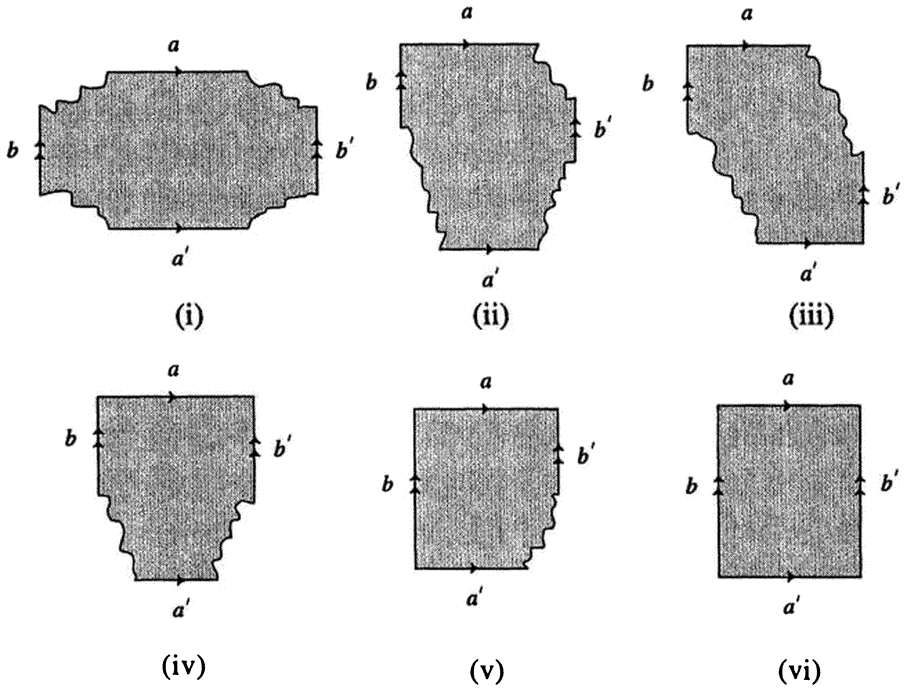


Figure 3.6.12

which, if any, of the edges a, a', b and b' touch each other; the possibilities are shown in Figure 3.6.12.

Option (vi) in Figure 3.6.12 is simply a torus, and there is nothing more to prove. The other five options are all quite similar, and we will only discuss option (i), leaving the details of options (ii)–(v) to the reader. We make four cuts in D, along lines p, q, r and s, as in Figure 3.6.13 (i), dividing D into five pieces labeled I–IV and W. Observe that the pairs of points in Figure 3.6.13 (i) that have the same labels are glued to each other when a is glued to a' and b is glued to b'. We now take pieces I–IV and join them by gluing all pairs of points that have the same label. See Figure 3.6.13 (ii). The result of this operation is a disk with a hole D_1. We see that D_1 has $n - 4$ edges, and all the edges of D_1 along the outside boundary are glued to one another via the gluing scheme. As before, the inductive hypothesis implies that gluing the outside edges of D_1 will yield a surface Q_1 that is homeomorphic to either a sphere with a hole or a connected sum of tori and projective planes with a hole. The piece labeled W

yields a punctured torus when a is glued to a' and b is glued to b', the details being left to the reader. The rest of the proof is similar to subcase (a) of Case #1. □

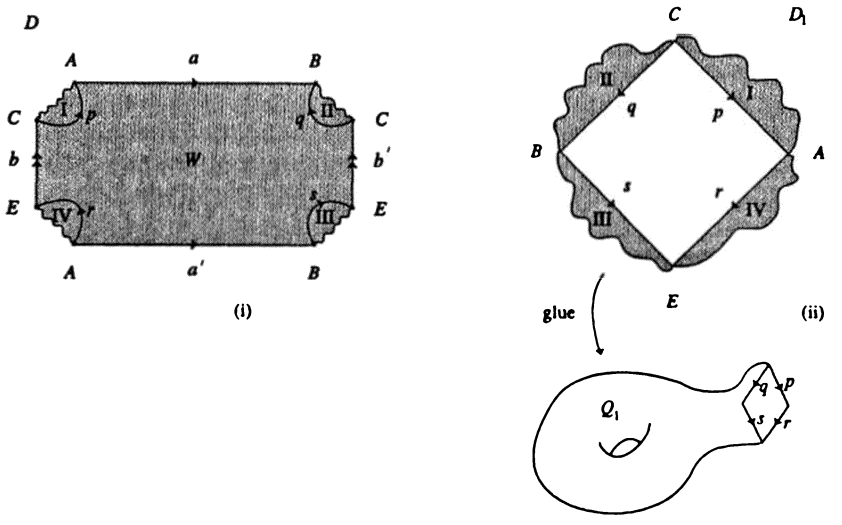


Figure 3.6.13

The following corollary can be deduced straightforwardly from the classification of surfaces.

Corollary 3.6.1. *Two compact surfaces $Q_1 \subset \mathbb{R}^n$ and $Q_2 \subset \mathbb{R}^m$ are homeomorphic iff (1) they are both orientable or both non-orientable, and (2) $\chi(Q_1) = \chi(Q_2)$.*

3.7 Simplicial Curvature and the Simplicial Gauss-Bonnet Theorem

In addition to using the structure of simplicial surfaces to give a proof of the Classification Theorem for Surfaces — a topological result — we can also use the simplicial structure to investigate geometric properties of surfaces in $\mathbb{R}^n$, which depends upon the particular way in which the surface sits in $\mathbb{R}^n$. Here we will look at the curvature of simplicial surfaces. Although the concept of

curvature is far more substantial (and subtle) for smooth surfaces, as we will see in Chapter 6, there is nonetheless a very simple but valid theory for simplicial surfaces, including an analog of the Gauss–Bonnet Theorem (to be proved in the smooth case in Chapter 8).

What properties should a formula for calculating the curvature of simplicial surfaces have? In contrast to the smooth case (for example S^2) where the surface is possibly curved at all points, in the simplicial case the only interesting points as far as curvature is concerned are the vertices; at the interiors of 2-simplices the surface is flat, and at the interiors of 1-simplices the surface always looks like a "ridge," which will also turn out to possess no curvature. Curvature of a simplicial surface will be given by assigning to each 0-simplex of the surface a number that will describe how the surface is curving at that point. If K is a simplicial surface in $\mathbb{R}^n$, we can thus think of curvature as a function $d\colon K^{((0))} \to \mathbb{R}$. However the function d is defined, we should expect d to have the following three properties.

(1) If a 0-simplex v of K has an open neighborhood in $|K|$ which is flat (that is, the neighborhood is contained in a plane), then we should have $d(v) = 0$.

(2) If v and w are 0-simplices of K such that v has an open neighborhood in $|K|$ that is, intuitively, more of a sharp peak than an open neighborhood of w, then we should have $d(v) > d(w)$. See Figure 3.7.1.

(3) The numbers $d(v)$ should be "intrinsic." This concept, touched on briefly in Section 2.5, helps make curvature useful. Imagine a small creature living on a surface that is the creature's whole universe. Though we in $\mathbb{R}^3$ can observe the surface from outside of it, this creature cannot see off the surface, or through it, but only along it; its lines of vision curve along the surface. The creature can make various geometric measurements on the surface such as lengths, angles and areas. A quantity associated with the above surface is called *intrinsic* if it could be calculated by such a creature from the measurements it is capable of making. In other words, a quantity is intrinsic if one does not have to step off the surface in order to calculate it. For example, the area of a surface is intrinsic. In a simplicial surface, a quantity that can be calculated using only the lengths of the 1-simplices of the surface is certainly intrinsic.

Our definition of the function d is actually quite simple. Note that if a 0-simplex v of K has a flat open neighborhood, then it is certainly the case that the sum of the angles at v in all the 2-simplices of K containing v is 2π. We can view curvature as a measure of how a surface deviates from being planar.

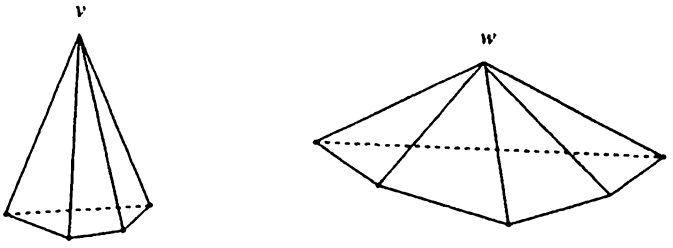


Figure 3.7.1

Definition. Let K be a simplicial surface, and let $v \in K$ be a 0-simplex. If $\sigma \in K$ is a 2-simplex containing v, let $\angle(v, \sigma)$ denote the angle at v in σ. The **curvature** of K at v is defined to be the number $d(v)$ given by

$$d(v) = 2\pi - \sum_{\eta v} \angle(v, \eta),$$

where the η are the 2-simplices of K containing v. $\Diamond$

Example 3.7.1. Let K be a regular tetrahedron, so that the 2-simplices of K are all equilateral triangles. We see that $d(v) = 2\pi - 3 \cdot \frac{\pi}{3} = \pi$ for each vertex v of K. $\Diamond$

It is not hard to see that all three properties for $d(v)$ mentioned above indeed hold (for property (3) the Law of Cosines is needed to compute the angles in the 2-simplices knowing the lengths of their sides). The definition for curvature given above, called the "angle defect," goes back at least as far as Descartes (see [FE]). In this manuscript Descartes discusses a rather remarkable fact: If K is any simplicial complex in $\mathbb{R}^3$ such that $|K|$ is homeomorphic to S^2, then the sum of the curvatures at all the 0-simplices of K (called the total curvature) is always 4π. The number 4π is not only independent of the way in which the simplicial surface K sits in $\mathbb{R}^3$, it is independent of which simplicial surface is used as long as the underlying surface is homeomorphic to S^2. The following theorem shows that Descartes' result can be generalized to simplicial surfaces with arbitrary underlying spaces as long as the 4π is appropriately modified.

Theorem 3.7.2 (Simplicial Gauss–Bonnet Theorem). *Let K be a simplicial surface in $\mathbb{R}^n$. Then*

$$\sum_{v \in K^{((0))}} d(v) = 2\pi \chi(K).$$

Proof. From Theorem 3.4.1 we know that in any simplicial surface each 1-simplex is the face of precisely two 2-simplices. Given that each 2-simplex has three 1-simplices as faces we see that

$$3f_2(K) = 2f_1(K). \tag{3.7.1}$$

We now compute

$$\begin{aligned}
\sum_{v\in K^{((0))}} d(v) &= \sum_{v\in K^{((0))}} \Big\{2\pi - \sum_{\eta v} \angle(v,\eta)\Big\} \\
&= \sum_{v\in K^{((0))}} 2\pi - \sum_{v\in K^{((0))}} \sum_{\eta v} \angle(v,\eta) = 2\pi f_0(K) - \sum_{\eta\in K^{((2))}} \sum_{v\in\eta} \angle(v,\eta) \\
&= 2\pi f_0(K) - \sum_{\eta\in K^{((2))}} \pi \quad \text{since the sum of the angles in a triangle is } \pi \\
&= 2\pi f_0(K) - \pi f_2(K) = 2\pi f_0(K) - 3\pi f_2(K) + 2\pi f_2(K) \\
&= 2\pi f_0(K) - 2\pi f_1(K) + 2\pi f_2(K) \qquad \text{by Equation 3.7.1} \\
&= 2\pi\chi(K). \quad \square
\end{aligned}$$

The following corollary is needed for the proof of Theorem 3.5.2.

Corollary 3.7.3. *Let K be a 2-complex in $\mathbb{R}^n$ and let L be a subdivision of K. Then $\chi(L) = \chi(K)$.*

Proof. Let K be a simplicial surface in $\mathbb{R}^n$ and let L be a subdivision of K. The curvature of L at each 0-simplex v of L is computed as follows: If v is a 0-simplex of K, then the curvature of L at v is the same as the curvature of K at v; if v is not a 0-simplex of K (so it is in the interior of a 1-simplex or 2-simplex of K), then the curvature of L at v is zero. It follows that the total curvatures for K and L are equal. From Theorem 3.7.2 we then deduce that $\chi(L) = \chi(K)$. $\square$

Exercises

In Exercises 1–4 the polyhedral version of the Gauss–Bonnet Theorem is developed.

3.7.1. If you have not already seen it, discover and prove the formula for the sum of the interior angles at the vertices of a polygonal disk with n sides.

3.7.2. Discover and prove the analog of Equation 3.7.1 for polyhedral surfaces.

3.7.3. We can define the curvature at the vertices of a polyhedral surface by the same angle defect formula as for simplicial surfaces. Prove the Gauss-Bonnet Theorem for polyhedral surfaces.

3.7.4*. Though it is standard to think of the curvature of simplicial surfaces as being entirely concentrated at the vertices, an even closer analogy between the simplicial Gauss–Bonnet Theorem and the smooth version of the theorem (to be proved in Chapter 8) can be constructed as follows. Let K be a simplicial surface in $\mathbb{R}^n$. For each 0-simplex $v \in K$ let

$$k(v) = \frac{\mathrm{d}(v)}{\frac{\text{Area of star}(v, K)}{3}}.$$

A simplexwise linear function $k\colon |K| \to \mathbb{R}$ is then defined by extending k affine linearly over each simplex of K. Show that

$$\int_{|K|} k(x)\, dA = 2\pi\chi(K),$$

where the integral is the standard Riemann integral of a continuous function.

3.7.5. Show that any simplicial surface has an even number of 2-simplices.

3.7.6. Let ζ be any integer such that $\zeta \leq 2$. In theory we can take the collection of all compact connected simplicial surfaces with Euler characteristic ζ and find the surface (or surfaces) in the collection with the fewest vertices. Call this minimal number of vertices V_ζ; thus every compact connected simplicial surface with Euler characteristic ζ has at least V_ζ vertices.

Let a, b and c be positive integers that satisfy

$$a \geq V_\zeta, \qquad a - b + c = \zeta, \qquad 3c = 2b.$$

Show that there is a simplicial surface K in some $\mathbb{R}^n$ such that

$$\chi(K) = \zeta, \qquad f_0(K) = a, \qquad f_1(K) = b, \qquad f_2(K) = c.$$

Are the simplicial surfaces you have found unique?

3.8 Simplicial Disks and the Brouwer Fixed Point Theorem

Our goal in this section is to present a proof of the two-dimensional version of one of the most famous theorems in topology, the Brouwer Fixed Point Theorem. We already saw a proof of the one-dimensional version of this theorem in Section 1.5, but in dimensions higher than 1 there is no similarly simple proof since there is no appropriate analog of the Intermediate Value Theorem. Our proof in the two-dimensional case uses a modification of the Sperner Lemmas approach.

We start with a brief discussion of simplicial complexes with underlying spaces that are disks. See Figure 3.8.1.

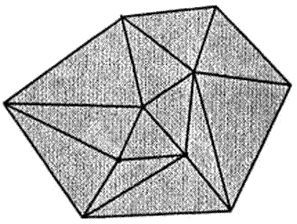


Figure 3.8.1

Definition. A **simplicial disk** is a simplicial complex K in $\mathbb{R}^n$ such that $|K|$ is a disk. The **simplicial boundary** of a simplicial disk K, denoted Bd K, is the collection of simplices $\eta \in K$ such that η is either a 1-simplex that is the face of precisely one 2-simplex, or a 0-simplex, the link of which has underlying space an arc. $\Diamond$

We have two different ways of looking at the "boundary" of K: the topological boundary $\partial|K|$ and the simplicial boundary Bd K. It is seen in Exercise 3.8.1 that these two approaches yield the same result. In order to use induction in our proof of the Brouwer Fixed Point Theorem we need simplicial disks from which we can remove 2-simplices one at a time.

Definition. Let K be a simplicial disk in $\mathbb{R}^n$. A **shelling** of K is a listing of the 2-simplices of K in an order $\sigma_1, \ldots, \sigma_m$ such that the collection $\{\sigma_1, \ldots, \sigma_k\}$ together with all the faces of these simplices forms a simplicial disk for all $k \in \{1, \ldots, m\}$. We say that K is **shellable** if it has a shelling. $\Diamond$

Example 3.8.1. For the simplicial disk K shown in Figure 3.8.2, the listing σ, τ, η is a shelling of K, whereas the listing σ, η, τ is not a shelling. Since K has at least one shelling, then it is a shellable simplicial disk. $\Diamond$

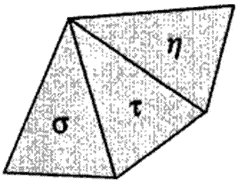


Figure 3.8.2

Three remarks on shellings. First, suppose that K is a shellable simplicial disk in $\mathbb{R}^n$, and $\sigma_1, \ldots, \sigma_m$ is a shelling of K. Then the simplicial disk formed by the collection $\{\sigma_1, \ldots, \sigma_k\}$ together with all the faces of these simplices is itself shellable for each $k \in \{1, \ldots, m\}$. Second, if Δ is a 2-simplex in $\mathbb{R}^n$ and $\delta > 0$ is any number, then there is a shellable subdivision of Δ so that the distance between any two points in a single simplex of the subdivision is less than δ; one way of obtaining such a subdivision is as in Figure 3.8.3, using sufficiently many parallel lines to slice up Δ. Finally, it can actually be shown that every simplicial disk is shellable; see [MO, p. 27] for details. However, we will not use this result, and hence will not prove it here. (Rather surprisingly, the analogous result does not hold in 3 dimensions; see [RD].)

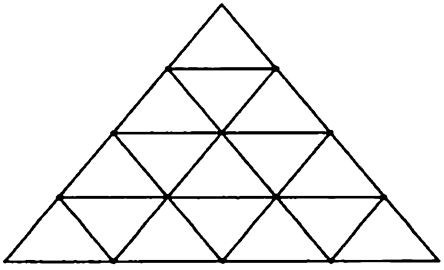


Figure 3.8.3

We now turn to the Brouwer Fixed Point Theorem.

Definition. Let $X \subset \mathbb{R}^n$ be a set, and let $f: X \to X$ be a function. A **fixed point** of f is a point $x \in X$ such that $f(x) = x$. $\Diamond$

Are there any restrictions on X and f that guarantee that every map $f: X \to X$ must have a fixed point? The function $g: S^1 \to S^1$ that rotates S^1 by 90° (either direction) has no fixed point, even though S^1 is both compact and path

connected. The map $h: D^2 \to D^2$ defined by

$$h\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{cases} \binom{-1}{0}, & \text{if } x \geq 0; \\ \binom{1}{0}, & \text{if } x < 0, \end{cases}$$

has an even simpler domain than g, but it also has no fixed point; of course, the map h is not continuous. The following remarkable result shows that the two types of problems we have just seen, namely that the function is not continuous or the space is not simple enough (for example S^1), are the essential problems.

Theorem 3.8.2 (Brouwer Fixed Point Theorem). *Any continuous map $f: D^2 \to D^2$ has a fixed point.*

This theorem is an existence theorem only; it does not tell us how many fixed points there are, nor how to find them. The Brouwer Fixed Point Theorem would work just as well if D^2 were replaced by any other disk; see Exercise 3.8.6.

One of the first encounters I had with an idea from topology was when as a boy I read about the following graphical interpretation of the Brouwer Fixed Point Theorem in the Time–Life book on mathematics. Lay two identical sheets of paper one precisely on top of the other. Take the top sheet, crumple it up any way you please (though do not tear it), and lay it on top of the bottom sheet (with no part of the top sheet off of the bottom sheet). Then the Brouwer Fixed Point Theorem implies that at least one point of the crumpled sheet must be exactly on top of its original location.

To prove the Brouwer Fixed Point Theorem we start off in the standard way by proving that this theorem is logically equivalent to the following result.

Theorem 3.8.3 (No-Retraction Theorem). *There is no continuous map $r: D^2 \to S^1$ such that $r(x) = x$ for all $x \in S^1$.*

The No-Retraction Theorem states the intuitively plausible fact that the skin of a drum cannot be pushed onto its rim without a hole being punched in the drum.

Proposition 3.8.4. *The Brouwer Fixed Point Theorem is true iff the No-Retraction Theorem is true.*

Proof. We show that the falsity of each theorem implies the falsity of the other. First assume that the No-Retraction Theorem is false, so that there is a continuous map $r: D^2 \to S^1$ such that $r(x) = x$ for all $x \in S^1$. Let $R: S^1 \to S^1$ denote

rotation of S^1 by 180°, and let $j: S^1 \to D^2$ denote the inclusion map. It is seen that $j \circ R \circ r: D^2 \to D^2$ has no fixed point, and hence the Brouwer Fixed Point Theorem is false.

Now suppose that the Brouwer Fixed Point Theorem is false, so that there is a continuous map $f: D^2 \to D^2$ with no fixed points. We define a map $r: D^2 \to S^1$ as follows. For each point $x \in D^2$, let $r(x)$ be the intersection with S^1 of the ray that starts at $f(x)$ and goes through x; if the ray intersects S^1 in two points, then one of the points of intersection is $f(x)$, and let $r(x)$ to be the other point of intersection. See Figure 3.8.4. This ray is well-defined for all points $x \in D^2$ precisely because f has no fixed points. It is not hard to see that r is continuous (because f is), and that $r(x) = x$ for all $x \in S^1$; hence the No-Retraction Theorem is false. □

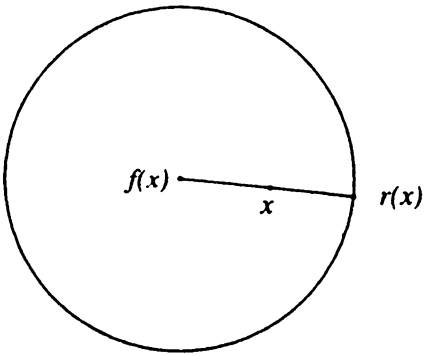


Figure 3.8.4

To prove the No-Retraction Theorem we will approximate arbitrary continuous maps with more well-behaved ones, as described in the following definition.

Definition. Let K be a simplicial complex in $\mathbb{R}^n$. A map $f: |K| \to \mathbb{R}^m$ is **simplexwise linear**, or **SL**, if the restriction of f to each simplex of K is an affine linear map. ◊

SL maps are not quite the same as simplicial maps; simplicial maps are from one simplicial complex to another and make use of the simplicial structure of both the domain and codomain, whereas SL maps have Euclidean space as the co-domain. An SL map is uniquely determined by what it does to the 0-simplices in its domain (use Lemma A.7). SL maps are always continuous (use

Exercise 1.3.10 and Lemma 1.3.6). The following lemma, a simplexwise linear version of the No-Retraction Theorem, is a variant on Sperner's First Lemma (see [LY, §43]).

Definition. Let $\Delta = \langle a, b, c \rangle$ denote a fixed 2-simplex in $\mathbb{R}^2$, let K be a simplicial disk in $\mathbb{R}^n$ and let $f: |K| \to \Delta$ be an SL map which sends each 0-simplex of K to a vertex of Δ. The map f is **boundary-odd** (respectively, **boundary-even**) if every 1-face of Δ is the image of an odd (respectively, even) number of 1-simplices of $\mathrm{Bd}\, K$. The map f is **interior-odd** (respectively, **interior-even**) if Δ is the image of an odd (respectively, even) number of 2-simplices of K. $\Diamond$

It is evident that any map as in the above definition is either interior-odd or interior-even. It is not obvious that a given map is necessarily either boundary-odd or boundary-even, though it is true (a fact we will not be using).

Lemma 3.8.5. *Let K be a shellable simplicial disk in $\mathbb{R}^n$, and let $f: |K| \to \Delta$ be an SL map which sends each 0-simplex of K to a vertex of Δ. If f is boundary-odd (respectively, even) then it is interior-odd (respectively, even).*

Proof. The proof proceeds by induction on the number p of 2-simplices of K. If $p = 1$ then K has one 2-simplex σ, and the result is quite straightforward; we will go over this case in detail nonetheless since it will save work later on. There are three cases:

Case (1). The map f sends distinct vertices of σ to distinct vertices of Δ. Hence f is injective on σ. In this case each 1-face of Δ is the image of exactly one 1-simplex of $\mathrm{Bd}\, K$, so f is boundary-odd. Since Δ is the image of σ it follows that f is also interior-odd.

Case (2). The map f sends two of the vertices of σ to the same vertex of Δ, and the third vertex of σ to a different vertex of Δ. In this case one 1-face of Δ is the image of two 1-simplices of $\mathrm{Bd}\, K$, and the other two 1-faces of Δ are not the images of any 1-simplices of $\mathrm{Bd}\, K$. Thus f is boundary-even. Since Δ is not the image of σ, it follows that f is also interior-even.

Case (3). The map f sends all vertices of σ to the same vertex of Δ. Hence neither the 1-faces of Δ nor Δ itself are contained in the image of f, and hence f is both boundary-even and interior-even.

We now assume that K has p 2-simplices (where $p \geq 2$), and that the result is true for all simplicial disks with fewer than p 2-simplices. Let $\sigma_1, \ldots, \sigma_p$ be a

shelling of K. If we let K' denote the collection $\{\sigma_1, \ldots, \sigma_{p-1}\}$ together with all the faces of these simplices, then as remarked earlier K' is a shellable simplicial disk with $p-1$ 2-simplices. Observe that the map $f|K'$ is an SL map $K' \to \Delta$ that takes each 0-simplex of K' to a vertex of Δ; for convenience we let f' denote $f|K'$. To prove the lemma it suffices to show that f' is either boundary-odd or boundary-even (using the fact that f is either boundary-odd or boundary-even), and that when it has the same (respectively, opposite) boundary-parity as f, then it has the same (respectively, opposite) interior-parity as f. Again there are three cases, corresponding to the three cases treated for $p = 1$, this time with respect to $f|\sigma_p$; we treat the first case and leave the details of the other two cases to the reader.

In the first case, suppose that f sends distinct vertices of σ_p to distinct vertices of Δ. It is easy to see that f' has the opposite interior-parity as f. For each 1-face of Δ, we observe that the number of 1-simplices of $\mathrm{Bd}(K')$ of which it is the image under f' is either one more or one less than the number of 1-simplices of $\mathrm{Bd}\, K$ of which it is the image under f (depending upon whether the face of σ_p that maps onto the 1-face of Δ is in $\mathrm{Bd}\, K$ or not). It now follows that f' is either boundary-even or boundary-odd and that its boundary-parity is the opposite of the boundary parity of f. □

The following now completes our proof of the Brouwer Fixed Point Theorem.

Proof of the No-Retraction Theorem. It suffices to prove the No-Retraction Theorem for any choice of a disk instead of D^2 (use an argument similar to Exercise 3.8.6). We will prove the No-Retraction Theorem for a 2-simplex $\Delta \subset \mathbb{R}^2$ that is an equilateral triangle with sides of length 1. Suppose that the No-Retraction Theorem were false for Δ, so that there is a continuous map $r: \Delta \to \partial\Delta$ such that $r(x) = x$ for all $x \in \partial\Delta$; we will derive a contradiction. Since Δ is compact it follows from Exercise 1.5.5 that the map r is uniformly continuous (see Exercise 1.3.7 for the definition of uniform continuity). In particular, we can find some number $\delta > 0$ such that if $x, y \in \Delta$ are any two points such that $\|x - y\| < \delta$ then $\|r(x) - r(y)\| < \frac{1}{8}$. As remarked above, we can now find a shellable subdivision K of Δ such that the distance between any two points in a single simplex of the subdivision is less than δ.

We now define an SL map $L: |K| \to \Delta \subset \mathbb{R}^2$ as follows. Pick three points $a' \in \mathrm{Int}\langle b, c\rangle$, $b' \in \mathrm{Int}\langle a, c\rangle$ and $c' \in \mathrm{Int}\langle a, b\rangle$ that are not the images under f of any 0-simplices of K and that are within $1/8$ of the midpoints of 1-faces

of Δ (given that there are only finitely many 0-simplices in K such a choice of points is always possible). See Figure 3.8.5. Divide $\partial\Delta$ into six line segments using the points a, c', b, a', c and b'. Then for each 0-simplex v of K let $L(v)$ equal a, b or c, respectively, if $r(v)$ is contained in one of the six line segments in $|\operatorname{Bd} K|$ that has a, b or c, respectively, as one of its endpoints. (Since $L(v)$ cannot be one of a', b' or c' this definition is unambiguous.) Extend L affine linearly over the 1-simplices and 2-simplices of K. We will show that L is boundary-odd and interior-even, a contradiction to Lemma 3.8.5.

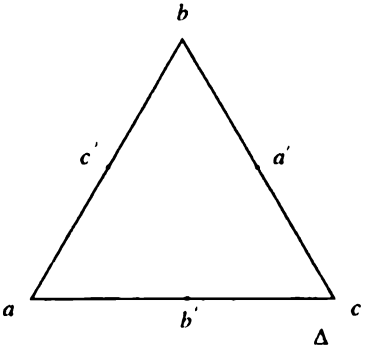


Figure 3.8.5

We first need to show that $L(|K|) \subset \partial\Delta$. It is straightforward to see that L sends any 0-simplex or 1-simplex of K into $\partial\Delta$; the only question concerns the 2-simplices of K. Let $\eta = \langle p, q, s\rangle$ be a 2-simplex of K. If two of the vertices of η are mapped to the same vertex of Δ then $L(\eta) \subset \partial\Delta$, so assume that all three vertices of η are mapped to distinct vertices of Δ. Without loss of generality assume that $L(p) = a$, $L(q) = b$ and $L(s) = c$. By choice of K it must be the case that $\|r(p) - r(q)\| < \frac{1}{8}$. It is not hard to verify that $r(p)$ and $r(q)$ are therefore both in $\langle a, b\rangle$, and are both within $1/8$ of c', and hence within $1/4$ of the midpoint of $\langle a, b\rangle$. Since $\|r(p) - r(s)\| < \frac{1}{8}$ and $\|r(q) - r(s)\| < \frac{1}{8}$, it now follows that $r(s)$ cannot be in $\langle b', c\rangle \cup \langle a', c\rangle$, a contradiction to the fact that $L(s) = c$. We therefore deduce that $L(|K|) \subset \partial\Delta$. In particular L is interior-even.

Next, we can divide the 1-sphere $\partial|K|$ into three arcs, namely $\partial|K| \cap \langle a, b\rangle$, $\partial|K| \cap \langle b, c\rangle$ and $\partial|K| \cap \langle c, a\rangle$. It is straightforward to verify from the definition of L that L maps each of these arcs into itself, and it fixes the endpoint of each of the arcs. It is not hard to see that Exercise 3.8.2 applied to each of these arcs

implies that L is boundary-odd. This completes the proof. □

Exercises

3.8.1*. Let K be a simplicial disk.

(i) Show that K is a 2-complex.

(ii) Show that every 1-simplex of K is the face of one or two 2-simplices, and that the underlying space of the link of every 0-simplex of K is either an arc or a 1-sphere.

(iii) Show that the collection of simplices $\operatorname{Bd} K$ is a subcomplex of K and $|\operatorname{Bd} K| = \partial|K|$.

(iv) Show that part (ii) of this exercise is not an "if and only if" statement, that is, there are 2-complexes which satisfy the criteria in part (ii), and yet do not have underlying spaces that are disks.

3.8.2*. This statement is the one-dimensional analog of Lemma 3.8.5. Let $a_0 < a_1 < \cdots < a_p$ be real numbers, and let $f: [a_0, a_p] \to [0, 1]$ be a map such that for each $i \in \{1, \ldots, p\}$ the value of $f(a_i)$ is either 0 or 1, and the map $f|[a_{i-1}, a_i]$ is affine linear. Show that if $f(a_0) = f(a_p)$ then $[0, 1]$ is the image of an even number of the intervals $[a_{i-1}, a_i]$, and if $f(a_0) \neq f(a_p)$ then $[0, 1]$ is the image of an odd number of the intervals $[a_{i-1}, a_i]$.

3.8.3. Does every continuous map $T^2 \to T^2$ have a fixed point? What about continuous maps $S^2 \to S^2$? (The question of whether a given continuous map of a surface (or any simplicial complex) to itself has a fixed point is treated by the Lefschetz Fixed Point Theorem; see [MU3, p. 125] for example.)

3.8.4*. Our goal is to show that the 1-sphere $C = S^1 \times \{0\} \subset S^1 \times \mathbb{R}$ is not contained in any subset of $S^1 \times \mathbb{R}$ homeomorphic to int D^2; fill in the details of each step.

Step 1: Suppose to the contrary that C is contained in a subset of $S^1 \times \mathbb{R}$ that is homeomorphic to int D^2. Show that C is the boundary of a disk B in $S^1 \times \mathbb{R}$. Let $f: D^2 \to B$ be a homeomorphism

Step 2: Show that there is a homeomorphism $H: S^1 \times \mathbb{R} \to \mathbb{R}^2 - O_2$ such that $H(C) = C$.

Step 3: Show that the homeomorphism $H \circ f|C: C \to C$ can be extended to a homeomorphism $G: \mathbb{R}^2 - O_2 \to \mathbb{R}^2 - O_2$.

Step 4: Consider the map $G^{-1} \circ H \circ f$, and derive a contradiction.

3.8.5*. Show that the surfaces $\mathbb{R}^2$ and $S^1 \times \mathbb{R}$ are not homeomorphic.

3.8.6*. Let $B \subset \mathbb{R}^n$ be a disk. Show that the Brouwer Fixed Point Theorem holds as stated iff it holds with B replacing D^2.

Endnotes

Notes for Section 3.3

Although going from a simplicial map of simplicial complexes to a continuous map of the underlying spaces is simple (see Lemma 3.3.5), going in the other direction is much trickier. Given an arbitrary continuous map from one underlying space to another, it would be very unlikely that this map was induced by a simplicial map, since an arbitrary continuous map is unlikely to be affine linear on simplices. Continuous maps can, however, be approximated arbitrarily closely by simplicial maps on subdivisions of the original complexes, a result known as the Simplicial Approximation Theorem (see [MU3, §16]).

Notes for Section 3.4

(A) An alternative proof of Theorem 3.4.1, making use of the Jordan Curve Theorem (Corollary 2.2.5 (i)) rather than Invariance of Domain (Theorem 2.2.1), can be found in [MO, Chapter 4].

(B) A very efficient proof of Theorem 3.4.5 (i) is found in [TH]).

Notes for Section 3.5

There is a disagreement among various authors about whether Euler was the first to discover the Euler characteristic, or whether Descartes (who lived well before Euler) had been aware of the Euler characteristic. In [F-F] it is argued that Descartes had been aware of the Euler characteristic, whereas in [S-F, §4] it is argued otherwise; a good survey of this argument is in [FE].

Notes for Section 3.6

There are a number of different proofs of the classification of compact connected surfaces (Theorem 2.6.7). One standard method is given in [MS1]. Another proof is via Morse Theory, as in [HR, Chapter 9]. An efficient recent proof using some ideas from graph theory is found in [TH]. The first rigorous proof of the classification theorem is often said to be in [BR], though [MS1] raises some question in this matter.

Notes for Section 3.7

There are a number of alternative approaches to defining curvature in simplicial surfaces. The most well-known of these is given in [BA1], [BA2] and [BA3]. Other approaches with a geometric flavor are in [C-M-S], [YU] and [BL], while combinatorial approaches can be found in [GR2] and [MC] among others. All these approaches are equivalent to the angle defect when applied to simplicial surfaces, but some can also be used with arbitrary simplicial complexes in all dimensions.

Notes for Section 3.8

(A) The Brouwer Fixed Point Theorem has many proofs, some using algebraic topology (for example [MU3, §21] and [MS2, p. 74]), and others using advanced Calculus (for example [MI4]). One of the most geometric elementary approaches is via the Sperner Lemmas, as in [LY].

(B) Not only is the Brouwer Fixed Point Theorem an inherently interesting geometric result, but it is also of interest to a number of applications such as economics (see [DE] or [CS]).

CHAPTER IV

Curves in $\mathbb{R}^3$

4.1 Introduction

Though our main topic of concern is surfaces, prior to studying smooth surfaces we take a small detour through the study of smooth curves in $\mathbb{R}^3$ to develop some important tools. Our treatment of curves will be brief; more about curves, including such results such as the pretty Milnor–Fary Theorem, can be found in [M-P] or [DO1].

For the rest of the book we will be in the realm of differentiable functions. Section 4.2 reviews some basic facts concerning such functions, including the Inverse Function Theorem and some existence and uniqueness theorems for the solutions of ordinary differential equations, which play a foundational role for smooth surfaces.

4.2. Smooth Functions

We start with some assumptions about differentiable functions.

Definition. Let $U \subset \mathbb{R}^n$ be a set, and let $F: U \to \mathbb{R}^m$ be a map. We say F is **smooth** if

(1) the set U is open in $\mathbb{R}^n$;

(2) all partial derivatives of F of all orders exist and are continuous.

We can write F using coordinate functions as

$$F(x) = \begin{pmatrix} F_1(x) \\ \vdots \\ F_m(x) \end{pmatrix},$$

where $x = \begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}$ and $F_1, \ldots, F_m: U \to \mathbb{R}$ are smooth functions. The

Jacobian matrix of F is the matrix of partial derivatives

$$DF = \begin{pmatrix} \frac{\partial F_1}{\partial x_1} & \cdots & \frac{\partial F_1}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial F_m}{\partial x_1} & \cdots & \frac{\partial F_m}{\partial x_n} \end{pmatrix}.$$

The openness of the set U in the above definition will often be unstated, but will be assumed nonetheless. The following definition is the smooth analog of the notion of homeomorphism.

Definition. Let $U, V \subset \mathbb{R}^n$ be open sets. A function $f: U \to V$ is a **diffeomorphism** if it is bijective, and if both f and f^{-1} are smooth.

If $f: U \to V$ is a diffeomorphism then the Jacobian matrix Df is non-singular at each point in U (see Exercise 4.2.2).

We now turn to the Inverse Function Theorem and differential equations; the reader should feel free to skip this material until it is needed is subsequent sections. The Inverse Function Theorem addresses the question of whether a smooth function $f: U \to \mathbb{R}^n$ has a smooth inverse (that is, whether it is a diffeomorphism). The one-dimensional case is simple. Let $f: J \to \mathbb{R}$ be a smooth function for some open interval J. If $f'(x_0) \neq 0$ for point $x_0 \in J$, then the function is either strictly increasing or strictly decreasing near x_0, and it follows that near x_0 the function has an inverse. Of course, having $f'(x_0) \neq 0$ does not imply that the whole function f has an inverse, but only that the function restricted to some (possibly very small) open neighborhood of x_0 has an inverse. Since the graph of an inverse function is simply the reflection in the line $y = x$ of the original graph, we see that if $f'(x_0) \neq 0$ then the inverse function of f restricted to a neighborhood of $f(x_0)$ will also be smooth. The Inverse Function Theorem is the higher-dimensional analog of what we have just discussed. The condition $f'(x_0) \neq 0$ is replaced by the condition that the Jacobian matrix has non-zero determinant at the given point.

Theorem 4.2.1 (Inverse Function Theorem). *Let $U \subset \mathbb{R}^n$ be an open set and let $F: U \to \mathbb{R}^n$ be a smooth map. If $p \in U$ is a point such that* $\det DF(p) \neq 0$, *then there is an open set $W \subset U$ containing p such that $F(W)$ is open in $\mathbb{R}^n$ and F is a diffeomorphism from W onto $F(W)$.*

See [SK1, p. 34] and [BO, p. 42] for proofs, as well as other information concerning the Inverse Function Theorem. We will also need the following

result, the proof of which is lengthy and might be skipped. This theorem is a special case of a more general result known as the Rank Theorem (see [BO, p. 47]); another special case of the Rank Theorem is given in Exercise 4.2.1.

Theorem 4.2.2. *Let $U \subset \mathbb{R}^2$ be an open set and let $f: U \to \mathbb{R}^3$ be a smooth map. If $p \in U$ is a point such that the matrix $Df(p)$ has rank 2, then there are open subsets $W \subset U$ and $V \subset \mathbb{R}^3$ containing p and $f(p)$ respectively and a smooth map $G: V \to \mathbb{R}^3$ such that $G(V)$ is open in $\mathbb{R}^3$, that G is a diffeomorphism from V onto $G(V)$, that $f(W) \subset V$ and that*

$$G \circ f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}$$

for all $\binom{x}{y} \in W$.

Proof of Theorem 4.2.2. Let $\overline{U} \subset \mathbb{R}^2$ be defined by $\overline{U} = \{x - p \mid x \in U\}$. We define a function $\overline{f}: \overline{U} \to \mathbb{R}^3$ by $\overline{f}(v) = f(v + p) - f(p)$ for all $v \in \overline{U}$. Observe that $\overline{U}$ is open in $\mathbb{R}^2$, that $\overline{f}$ is smooth, that $D\overline{f}(v) = Df(v + p)$, that $O_2 \in \overline{U}$, that $\overline{f}(O_2) = O_3$ and that $D\overline{f}(O_2)$ has rank 2. If the function $\overline{f}$ is given in coordinates by

$$\overline{f}(\bar{u}) = \begin{pmatrix} \overline{f}_1(\bar{u}) \\ \overline{f}_2(\bar{u}) \\ \overline{f}_3(\bar{u}) \end{pmatrix},$$

where $\bar{u} = \binom{u_1}{u_2}$, then the Jacobian matrix of $\overline{f}$ is

$$D\overline{f} = \begin{pmatrix} \frac{\partial \overline{f}_1}{\partial u_1} & \frac{\partial \overline{f}_1}{\partial u_2} \\ \frac{\partial \overline{f}_2}{\partial u_1} & \frac{\partial \overline{f}_2}{\partial u_2} \\ \frac{\partial \overline{f}_3}{\partial u_1} & \frac{\partial \overline{f}_3}{\partial u_2} \end{pmatrix}.$$

Since the rank of the matrix $D\overline{f}(O_2)$ is 2, it follows from standard results in linear algebra that $D\overline{f}(O_2)$ has a 2×2 submatrix with non-zero determinant. By relabeling the coordinates of $\mathbb{R}^3$ if necessary, we may assume without loss of generality that the top two rows of $D\overline{f}(O_2)$ have non-zero determinant, that is

$$\det \begin{pmatrix} \frac{\partial \overline{f}_1}{\partial u_1}|_{\bar{u}=O_2} & \frac{\partial \overline{f}_1}{\partial u_2}|_{\bar{u}=O_2} \\ \frac{\partial \overline{f}_2}{\partial u_1}|_{\bar{u}=O_2} & \frac{\partial \overline{f}_2}{\partial u_2}|_{\bar{u}=O_2} \end{pmatrix} \neq 0. \tag{4.2.1}$$

We now define a function $H: \overline{U} \times \mathbb{R} \to \mathbb{R}^3$ by

$$H\left(\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}\right) = \overline{f}(\bar{u}) + \begin{pmatrix} 0 \\ 0 \\ u_3 \end{pmatrix} = \begin{pmatrix} \overline{f}_1(\bar{u}) \\ \overline{f}_2(\bar{u}) \\ \overline{f}_3(\bar{u}) + u_3 \end{pmatrix},$$

where $\bar{u}$ is as above. The domain of H is an open subset of $\mathbb{R}^3$, and since the coordinate functions of $\overline{f}$ are smooth, so is the map H. Further, note that $H(O_3) = O_3$. The Jacobian matrix of H is

$$DH = \begin{pmatrix} \frac{\partial \overline{f}_1}{\partial u_1} & \frac{\partial \overline{f}_1}{\partial u_2} & 0 \\ \frac{\partial \overline{f}_2}{\partial u_1} & \frac{\partial \overline{f}_2}{\partial u_2} & 0 \\ \frac{\partial \overline{f}_3}{\partial u_1} & \frac{\partial \overline{f}_3}{\partial u_2} & 1 \end{pmatrix}.$$

It follows from Equation 4.2.1 that $\det DH(O_3) \neq 0$. Applying the Inverse Function Theorem to H at the point O_3, we conclude that there is an open set $T \subset \overline{U} \times \mathbb{R}$ containing O_3 such that $H(T)$ is open in $\mathbb{R}^3$ and H is a diffeomorphism from T onto $H(T)$. Observe that $O_3 \in H(T)$.

We now define the sets V and W and the map G as follows. Let $V = \{x + f(p) \mid x \in H(T)\}$. Note that V is open in $\mathbb{R}^3$ and that $f(p) \in V$. Next, define

$$W = f^{-1}(V) \cap \{x + p \mid x \in H(T) \cap \mathbb{R}^2\}.$$

Observe that W is open in $\mathbb{R}^2$, that $p \in W$ and that $f(W) \subset V$. We now define $G: V \to \mathbb{R}^3$ by

$$G(v) = (H|T)^{-1}(v - f(p)) + \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix}$$

for all $v \in V$, where $p = \binom{p_1}{p_2}$. Since $H|T$ is a diffeomorphism so is $(H|T)^{-1}$, and it follows that $G(V)$ is an open subset of $\mathbb{R}^3$ and that G is a diffeomorphism from V onto $G(V)$.

From the definitions of $\overline{f}$ and H it follows that $f(v) = \overline{f}(v-p) + f(p)$ for all $v \in U$, and that $\overline{f}\left(\binom{u_1}{u_2}\right) = H\left(\begin{pmatrix} u_1 \\ u_2 \\ 0 \end{pmatrix}\right)$ for all $\binom{u_1}{u_2} \in \overline{U}$. For each $\binom{x}{y} \in W$,

we now compute

$$G \circ f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) = H^{-1}\left(f\left(\begin{pmatrix} x \\ y \end{pmatrix}\right) - f(p)\right) + \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix}$$

$$= H^{-1}\left(\overline{f}\left(\begin{pmatrix} x \\ y \end{pmatrix} - p\right) + f(p) - f(p)\right) + \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix}$$

$$= H^{-1}\left(\overline{f}\left(\begin{pmatrix} x - p_1 \\ y - p_2 \end{pmatrix}\right)\right) + \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix}$$

$$= H^{-1}\left(H\left(\begin{pmatrix} x - p_1 \\ y - p_2 \\ 0 \end{pmatrix}\right)\right) + \begin{pmatrix} p_1 \\ p_2 \\ 0 \end{pmatrix} = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}. \quad \square$$

The following result can be deduced from the above theorem.

Corollary 4.2.3. *Let $U \subset \mathbb{R}^2$ be an open set and let $f: U \to \mathbb{R}^3$ be a smooth map. If $p \in U$ is a point such that the matrix $Df(p)$ has rank 2, then there is an open set $W \subset U$ containing p such that $f|W$ is injective, and $Df(q)$ has rank 2 for all $q \in W$.*

Proof. Exercise 4.2.5. $\square$

The other foundational material we need is the following three existence and uniqueness theorems for the solutions of ordinary differential equations. The first of these results is the standard such existence and uniqueness theorem; the second is a stronger version, which shows how solutions of ordinary differential equations depend upon the initial conditions; the third is a theorem concerning the special case of linear differential equations, where we have a slightly better result than for arbitrary differential equations. See [LA2, Chapter XVIII] for proofs of all three theorems, or [HZ] for the first two.

Theorem 4.2.4 (Existence and uniqueness of solutions of ordinary differential equations). *Let $U \subset \mathbb{R}^n$ be an open set, let $F: U \to \mathbb{R}^n$ be a smooth map and let $t_0 \in \mathbb{R}$ and $v_0 \in U$ be points. Then there is a number $\epsilon > 0$ and a smooth map $c: (t_0 - \epsilon, t_0 + \epsilon) \to U$ such that*

$$c'(t) = F(c(t)) \quad \text{and} \quad c(t_0) = v_0 \tag{4.2.2}$$

for all $t \in (t_0 - \epsilon, t_0 + \epsilon)$*; if* $\widetilde{c}\colon (t_0 - \delta, t_0 + \delta) \to U$ *is any other map that satisfies Equation 4.2.2 for some* $\delta > 0$*, then* $\widetilde{c}(t) = c(t)$ *for all* t *in the intersection of the domains of the two maps.*

Theorem 4.2.5. *Let* $U \subset \mathbb{R}^n$ *be an open subset, let* $F\colon U \to \mathbb{R}^n$ *be a smooth map and let* $t_0 \in \mathbb{R}$ *and* $v_0 \in U$ *be points. Then there is a number* $\epsilon > 0$*, an open subset* $V \subset \mathbb{R}^n$ *containing* v_0 *and a smooth map* $C\colon (t_0 - \epsilon, t_0 + \epsilon) \times V \to U$ *such that*

$$C'(t, v) = F(C(t, v)) \quad \textit{and} \quad C(t_0, v) = v$$

for all $(t, v) \in (t_0 - \epsilon, t_0 + \epsilon) \times V$.

Let $M_{nn}(\mathbb{R})$ denote the set of real $n \times n$ matrices.

Theorem 4.2.6. *Let* (a, b) *be an open interval, let* $A\colon (a, b) \to M_{nn}(\mathbb{R})$ *be a smooth map and let* $t_0 \in (a, b)$ *and* $v_0 \in \mathbb{R}^n$ *be points. Then there is a unique smooth function* $c\colon (a, b) \to \mathbb{R}^n$ *such that*

$$c'(t) = A(t)c(t) \quad \textit{and} \quad c(t_0) = v_0$$

for all $t \in (a, b)$.

Exercises

4.2.1*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth map. Suppose that $p \in (a, b)$ is a point such that the matrix $c'(p) \neq 0$. Show that there is a number $\epsilon > 0$, an open subset $V \subset \mathbb{R}^3$ containing $c(p)$ and a smooth map $G\colon V \to \mathbb{R}^3$ such that $G(V)$ is open in $\mathbb{R}^3$, that G is a diffeomorphism from V onto $G(V)$, that $c((p - \epsilon, p + \epsilon)) \subset V$ and that

$$G \circ c(t) = \begin{pmatrix} t \\ 0 \\ 0 \end{pmatrix}$$

for all $t \in (p - \epsilon, p + \epsilon)$.

4.2.2*. Let $U, V \subset \mathbb{R}^n$ be open sets, and suppose $f\colon U \to V$ is a diffeomorphism. Show that $Df(p)$ is a non-singular matrix for all $p \in U$.

4.2.3*. Let $U, V \subset \mathbb{R}^n$ be open sets, and suppose $f\colon U \to V$ is a smooth bijective map. Show that if $Df(p)$ is a non-singular matrix for all $p \in U$, then f is a diffeomorphism.

4.2.4*. Let $c: (a, b) \to \mathbb{R}^2$ be a smooth function. Suppose that the tangent vectors to c are never the zero vector and are never parallel to the y-axis. Show that the image of c is the graph of a function of the form $y = f(x)$ for some smooth function $f: (p, q) \to \mathbf{R}$.

4.2.5*. Prove Corollary 4.2.3. State and prove the analog of this corollary for smooth functions $c: (a, b) \to \mathbb{R}^3$.

4.2.6. Give an example of a function $G: \mathbb{R}^2 \to \mathbb{R}^2$ that has non-zero Jacobian matrix at all points, and yet is not injective in every neighborhood of any point.

4.3 Curves in $\mathbb{R}^3$

The concept of a curve in $\mathbb{R}^3$ is intuitively quite simple; imagine a twisted piece of string, as in Figure 4.3.1. A smooth curve is, pictorially, one that bends nicely and has no kinks or corners. To deal with smooth curves rigorously, however, we need to think of a curve slightly differently; rather than thinking of a curve as an object that simply sits in $\mathbb{R}^3$, we should view it as the path of a moving object. Every point on the curve corresponds to the location of the moving object at a particular time. We could imagine traversing the same path at a variety of different speeds, not to mention changing direction; we will deal with this issue shortly. Finally, rather than thinking about the points on the curve as simply points in $\mathbb{R}^3$, it is technically more useful to think of points on a curve as the endpoints of vectors starting at the origin. Putting these observations together we arrive at the following definition.

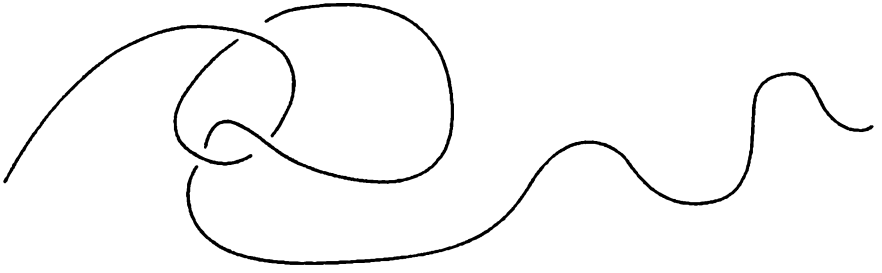


Figure 4.3.1

Definition. A **smooth curve** (or simply **curve**) in $\mathbb{R}^3$ is a smooth function $c\colon (a, b) \to \mathbb{R}^3$, where (a, b) is an interval (possibly infinite) in $\mathbb{R}$. For each $t \in (a, b)$ the **velocity vector** of the curve at t is the vector $c'(t)$, and the **speed** at t is the real number $\|c'(t)\|$. A curve is **unit speed** if $\|c'(t)\| = 1$ for all $t \in (a, b)$.

Example 4.3.1. Consider the curve $c\colon (0, 1) \to \mathbb{R}^3$ given by

$$c(t) = \begin{pmatrix} \cos t \\ \sin t \\ t^2 \end{pmatrix}.$$

Then

$$c'(t) = \begin{pmatrix} -\sin t \\ \cos t \\ 2t \end{pmatrix} \quad \text{and} \quad \|c'(t)\| = \sqrt{1 + 4t^2},$$

so that c is not unit speed. ◊

The above definition is actually not quite enough to insure that the image of the curve will look geometrically "smooth." Imagine a bug flying around in $\mathbb{R}^3$, and assume that the flight is smooth (in the sense of infinite differentiability). While maintaining smooth motion the bug could slow down till it stops altogether, turn 90° in some direction, and then take off again, gradually accelerating from its initial speed of zero. The path taken by the bug after executing this maneuver has a corner in it, even though its flight could be described as a smooth curve as we have defined it. The following definition eliminates the problem, and describes the class of curves with which we will be working.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth curve. The curve c is **regular** if $c'(t) \neq 0$ for all $t \in (a, b)$, that is, if $\|c'(t)\| \neq 0$ for all $t \in (a, b)$.

Example 4.3.2. The curve in Example 4.3.1 is regular, since $\|c'(t)\|$ is never zero. ◊

The following definition definition is the formal relation of "different ways of traversing a string."

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ and $\widetilde{c}\colon (d, e) \to \mathbb{R}^3$ be smooth curves. We say that $\widetilde{c}$ is a **reparametrization** of c if there is a diffeomorphism $h\colon (d, e) \to (a, b)$ such that $\widetilde{c} = c \circ h$.

Observe that a curve and any reparametrization of it have the same image set in $\mathbb{R}^3$.

Example 4.3.3. Let $c\colon (1, 5) \to \mathbb{R}^3$ and $\tilde{c}\colon (0, 2) \to \mathbb{R}^3$ be defined by

$$c(t) = \begin{pmatrix} t^2 + 3 \\ t - 7 \\ \sin t \end{pmatrix}, \quad \text{and} \quad \tilde{c}(t) = \begin{pmatrix} 4t^2 + 4t + 4 \\ 2t - 6 \\ \sin(2t + 1) \end{pmatrix}.$$

Then $\tilde{c} = c \circ h$, where $h\colon (0, 2) \to (1, 5)$ is given by $h(t) = 2t + 1$. It is straightforward to verify that h is smooth, bijective, and has a smooth inverse, so that h is a diffeomorphism. $\Diamond$

The following lemma shows that any regular curve can be reparametrized in a particularly simple way. The proof of the lemma might at first appear to be pulled out of thin air, though there is actually an intuitive idea behind it, namely that a curve will be unit speed if the parameter corresponds to arc-length along the curve.

Proposition 4.3.4. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve.*

(i) *There is a reparametrization of c that is a unit speed curve.*

(ii) *Let $c \circ h_1$ and $c \circ h_2$ be unit speed reparametrizations of c, for appropriate functions $h_1\colon (d_1, e_1) \to (a, b)$ and $h_2\colon (d_2, e_2) \to (a, b)$. Then the function $h_2^{-1} \circ h_1\colon (d_1, e_1) \to (d_2, e_2)$ has the form $h_2^{-1} \circ h_1(s) = \pm s + k$ for some constant k.*

Proof. (i) Pick some point $t_0 \in (a, b)$. Define a function $q\colon (a, b) \to \mathbb{R}$ by

$$q(t) = \int_{t_0}^{t} \|c'(u)\|\, du.$$

By the Fundamental Theorem of Calculus the function q is smooth and $q'(t) = \|c'(t)\| > 0$, the inequality following from the regularity of c. Hence q is a strictly increasing function, and q is therefore a bijective map from (a, b) onto its image. The image of q will be the interval (d, e), where

$$d = \int_{t_0}^{a} \|c'(u)\|\, du\,, \qquad e = \int_{t_0}^{b} \|c'(u)\|\, du.$$

Let $h\colon (d, e) \to (a, b)$ be the inverse function of q. Since the derivative of q is never zero it follows from a standard theorem in Calculus that h is also smooth, and $h'(s) = 1/q'(s)$ for all $s \in (d, e)$. Let $\tilde{c}\colon (d, e) \to \mathbb{R}^3$ be defined by $\tilde{c} = c \circ h$. By definition $\tilde{c}$ is a reparametrization of c. Further, for each

$s \in (d, e)$ we have

$$\tilde{c}'(s) = c'(h(s))\, h'(s) = c'(h(s))\, \frac{1}{q'(h(s))} = c'(h(s))\, \frac{1}{\|c'(h(s))\|}.$$

Hence $\|\tilde{c}'(s)\| = 1$ for all $s \in (d, e)$.

(ii) For each $i = 1, 2$ we have

$$1 = \|(c \circ h_i)'(s)\| = \|c'(h_i(s))\|\, |h_i'(s)|$$

for $s \in (d_i, e_i)$. Hence

$$h_1'(s) = \pm \frac{1}{\|c'(h_1(s))\|} = \pm \frac{1}{\|c'(h_2(h_2^{-1} \circ h_1(s)))\|} = \pm h_2'(h_2^{-1} \circ h_1(s))$$

for each $s \in (d_1, e_2)$, and thus

$$(h_2^{-1} \circ h_1)'(s) = (h_2^{-1})'(h_1(s))\, h_1'(s) = \frac{h_1'(s)}{h_2'(h_2^{-1} \circ h_1(s))} = \pm 1.$$

Since $h_2^{-1} \circ h_1$ is smooth, then it is either constantly 1 or constantly -1. The desired result now follows. □

Though in theory the proof of part (i) of the above lemma gives a procedure for finding unit speed reparametrizations, in practice doing so is not always possible since it involves computing integrals and inverses of functions.

Example 4.3.5. The unit right circular helix is the curve $c: (-\infty, \infty) \to \mathbb{R}^3$ given by

$$c(t) = \begin{pmatrix} \cos t \\ \sin t \\ t \end{pmatrix}.$$

See Figure 4.3.2. It is not hard to see that $\|c'(t)\| = \sqrt{2}$ for all t. Choosing $t_0 = 0$, we have

$$q(t) = \int_0^t \sqrt{2}\, du = \sqrt{2}t,$$

and hence

$$h(t) = \frac{t}{\sqrt{2}}.$$

Thus our unit speed reparametrization is

$$\tilde{c}(t) = (c \circ h)(t) = \begin{pmatrix} \cos \frac{t}{\sqrt{2}} \\ \sin \frac{t}{\sqrt{2}} \\ \frac{t}{\sqrt{2}} \end{pmatrix}. \quad \Diamond$$

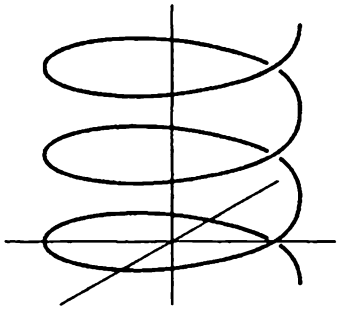


Figure 4.3.2

Suppose two smooth curves $c\colon (a, b) \to \mathbb{R}^3$ and $\tilde{c}\colon (d, e) \to \mathbb{R}^3$ have the same image. Can we realize $\tilde{c}$ as a reparametrization of c? Although the situation could be tricky if the curves were not injective, we do have the following lemma, which will suffice for our purposes.

Lemma 4.3.6. *Let $c\colon (a, b) \to \mathbb{R}^3$ and $\tilde{c}\colon (d, e) \to \mathbb{R}^3$ be injective regular curves with the same image. Then $\tilde{c}$ is a reparametrization of c.*

Proof. Since c is injective it must be a bijection onto its image, and hence there is a function $c^{-1}\colon c((a, b)) \to (a, b)$. Define the function $h\colon (d, e) \to (a, b)$ by letting $h = c^{-1} \circ \tilde{c}$. The function h is a bijection, and by Exercise 4.3.11 it is smooth. Doing this whole procedure in the other direction also shows that h^{-1} is smooth. Evidently $\tilde{c} = c \circ h$, and thus $\tilde{c}$ is a reparametrization of c. □

We now calculate the length of the image of a curve, which for convenience we will refer to as "length of a curve." It ought to be the case that the length depends only upon the image of the curve, and not upon the particular parametrization used. However, it is easier to make use of parametrizations in our definition of the length of a curve, and then to show that the quantity defined in fact does not depend upon the parametrization used. The idea is to approximate the image of the curve with a finite number of small straight line segments, add up the lengths of the segments to get an approximate length of the curve, and take the limit of these sums as smaller and smaller segments are used. In the limit the sum becomes an integral, and the term $\|c'(t)\|\, dt$ in the definition below comes from the lengths of the line segments. Such argumentation does not "prove" that our formula for the length of a curve equals our intuitive notion of what is meant by the length of such curves; it really only pushes back where the leap of faith is made.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth curve. The **length** of c is defined to be the number Length(c) given by

$$\text{Length}(c) = \int_a^b \|c'(t)\|\, dt, \tag{4.3.1}$$

provided the integral exists.

Example 4.3.7. Let $c\colon (1, 2) \to \mathbb{R}^3$ be given by

$$c(t) = \begin{pmatrix} \frac{t^2}{2} \\ 1 \\ \frac{t^3}{3} \end{pmatrix}.$$

It can be computed that $\|c'(t)\| = t\sqrt{1 + t^2}$ (observe that $t > 0$). The length of c is thus

$$\text{Length}(c) = \int_1^2 t\sqrt{1 + t^2}\, dt = \frac{5^{3/2} - 2^{3/2}}{3}. \quad \Diamond$$

The following lemma says that our definition of the length of curves behaves as we hoped it would with respect to parametrizations.

Lemma 4.3.8. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth curve. If $\tilde{c}\colon (d, e) \to \mathbb{R}^3$ is a reparametrization of c, then Length($\tilde{c}$) = Length(c).*

Proof. Exercise 4.3.6. □

Exercises

4.3.1. Which of the following curves are regular?

(i) $c\colon (-\infty, \infty) \to \mathbb{R}^3$ given by $c(t) = \begin{pmatrix} 1 \\ t^3 \\ t^4 \end{pmatrix}$;

(ii) $d\colon (0, \infty) \to \mathbb{R}^3$ given by $d(t) = \begin{pmatrix} t \ln t - t \\ 5 \\ 2t \ln t - 2t \end{pmatrix}$.

4.3.2. The curve $c\colon (-\infty, \infty) \to \mathbb{R}^3$ defined by

$$c(t) = \begin{pmatrix} Be^{kt} \cos t \\ Be^{kt} \sin t \\ 0 \end{pmatrix}$$

is called the **logarithmic spiral**; this curve appears to appear in nature, describing, for example, the shape of a nautilus shell. Show that this curve has the property that the angle between the vector $c(t)$ and the vector $c'(t)$ is a constant. (This property in fact characterizes the logarithmic spiral.)

4.3.3*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth curve. Show that there is a diffeomorphism $h\colon (d, e) \to (a, b)$ for some interval (d, e) in $\mathbb{R}$ such that $\tilde{c} = c \circ h$ is unit speed and $h'(t) > 0$ for all $t \in (d, e)$.

4.3.4. Find unit speed reparametrizations of the following curves.

(i) $c\colon (0, \infty) \to \mathbb{R}^3$ given by $c(t) = \frac{1}{2}\begin{pmatrix} t \\ 1/t \\ \sqrt{2}\ln t \end{pmatrix}$.

(ii) The logarithmic spiral in Exercise 4.3.2.

4.3.5. The logarithmic spiral can be broken into segments from $t = 2n\pi$ to $t = 2(n + 1)\pi$ for each $n \in \mathbb{Z}$. Find the length of such a segment. What is the ratio of the length of one such segment to the length of the previous segment? Intuitively, why would a nautilus shell would have this property?

4.3.6*. Prove Lemma 4.3.8.

4.3.7. Let $x = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}$ and $y = \begin{pmatrix} y_1 \\ y_2 \\ y_3 \end{pmatrix}$ be points in $\mathbb{R}^3$. Choose a parametrization of the line segment from x to y and calculate the length of this curve. (There are many such parametrizations, so chose one you think will be most convenient to work with.)

4.3.8. Show that the circumference of a circle of radius r is $2\pi r$.

4.3.9. Let $y = f(x)$ be a function $f\colon (a, b) \to \mathbb{R}$. The graph of this function can be parametrized by the curve $c\colon (a, b) \to \mathbb{R}^3$ given by

$$c(t) = \begin{pmatrix} t \\ f(t) \\ 0 \end{pmatrix}.$$

Find a formula for the length of this curve. How does it compare to the standard formula for arc-length found in most Calculus texts?

4.3.10*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve. Suppose that $c|[p, q]$ is injective for some closed interval $[p, q] \subset (a, b)$. Show that there exists a number $\epsilon > 0$ such that $c|(p-\epsilon, q+\epsilon)$ is a homeomorphism from $(p-\epsilon, q+\epsilon)$ to $c((p - \epsilon, q + \epsilon))$.

4.3.11*. Prove that $h = c^{-1} \circ \tilde{c}$ in the proof of Lemma 4.3.6 is smooth.

4.4 Tangent, Normal and Binormal Vectors

The tangent vector to a curve is the vector that best approximates the curve at the point of tangency. See Figure 4.4.1. Given a smooth curve $c\colon (a, b) \to \mathbb{R}^3$, the tangent vector at point $t \in (a, b)$ turns out to be nothing other than the velocity vector $c'(t)$ defined previously. However, whereas we would like to think of a tangent vector as "starting" at the point of tangency on the curve, in our present situation the tangent vector is translated so that it starts at the origin. The use of the following definition will become apparent shortly.

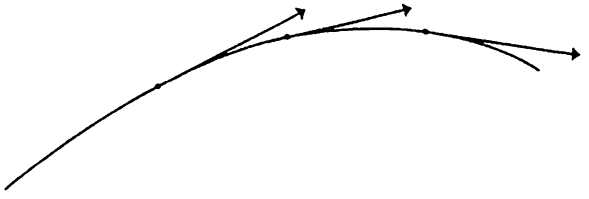


Figure 4.4.1

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth curve. For each $t \in (a, b)$ such that $\|c'(t)\| \neq 0$ the **unit tangent vector** to the curve at t is the vector

$$T(t) = \frac{c'(t)}{\|c'(t)\|}.$$

If a curve is regular then the unit tangent vector is defined at all points. Also, if a curve is unit speed then the unit tangent vector is just the velocity vector.

Example 4.4.1. Let $c\colon (-\infty, \infty) \to \mathbb{R}^3$ be given by

$$c(t) = \begin{pmatrix} 1 \\ t \\ t^2/2 \end{pmatrix}.$$

Then

$$c'(t) = \begin{pmatrix} 0 \\ 1 \\ t \end{pmatrix}, \quad \|c'(t)\| = \sqrt{1 + t^2} \quad \text{and} \quad T(t) = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{1+t^2}} \\ \frac{t}{\sqrt{1+t^2}} \end{pmatrix}. \quad \Diamond$$

Consider a regular curve $c\colon (a, b) \to \mathbb{R}^3$. Although the image of the curve need not lie in a single plane, at any point $c(t)$ on the curve there is a plane that is the closest thing to a plane containing the curve. See Figure 4.4.2. The unit tangent vector to the curve will be contained in this plane; we need to find another unit vector contained in the plane and linearly independent from the unit tangent vector. To find this other unit vector, we start by noting that the unit tangent vector function $T\colon (a, b) \to \mathbb{R}^3$ is also smooth. Observing that $\|T(t)\| = 1$ for all t, we have

$$\langle T(t), T(t)\rangle = 1,$$

where $\langle\ ,\ \rangle$ is the standard inner product in $\mathbb{R}^3$. Taking the derivative of both sides, and using the standard properties of derivatives and inner products (see Lemma 5.6.1), we deduce that

$$2\langle T'(t), T(t)\rangle = 0.$$

Thus $T'(t)$ is orthogonal to $T(t)$ for all t. If $T'(t) = 0$ then this whole business does not do us much good, so we will generally assume that $T'(t) \neq 0$. (This last assumption rules out the usual parametrization of a straight line, for example.) We can now define a new vector that is always orthogonal to $T(t)$.

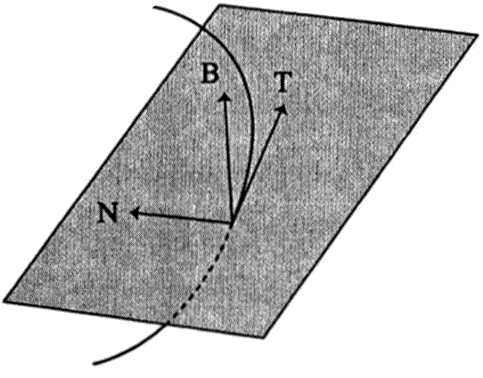


Figure 4.4.2

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve. For each $t \in (a, b)$ such that $\|T'(t)\| \neq 0$, the **unit normal vector** to the curve at t is the vector

$$N(t) = \frac{T'(t)}{\|T'(t)\|}.$$

Whenever the vectors $T(t)$ and $N(t)$ are both defined, we consider the plane that they span to be the plane that best fits the curve (referred to as the "osculating plane"), just as the tangent line is the line that best fits the curve.

Example 4.4.2. We continue Example 4.4.1, computing

$$T'(t) = \begin{pmatrix} 0 \\ \frac{-t}{(1+t^2)^{3/2}} \\ \frac{1}{(1+t^2)^{3/2}} \end{pmatrix}, \quad \|T'(t)\| = \frac{1}{1+t^2} \quad \text{and} \quad N(t) = \begin{pmatrix} 0 \\ \frac{-t}{\sqrt{1+t^2}} \\ \frac{1}{\sqrt{1+t^2}} \end{pmatrix}.$$

◊

It is often inconvenient to verify whether $\|T'(t)\| \neq 0$, since $T(t)$ is often a fraction with a complicated denominator. The following lemma makes life a bit easier.

Lemma 4.4.3. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve. For each $t \in (a, b)$ the following are equivalent:*

(1) $\|T'(t)\| \neq 0$;
(2) *The vectors $c'(t)$ and $c''(t)$ are linearly independent;*
(3) $c'(t) \times c''(t) \neq 0$.

Proof. Exercise 4.4.3. □

For convenience we adopt the following terminology.

Definition. A regular curve $c\colon (a, b) \to \mathbb{R}^3$ is **strongly regular** if any of the three equivalent conditions in Lemma 4.4.3 holds for all $t \in (a, b)$.

Example 4.4.4. For the curve in Example 4.4.1 we compute

$$c''(t) = \begin{pmatrix} 0 \\ 0 \\ 2 \end{pmatrix}, \quad \text{and} \quad c'(t) \times c''(t) = \begin{pmatrix} 2 \\ 0 \\ 0 \end{pmatrix},$$

and thus the curve is strongly regular. ◊

For every t such that $\|T'(t)\| \neq 0$, we have now defined two orthogonal unit vectors $T(t)$ and $N(t)$. Given that our curve is in $\mathbb{R}^3$, and that three orthonormal vectors in $\mathbb{R}^3$ form a basis, we complete the picture by defining for each appropriate t a third unit vector orthogonal to both $T(t)$ and $N(t)$.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve. For each $t \in (a, b)$ such that $\|T'(t)\| \neq 0$, the **unit binormal vector** to the curve at t is the vector

$$B(t) = T(t) \times N(t).$$

A few observations about the above definition. First, except for the sign, there is really no choice in the definition of $B(t)$ if we want the set of vectors $\{T(t), N(t), B(t)\}$ to form an orthonormal set. Second, the definition of $B(t)$ makes crucial use of the fact that our curve is in $\mathbb{R}^3$, since the cross product is only defined in three dimensions (in higher dimensions, by contrast, there are many possible choices for a unit vector orthogonal to any two given vectors). The vectors $\{T(t), N(t), B(t)\}$ are defined for all t in the domain of a strongly regular curve. These three vectors are often called the **Frenet frame** of the curve.

Example 4.4.5. Continuing Example 4.4.1, we compute

$$B(t) = \begin{pmatrix} 0 \\ \frac{1}{\sqrt{1+t^2}} \\ \frac{t}{\sqrt{1+t^2}} \end{pmatrix} \times \begin{pmatrix} 0 \\ \frac{-t}{\sqrt{1+t^2}} \\ \frac{1}{\sqrt{1+t^2}} \end{pmatrix} = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}.$$

The significance of the fact that $B(t)$ turns out to be a constant in this example will be clarified by Exercise 4.4.4. $\Diamond$

It would be nice to have a simpler way to compute the Frenet frame of a curve, since the often complicated denominator in the expression for $T(t)$ can make finding the necessary derivatives quite messy. An alternate method will be given in Lemma 4.5.7; although the statement of the relevant parts of this lemma could be given now, some additional concepts and results are needed prior to the proof of the lemma.

Exercises

4.4.1. For each of the following curves, determine whether the curve is strongly regular, and, if so, find T, N and B.

(i) The circle in the x-y plane of radius 2 centered at the origin which we parametrize by the curve $g\colon (-\infty, \infty) \to \mathbb{R}^3$ given by $g(t) = \begin{pmatrix} 2\cos(t/2) \\ 2\sin(t/2) \\ 0 \end{pmatrix}$;

(ii) $c\colon (-\infty, \infty) \to \mathbb{R}^3$ given by $c(t) = \begin{pmatrix} 1 \\ t \\ 3t \end{pmatrix}$;

(iii) $d\colon (0, \infty) \to \mathbb{R}^3$ given by $d(t) = \begin{pmatrix} \ln t \\ t \\ 0 \end{pmatrix}$.

4.4.2*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve lying entirely in a plane. Show that whenever $T(t)$ and $N(t)$ are both defined they are parallel to the plane containing the curve.

4.4.3*. Prove Lemma 4.4.3.

4.4.4*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular curve whose image lies entirely in a plane. Show that $B(t)$ is a constant.

4.5 Curvature and Torsion

If we look at the image of a curve in $\mathbb{R}^3$, as in Figure 4.5.1, we see that there are points on the curve at which the curve is bending more (point A) and others at which the curve is bending less (point B). We wish to quantify this bending. As in Section 3.9, we begin with a discussion of the expected properties of curvature before stating our definition. Curvature ought to be an assignment of a number to each point of the curve to tell us how much the curve is bending at that point. Although curvature should only depend upon the image of the curve, and not upon any particular choice of parametrization, it will be much more convenient to assign the curvature to each value t in the domain of the curve $c\colon (a, b) \to \mathbb{R}^3$. Thus curvature will be a function of the form $\kappa\colon (a, b) \to \mathbb{R}$. The function κ should be smooth, and it should have the property that whenever the image of the curve is a straight line in a neighborhood of a point $c(t)$, then $\kappa(t)$ should be zero.

Consider the velocity vector to a curve $c\colon (a, b) \to \mathbb{R}^3$. The faster the velocity vector changes direction as we move along the curve, the more the curve appears to be bending. Thus the measure of curvature ought to be something like the derivative of the velocity vector, or, better, the length of the derivative of the velocity vector (since curvature ought to be a scalar). The problem with this proposed definition is that it does depend upon the parametrization of the curve, since if we traverse a curve faster the derivative of the velocity vector will be larger. To overcome this problem, we first look at unit speed curves,

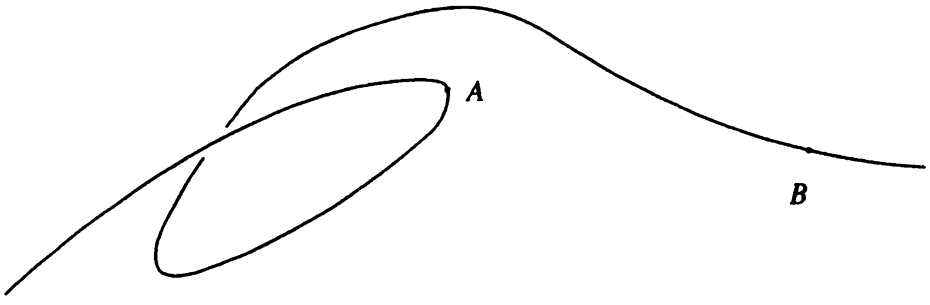


Figure 4.5.1

which gets rid of the problem of traversing a curve at differing speeds. For the following definition recall that for a unit speed curve $c\colon (a, b) \to \mathbb{R}^3$ the unit tangent vector $T(t)$ equals the velocity vector $c'(t)$.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a unit speed curve. For each $t \in (a, b)$ the **curvature** of the curve at t is the number

$$\kappa(t) = \|T'(t)\| = \|c''(t)\|.$$

Observe that curvature is a smooth function $\kappa\colon (a, b) \to \mathbb{R}$, and that $\kappa(t) \geq 0$.

Example 4.5.1. (1) Any straight line in $\mathbb{R}^3$ can be parametrized by $c\colon (-\infty, \infty) \to \mathbb{R}^3$ of the form

$$c(t) = \begin{pmatrix} a_1 t + b_1 \\ a_2 t + b_2 \\ a_3 t + b_3 \end{pmatrix};$$

the added condition that $a_1^2 + a_2^2 + a_3^2 = 1$ insures that this curve is unit speed. Clearly $c''(t)$ is the zero vector for all t, so $\kappa(t) = 0$ for all t. Hence we see that condition (2) for curvature suggested above is satisfied for this parametrization of a straight line.

(2) The circle of radius 2 in the x-y plane with center at the origin can be parametrized by the curve $d\colon (-\infty, \infty) \to \mathbb{R}^3$ given by

$$d(t) = \begin{pmatrix} 2\cos\frac{t}{2} \\ 2\sin\frac{t}{2} \\ 0 \end{pmatrix}.$$

It is seen that this curve is unit speed. We then compute

$$d'(t) = \begin{pmatrix} -\sin\frac{t}{2} \\ \cos\frac{t}{2} \\ 0 \end{pmatrix} \quad \text{and} \quad d''(t) = \frac{1}{2}\begin{pmatrix} -\cos\frac{t}{2} \\ -\sin\frac{t}{2} \\ 0 \end{pmatrix}.$$

It follows that $\kappa(t) = \frac{1}{2}$ for all t. The symmetry of the circle makes it reasonable that the curvature to be constant. ◊

The curvature function need not be constant, as seen in Exercise 4.5.1 (iii).

For a non-unit speed curve, we use reparametrization to reduce the problem to the previous definition.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve. Let $\widetilde{c} = c \circ h$ be a unit speed reparametrization of c for some diffeomorphism $h\colon (d, e) \to (a, b)$, and let $\widetilde{\kappa}$ be the curvature function for $\widetilde{c}$. For each $t \in (a, b)$ the **curvature** of the curve c at t is the number $\kappa(t) = \widetilde{\kappa}(h^{-1}(t))$.

The following lemma shows that the choice of unit speed reparametrization in the above definition does not affect the computation of curvature.

Lemma 4.5.2. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a regular curve, and let $c \circ h_1$ and $c \circ h_2$ be unit speed reparametrizations of c, where $h_1\colon (d_1, e_1) \to (a, b)$ and $h_2\colon (d_2, e_2) \to (a, b)$ are diffeomorphisms. If $\kappa_1(t)$ and $\kappa_2(t)$ are the curvature functions for $c \circ h_1$ and $c \circ h_2$ respectively, then*

$$\kappa_1(h_1^{-1}(t)) = \kappa_2(h_2^{-1}(t))$$

for all $t \in (a, b)$.

Proof. It follows from Proposition 4.3.4 (ii) that $h_2^{-1} \circ h_1(s) = \pm s + k$ for some constant k. Thus $h_1(s) = h_2(\pm s + k)$, so $c \circ h_1(s) = c \circ h_2(\pm s + k)$. Differentiating twice yields $(c \circ h_1)''(s) = (c \circ h_2)''(\pm s + k)$, so $\kappa_1(s) = \kappa_2(\pm s + k)$. If we let $s = h_1^{-1}(t)$ then $\kappa_1(h_1^{-1}(t)) = \kappa_2(\pm h_1^{-1}(t) + k)$. It is straightforward to verify that $h_2^{-1}(t) = \pm h_1^{-1}(t) + k$, and the result follows. □

Example 4.5.3. We compute the curvature for the curve in Example 4.3.5. Using the formula obtained for $\widetilde{c}(t)$, we see

$$\widetilde{c}'(t) = \begin{pmatrix} \frac{-\sin(t/\sqrt{2})}{\sqrt{2}} \\ \frac{\cos(t/\sqrt{2})}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad \widetilde{c}''(t) = \begin{pmatrix} \frac{-\cos(t/\sqrt{2})}{2} \\ \frac{-\sin(t/\sqrt{2})}{2} \\ 0 \end{pmatrix} \quad \text{and} \quad \kappa(t) = \|\widetilde{c}''(t)\| = \frac{1}{2}. \ \Diamond$$

Finally, we need to verify that curvature is a function of the points in the image of the curve in $\mathbb{R}^3$, rather than a function of the particular choice of parametrization. We will need to assume, however, that our parametrization is injective, since at a point where the curve intersects itself there is not necessarily a single value for curvature (this problem does not arise when curvature is a function of the parametrization). Thus, we need to show that if we have two injective parametrized curves with the same images, then they yield the same curvature at each point in the image. This fact can be seen to follow from Lemmas 4.3.6 and 4.5.2; details are left to the reader.

It would be nice to have a formula for curvature for arbitrary regular curves that does not involve reparametrization (which can be difficult to carry out in practice). Such a formula will be given in Lemma 4.5.7. For later use we note that, combining the definitions of $N(t)$ and $\kappa(t)$, we obtain

$$T'(t) = \kappa(t)N(t). \tag{4.5.1}$$

Though curvature tells us a great deal about curves, it does not tell us all we need to know. There are different curves with the same curvature functions, for example the curves in part (2) of Example 4.5.1 and Example 4.5.3. Observe that one of these curves is contained in a plane whereas the other is not. What we wish to measure is the extent to which a curve is twisting out of the plane spanned by $T(t)$ and $N(t)$ for each t in the domain of the curve. Just as the bending of the curve is measured by the change in $T(t)$, using the length of $T'(t)$, it seems plausible that the change in the length of $B'(t)$ will tell us something about how the curve is twisting out of the plane spanned by $T(t)$ and $N(t)$. The quantity $\|B'(t)\|$ almost works, but like curvature it would always be non-negative, and it turns out that in the present case we can do a bit better and get a signed quantity. What we need is the analog for $B'(t)$ of Equation 4.5.1.

Recall the proof of the fact that $T'(t)$ is perpendicular to $T(t)$. Since $B(t)$ is also a unit vector, we can similarly deduce that $B'(t)$ is perpendicular to $B(t)$. Since $\{T(t), N(t), B(t)\}$ form an orthonormal basis for $\mathbb{R}^3$ for all t at which all three vectors are defined, it follows that $B'(t)$ is a linear combination of $T(t)$ and $N(t)$. Next, taking the derivative of both sides of the equation $\langle B(t), T(t)\rangle = 0$ yields

$$\begin{aligned} 0 &= \langle B'(t), T(t)\rangle + \langle B(t), T'(t)\rangle \\ &= \langle B'(t), T(t)\rangle + \langle B(t), \kappa(t)N(t)\rangle = \langle B'(t), T(t)\rangle, \end{aligned}$$

making use of Equation 4.5.1 and the fact that $\langle B(t), N(t)\rangle = 0$. It follows that

$B'(t)$ is a multiple of $N(t)$, which leads us to the following definition.

Definition. Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular unit speed curve. For each $t \in (a, b)$ the **torsion** of the curve at t is the unique real number $\tau(t)$ such that

$$B'(t) = -\tau(t)N(t). \tag{4.5.2}$$

The minus sign in the above equation is chosen for later convenience. Observe that $\tau(t) = -\langle B'(t), N(t)\rangle$, and thus torsion is a smooth function $\tau\colon (a, b) \to \mathbb{R}$. Finally, note that $|\tau(t)| = \|B'(t)\|$, which is analogous to the definition of $\kappa(t)$, though torsion can be negative.

Example 4.5.4. (1) We continue Example 4.5.1 part (2). It is not hard to see that

$$T(t) = \begin{pmatrix} -\sin\frac{t}{2} \\ \cos\frac{t}{2} \\ 0 \end{pmatrix}, \quad N(t) = \begin{pmatrix} -\cos\frac{t}{2} \\ -\sin\frac{t}{2} \\ 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Hence $B'(t) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$ for all t, and therefore $\tau(t) = 0$ for all t.

(2) We continue Example 4.5.3, where our calculations refer to the unit speed reparametrization $\widetilde{c}$. It can be computed that

$$T(t) = \begin{pmatrix} \frac{-\sin(t/\sqrt{2})}{\sqrt{2}} \\ \frac{\cos(t/\sqrt{2})}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}, \quad N(t) = \begin{pmatrix} -\cos\frac{t}{\sqrt{2}} \\ -\sin\frac{t}{\sqrt{2}} \\ 0 \end{pmatrix}, \quad B(t) = \begin{pmatrix} \sin\frac{t}{\sqrt{2}} \\ -\cos\frac{t}{\sqrt{2}} \\ 1 \end{pmatrix}.$$

Hence $B'(t) = \frac{1}{\sqrt{2}} \begin{pmatrix} \cos(t/\sqrt{2}) \\ -\sin(t/\sqrt{2}) \\ 0 \end{pmatrix}$ for all t, and therefore $\tau(t) = \frac{1}{\sqrt{2}}$ for all t.

Now suppose we start with the mirror image of the unit right circular helix, obtained by reflecting the unit right circular helix in the y-z plane, resulting in the curve $f\colon (-\infty, \infty) \to \mathbb{R}^3$ given by

$$f(t) = \begin{pmatrix} -\cos t \\ \sin t \\ t \end{pmatrix}.$$

It can be found by a similar computation that the torsion for this curve is constantly $-\frac{1}{\sqrt{2}}$, which is the negative of the torsion for the original helix. It is to

detect such differences of handedness that we made sure to allow torsion to be positive or negative. ◊

Although the torsion functions in the above example are constant, because we chose simple curves, torsion is not constant in general, as will be seen in some of the exercises. Just as for curvature, torsion is independent of the choice of parametrizations, and it can be computed for non-unit speed curves either by reparametrization or by the formula that will be given in Lemma 4.5.7.

Consider Equations 4.5.1 and 4.5.2. You will notice that we are missing a third equation, namely one giving the derivative of $N(t)$. The following theorem, which for completeness includes the two equations just mentioned, completes the picture, and really sums up much of what there is to say about curves in $\mathbb{R}^3$. For convenience we drop the argument t in the statement and proof of the following theorem.

Theorem 4.5.5 (Frenet–Serret Theorem). *Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular unit speed curve. Then*

$$\begin{aligned} T' &= \kappa N \\ N' &= -\kappa T + \tau B \\ B' &= -\tau N. \end{aligned}$$

Proof. Only the second equation remains to be proved. Just as we saw in Section 4.4 that $\langle T', T\rangle = 0$, the same argument shows that $\langle N', N\rangle = 0$, since N is a unit vector. Hence N' is a linear combination of T and B. If we write $N' = aT + bB$, take the inner product of this equation with each of T and B, and solve for a and b, we deduce that

$$N' = \langle N', T\rangle\, T + \langle N', B\rangle\, B.$$

Since $\langle N, T\rangle = 0$, we compute

$$0 = \langle N, T\rangle' = \langle N', T\rangle + \langle N, T'\rangle = \langle N', T\rangle + \langle N, \kappa N\rangle = \langle N', T\rangle + \kappa,$$

using Equation 4.5.1. Hence $\langle N', T\rangle = -\kappa$. Since $\langle N, B\rangle = 0$, we can similarly deduce that $\langle N', B\rangle = \tau$. □

The formulas in the above theorem are called the Frenet–Serret formulas. An easy way of remembering these formulas is to write them

$$\begin{pmatrix} T \\ N \\ B \end{pmatrix}' = \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} T \\ N \\ B \end{pmatrix},$$

though this expression is not strictly meaningful because the terms T, N and B are not numbers but column vectors.

A typical application of the Frenet–Serret Theorem is the following result.

Proposition 4.5.6. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular unit speed curve. The following are equivalent:*

(1) *The image of c lies in a plane;*
(2) *$B(t)$ is a constant vector;*
(3) *$\tau(t) = 0$ for all $t \in (a, b)$.*

Proof. (1) $\Rightarrow$ (2). This follows from Exercise 4.4.4.

(2) $\Rightarrow$ (1). Let $p \in (a, b)$ be a point. We compute

$$\begin{aligned}\frac{d}{dt}\langle c(t) - c(p), B(t)\rangle &= \langle c'(t), B(t)\rangle + \langle c(t) - c(p), B'(t)\rangle \\ &= \langle T(t), B(t)\rangle = 0,\end{aligned}$$

using the fact that $B(t)$ is constant. It follows that $\langle c(t) - c(p), B(t)\rangle$ is constant for all $t \in (a, b)$. If we plug in $t = p$ we deduce that this constant must be zero. Hence $c(t) - c(p)$ is perpendicular to the constant vector $B(t)$, and therefore the image of c lies entirely in the plane containing the point $c(p)$ and perpendicular to the constant vector $B(t)$.

(2) $\Leftrightarrow$ (3). This follows immediately from the third of the Frenet–Serret equations. $\square$

Finally, we give the promised formulas for computing the Frenet frame, curvature and torsion of a non-unit speed curve that avoids reparametrization. Once again we drop the argument t in the following lemma.

Lemma 4.5.7. *Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular curve. Then*

(i) $T = \dfrac{c'}{\|c'\|}$

(ii) $B = \dfrac{c' \times c''}{\|c' \times c''\|}$

(iii) $N = B \times T$

(iv) $\kappa = \dfrac{\|c' \times c''\|}{\|c'\|^3}$

(v) $\tau = \dfrac{\langle c' \times c'', c'''\rangle}{\|c' \times c''\|^2}.$

Proof. Part (i) is true by definition. We prove part (iv), leaving the other parts to the reader as Exercise 4.5.6. Let $\widetilde{c} = c \circ h$ be a unit speed reparametrization of c, where h is an appropriate diffeomorphism. Let $g = h^{-1}$, so that $c = \widetilde{c} \circ g$; let $\widetilde{T}$, $\widetilde{N}$ and $\widetilde{B}$ denote the Frenet frame for $\widetilde{c}$. Then

$$c'(t) = \widetilde{c}'(g(t))\, g'(t) = \widetilde{T}(g(t))\, g'(t),$$

and hence

$$\|c'(t)\| = \|\widetilde{T}(g(t))\|\, |g'(t)| = |g'(t)|.$$

We now have

$$\begin{aligned} c''(t) &= \widetilde{T}\,'(g(t))\,(g'(t))^2 + \widetilde{T}(g(t))\, g''(t) \\ &= \widetilde{\kappa}(g(t))\widetilde{N}(g(t))\,(g'(t))^2 + \widetilde{T}(g(t))\, g''(t)\,' \end{aligned}$$

and thus

$$\begin{aligned} c'(t) \times c''(t) &= g'(t)\, \widetilde{T}(g(t)) \times \left\{\widetilde{\kappa}(g(t))\widetilde{N}(g(t))\,(g'(t))^2 + \widetilde{T}(g(t))\, g''(t)\right\} \\ &= \widetilde{\kappa}(g(t))\widetilde{B}(g(t))\,(g'(t))^3. \end{aligned}$$

Therefore

$$\|c'(t) \times c''(t)\| = \widetilde{\kappa}(g(t))|g'(t)|^3 = \widetilde{\kappa}(g(t))\|c'(t)\|^3,$$

and hence

$$\widetilde{\kappa}(g(t)) = \frac{\|c'(t) \times c''(t)\|}{\|c'(t)\|^3}.$$

The desired result now follows, since by definition $\kappa(t) = \widetilde{\kappa}(h^{-1}(t)) = \widetilde{\kappa}(g(t))$. □

Exercises

4.5.1. Compute the curvature and torsion for the following curves.

(i) A circle of radius R (without loss of generality in the x-y plane, centered at the origin).

(ii) c: $(0, \infty) \to \mathbb{R}^3$ given by $c(t) = \frac{1}{2}\begin{pmatrix} t \\ 1/t \\ \sqrt{2}\ln t \end{pmatrix}$.

(iii*) The logarithmic spiral in Exercise 4.3.2.

(iv) d: $(0, \infty) \to \mathbb{R}^3$ given by $d(t) = \frac{1}{2}\begin{pmatrix} t \\ t^2 \\ t^3 \end{pmatrix}$.

4.5.2*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular curve. If A is a rotation matrix for $\mathbb{R}^3$ (that is, A is an orthogonal matrix with positive determinant), and if q is a vector in $\mathbb{R}^3$, then the curve $\widehat{c}\colon (a, b) \to \mathbb{R}^3$ given by

$$\widehat{c}(t) = Ac(t) + q$$

is the result of rotating and translating the image of c by A and q respectively. Show that the curvature and torsion of functions of $\widehat{c}$ are the same as for c.

4.5.3*. Let K and T be any real numbers such that $K > 0$. Show that there are numbers $a > 0$ and b such that the right circular helix

$$c(t) = \begin{pmatrix} a\cos t \\ a\sin t \\ bt \end{pmatrix}$$

has constant curvature K and constant torsion T.

4.5.4. Let $c\colon (a, b) \to \mathbb{R}^3$ be a unit speed curve such that $\kappa(t) = 0$ for all $t \in (a, b)$. Show that the image of c lies in a straight line.

4.5.5. Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular unit speed curve. Show that the image of c lies in a straight line iff there is a point $x_0 \in \mathbb{R}^3$ such that every tangent line to c goes through x_0.

4.5.6*. Prove Lemma 4.5.7 parts (ii), (iii) and (v).

4.5.7. Let $c\colon (a, b) \to \mathbb{R}^3$ be a strongly regular curve. Show that

$$\begin{aligned} T' &= \|c'\|\kappa N \\ N' &= -\|c'\|\kappa T + \|c'\|\tau B \\ B' &= -\|c'\|\tau N, \end{aligned}$$

where for convenience we drop the argument t.

4.6 Fundamental Theorem of Curves

Given a curve, we can clearly compute its curvature and torsion; can we go the other way? That is, if we are given curvature and torsion functions, is there a curve which has these values of curvature and torsion? We see from Exercise 4.5.5 that for any constant curvature and torsion functions there is at least one curve with the given curvature and torsion. The following theorem shows that

in fact arbitrary curvature and torsion functions completely determine the curve up to translation and rotation of the image of the curve. Note the restriction in the theorem to positive curvature, to avoid things like straight lines, for which torsion is not defined.

Theorem 4.6.1 (Fundamental Theorem of Curves). *Let $\bar{\kappa}, \bar{\tau}\colon (a, b) \to \mathbb{R}$ be smooth functions with $\bar{\kappa}(t) > 0$ for all $t \in (a, b)$. Then there is a strongly regular unit speed curve $c\colon (a, b) \to \mathbb{R}^3$ whose curvature and torsion functions are $\bar{\kappa}$ and $\bar{\tau}$ respectively. If $c_1, c_2\colon (a, b) \to \mathbb{R}^3$ are two such curves, then c_2 can be obtained from c_1 by a rotation and translation of $\mathbb{R}^3$.*

Proof. We essentially follow [M-P] (though the idea of the proof is standard). Let $p \in (a, b)$ be a point. We will show that there exists a unique strongly regular unit speed curve $c\colon (a, b) \to \mathbb{R}^3$ whose curvature and torsion functions are $\bar{\kappa}$ and $\bar{\tau}$ respectively, and such that

$$c(p) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad T(p) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad N(p) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad B(p) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{4.6.1}$$

The precise statement of the theorem then follows straightforwardly using Exercise 4.5.2.

The idea of the proof is to solve the Frenet–Serret equations, which are differential equations, in order to find the purported tangent, normal and binormal vectors of the desired curve. Integrating the tangent vector will then give us the curve we are looking for. More precisely, consider the following system of linear differential equations with initial conditions, which are just the Frenet–Serret equations written out in coordinates:

$$\begin{aligned}
u_1'(t) &= \bar{\kappa}(t)\, u_4(t) \\
u_2'(t) &= \bar{\kappa}(t)\, u_5(t) \\
u_3'(t) &= \bar{\kappa}(t)\, u_6(t) \\
u_4'(t) &= -\bar{\kappa}(t)\, u_1(t) + \bar{\tau}(t)\, u_7(t) \\
u_5'(t) &= -\bar{\kappa}(t)\, u_2(t) + \bar{\tau}(t)\, u_8(t) \\
u_6'(t) &= -\bar{\kappa}(t)\, u_3(t) + \bar{\tau}(t)\, u_9(t) \\
u_7'(t) &= -\bar{\tau}(t)\, u_4(t) \\
u_8'(t) &= -\bar{\tau}(t)\, u_5(t) \\
u_9'(t) &= -\bar{\tau}(t)\, u_6(t),
\end{aligned} \tag{4.6.2}$$

$$\begin{aligned} u_1(p) &= 1, u_2(p) = 0, u_3(p) = 0, \\ u_4(p) &= 0, u_5(p) = 1, u_6(p) = 0, \\ u_7(p) &= 0, u_8(p) = 0, u_9(p) = 1. \end{aligned} \tag{4.6.3}$$

By Theorem 4.2.6 there are smooth functions $u_1, \ldots, u_9$: $(a, b) \to \mathbb{R}$ satisfying Equations 4.6.2 and 4.6.3, and these functions are unique. For convenience, we define smooth vector-valued functions X_1, X_2, X_3: $(a, b) \to \mathbb{R}^3$ by

$$X_1(t) = \begin{pmatrix} u_1(t) \\ u_2(t) \\ u_3(t) \end{pmatrix}, \quad X_2(t) = \begin{pmatrix} u_4(t) \\ u_5(t) \\ u_6(t) \end{pmatrix}, \quad X_3(t) = \begin{pmatrix} u_7(t) \\ u_8(t) \\ u_9(t) \end{pmatrix},$$

where we think of X_1, X_2 and X_3 as the tangent, normal and binormal vectors respectively. Since the u_i satisfy Equations 4.6.2 and 4.6.3, we have

$$\begin{aligned} X_1'(t) &= \bar{\kappa}(t)\, X_2(t) \\ X_2'(t) &= -\bar{\kappa}(t)\, X_1(t) + \bar{\tau}(t)\, X_3(t) \\ X_3'(t) &= -\bar{\tau}(t)\, X_2(t), \end{aligned} \tag{4.6.4}$$

$$X_1(p) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \quad X_2(p) = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \quad X_3(p) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}. \tag{4.6.5}$$

We see in Exercise 4.6.2 that for all $t \in (a, b)$ the three vectors $\{X_1(t), X_2(t), X_3(t)\}$ form an orthonormal basis for $\mathbb{R}^3$.

We now define a curve c: $(a, b) \to \mathbb{R}^3$ by

$$c(t) = \int_p^t X_1(s)\, ds.$$

From the Fundamental Theorem of Calculus it follows that $c'(t) = X_1(t)$. Thus c is smooth (since X_1 is smooth) and is unit speed (since $X_1(t)$ is a unit vector for all t by the claim). Let T and κ denote the unit tangent vector and curvature of c respectively (for convenience we will drop the argument t throughout most of this proof). Evidently $T = X_1$. Using Equation 4.6.4 we further compute that

$$T' = X_1' = \bar{\kappa}\, X_2. \tag{4.6.6}$$

Since $\bar{\kappa} > 0$ and X_2 is never the zero vector (by the claim) we deduce that T' is never the zero vector. Hence c is a strongly regular curve, so the unit normal,

unit binormal and torsion of c are all defined; we denote these quantities N, B and τ respectively. Using the Frenet–Serret Theorem (Theorem 4.5.5) and Equation 4.6.4, once again, we have

$$\kappa N = T' = X_1' = \bar{\kappa} X_2. \tag{4.6.7}$$

Taking the norm of both sides, and using the facts that $\|N\| = \|X_2\| = 1, \kappa \geq 0$ and $\bar{\kappa} > 0$, we deduce that $\kappa = \bar{\kappa}$. Cancelling by κ on both sides of Equation 4.6.7 yields $N = X_2$.

Since $\{T, N, B\}$ and $\{X_1, X_2, X_3\}$ are both orthonormal bases for $\mathbb{R}^3$, and since $T = X_1$ and $N = X_2$, it follows that $B(t) = \pm X_3(t)$; the continuity of $B(t)$ and $X_3(t)$ imply that the $\pm$ sign is independent of t. However, we observe that $B(p) = T(p) \times N(p)$ by definition and $X_3(p) = X_1(p) \times X_2(p)$ by Equation 4.6.5; hence $B(p) = X_3(p)$, and it follows that $B = X_3$ for all $t \in (a, b)$. Finally, using the Frenet-Serret Theorem and Equation 4.6.4 yet again, we have

$$-\tau N = B' = X_3' = -\bar{\tau} X_2. \tag{4.6.8}$$

Since $N = X_2$, and this vector is never the zero vector, we deduce that $\tau = \bar{\tau}$. We thus see that the curvature and torsion of the curve c are as desired. That c satisfies Equation 4.6.1 follows from the definition of c and Equation 4.6.5. Thus c has all the properties it is supposed to have. As for the uniqueness of c, we note that the functions $\{T, N, B\}$ are uniquely determined by the differential equation and initial conditions given in Equations 4.6.2 and 4.6.3. Thus c is uniquely determined since it is the unique solution to the differential equation and initial condition

$$c' = T, \quad c(p) = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}. \quad \square$$

Although in theory the above proof actually gives a procedure for finding the curve if the curvature and torsion are known, the bulk of the procedure involves solving some differential equations, which in practice can be quite difficult. A worked out example may be found in [M-P, §2-5].

Exercises

4.6.1. Show that curvature alone does not determine a curve up to rotation and translation.

4.6.2*. In this exercise we complete the missing piece of the proof of Theorem 4.6.1, namely to show that for all $t \in (a, b)$ the three vectors $\{X_1(t), X_2(t), X_3(t)\}$ form an orthonormal basis for $\mathbb{R}^3$. This proof has a few steps.

Step 1: For each pair of numbers $i, j \in \{1, 2, 3\}$, define a function $p_{ij}: (a, b) \to \mathbb{R}$ by

$$p_{ij}(t) = \langle X_i(t), X_j(t) \rangle.$$

Show that

$$\begin{aligned}
p'_{11} &= \bar{\kappa}\, p_{21} + \bar{\kappa}\, p_{12} \\
p'_{12} &= \bar{\kappa}\, p_{22} - \bar{\kappa}\, p_{11} + \bar{\tau}\, p_{13} \\
p'_{13} &= \bar{\kappa}\, p_{23} - \bar{\tau}\, p_{12} \\
p'_{21} &= -\bar{\kappa}\, p_{11} + \bar{\tau}\, p_{31} + \bar{\kappa}\, p_{22} \\
p'_{22} &= -\bar{\kappa}\, p_{12} + \bar{\tau}\, p_{32} - \bar{\kappa}\, p_{21} + \bar{\tau}\, p_{23} \\
p'_{23} &= -\bar{\kappa}\, p_{13} + \bar{\tau}\, p_{33} - \bar{\tau}\, p_{22} \\
p'_{31} &= -\bar{\tau}\, p_{21} + \bar{\kappa}\, p_{32} \\
p'_{32} &= -\bar{\tau}\, p_{22} - \bar{\kappa}\, p_{31} + \bar{\tau}\, p_{33} \\
p'_{33} &= -\bar{\tau}\, p_{23} - \bar{\tau}\, p_{32},
\end{aligned} \tag{4.6.9}$$

and

$$p_{ij}(p) = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases} \tag{4.6.10}$$

Step 2: For each pair of numbers $i, j \in \{1, 2, 3\}$ define a function $\delta_{ij}: (a, b) \to \mathbb{R}$ by

$$\delta_{ij}(t) = \begin{cases} 1 & \text{if } i = j; \\ 0 & \text{if } i \neq j. \end{cases}$$

Show that these functions satisfy Equations 4.6.9 and 4.6.10.

Step 3: Deduce the desired result.

4.7 Planar Curves

In our discussion of surfaces we will encounter certain curves with images in planes in $\mathbb{R}^3$. Without loss of generality we will assume throughout this section

that the curves under consideration have their images in $\mathbb{R}^2$, except when stated otherwise. Everything that we have said about curves in $\mathbb{R}^3$ certainly applies to curves whose images lie in planes. Planar curves all have zero torsion by Proposition 4.5.6, so we loose torsion as a useful concept. On the other hand, we can take advantage of planarity to strengthen the concept of curvature. By definition the curvature function of curves in $\mathbb{R}^3$ is always non-negative; there is no meaningful geometric way to define positive versus negative curvature for a curve in $\mathbb{R}^3$, since there is no way to say that it is bending in a particular direction. We cannot say that the curve is bending "away" from itself as opposed to bending "toward" itself, since such a description depends entirely upon how we look at the curve. For curves in $\mathbb{R}^2$, however, there is an inherent way to describe bending, namely as either clockwise or counterclockwise.

A curve, being parametrized, comes with a direction in which it is traversed; in Figure 4.7.1 (i) we see a curve with a given direction, and in Figure 4.7.1 (ii) is a curve with the same image, but parametrized in the other direction. The curve in Figure 4.7.1 (i) is bending in a counterclockwise direction from the point of view of a bug walking along the curve in the given direction; from the point of view of a bug walking along the curve in Figure 4.7.1 (ii), the bending is clockwise. The notion of clockwise vs. counterclockwise bending, which will give us positive or negative planar curvature, is thus seen to depend upon the given parametrization of the curve.

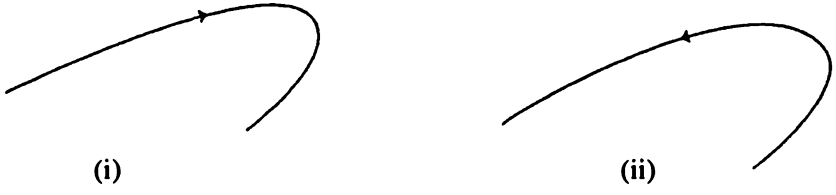


Figure 4.7.1

Technically, we proceed by defining variants of $T(t)$ and $N(t)$ for planar curves. For a planar curve $B(t)$ is constant, so we will not make use of it.

Definition. Let $c\colon (a, b) \to \mathbb{R}^2$ be a smooth curve. For each $t \in (a, b)$ such that $\|c'(t)\| \neq 0$, the **planar unit tangent vector** and **planar unit normal vector** to the curve at t, denoted $\bar{T}(t)$ and $\bar{N}(t)$ respectively, are defined by

letting $\bar{T}(t) = T(t)$ and letting $\bar{N}(t)$ be the unit vector obtained by rotating $\bar{T}(t)$ counterclockwise by 90°. (See Figure 4.7.2.)

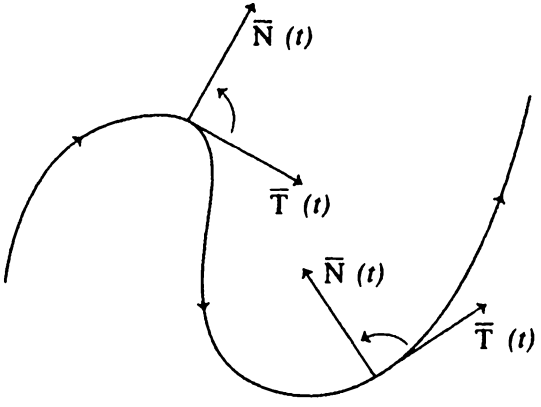


Figure 4.7.2

Note that $\bar{N}(t) = \pm N(t)$. Recall that rotating a vector in $\mathbb{R}^2$ counterclockwise by 90° is obtained by multiplying the vector (when written as a column vector with respect to the standard basis) by the matrix $\begin{pmatrix} 0 & -1 \\ 1 & 0 \end{pmatrix}$, and is thus a smooth operation.

Example 4.7.1. The unit circle in $\mathbb{R}^2$ centered at the origin can be parametrized in various ways; consider two such parametrizations, namely c_a, c_b: $(-\infty, \infty) \to \mathbb{R}^2$ given by

$$c_a(t) = \begin{pmatrix} \cos t \\ \sin t \end{pmatrix} \quad \text{and} \quad c_b(t) = \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}.$$

Both these parametrizations are unit speed, but c_a traverses the unit circle in the counterclockwise direction, whereas c_b traverses the unit circle in the clockwise direction. For c_a we compute

$$\bar{T}(t) = \begin{pmatrix} -\sin t \\ \cos t \end{pmatrix} \quad \text{and} \quad \bar{N}(t) = \begin{pmatrix} -\cos t \\ -\sin t \end{pmatrix},$$

and for c_b we compute

$$\bar{T}(t) = \begin{pmatrix} -\sin t \\ -\cos t \end{pmatrix} \quad \text{and} \quad \bar{N}(t) = \begin{pmatrix} \cos t \\ -\sin t \end{pmatrix}. \quad \diamond$$

It can be seen that $\bar{T}'(t)$ is perpendicular to $\bar{T}(t)$, using the same proof as for $T(t)$. Since the vectors $\{\bar{T}(t), \bar{N}(t)\}$ form an orthonormal basis for $\mathbb{R}^2$, it follows that $\bar{T}'(t)$ is a multiple of $\bar{N}(t)$. We are thus led to the following definition, which is analogous to the definition of torsion for curves in $\mathbb{R}^3$.

Definition. Let $c\colon (a, b) \to \mathbb{R}^2$ be a unit speed curve. The **planar curvature** of c at $t \in (a, b)$ is defined to be the unique real number $\bar{\kappa}(t)$ such that

$$\bar{T}'(t) = \bar{\kappa}(t)\bar{N}(t). \tag{4.7.1}$$

The above equation is entirely analogous to Equation 4.5.1, although in this case the equation is taken as the definition of planar curvature $\bar{\kappa}(t)$. Observe that $\bar{\kappa}(t)$ can be negative. However, since $\|\bar{N}(t)\| = 1$ for all t it follows that

$$|\bar{\kappa}(t)| = \|\bar{T}'(t)\| = \|T'(t)\| = \kappa(t).$$

Hence the only new information $\bar{\kappa}(t)$ brings is that it takes into account the direction of bending by being positive or negative.

Example 4.7.2. We continue Example 4.7.1. For c_a we have

$$\bar{T}'(t) = \begin{pmatrix} -\cos t \\ -\sin t \end{pmatrix},$$

and hence $\bar{\kappa}(t) = 1$ for all t. For c_b we have

$$\bar{T}'(t) = \begin{pmatrix} -\cos t \\ \sin t \end{pmatrix},$$

and hence $\bar{\kappa}(t) = -1$ for all t. Observe that the difference in sign of the planar curvature of these two parametrizations of the unit circle corresponds to the difference in the orientation of the two parametrizations. ◊

As before, if we start with a non-unit speed curve we can calculate planar curvature by reparametrizing the curve so that it is unit speed and then calculating the planar curvature. See Exercises 4.7.3 and 4.7.4 for formulas for the planar curvature of a non-unit speed plane curve.

We will need to be able to measure the planar curvature of curves in arbitrary planes in $\mathbb{R}^n$, not just in $\mathbb{R}^2$. In an arbitrary plane there is no inherent notion of which direction of rotation is "clockwise" and which is "counterclockwise," since it depends upon how we look at the plane. So, for each plane we can arbitrarily choose a direction of rotation and call it clockwise. Such a choice

is equivalent to choosing an ordered basis for the plane; if the plane does not contain the origin, consider a plane parallel to it that does contain the origin, and choose an ordered basis for this parallel plane. If the ordered basis is $\{x_1, x_2\}$, then we consider that counterclockwise rotation is given by a rotation of the plane taking x_1 to x_2 via the angle between the vectors that is less than π. If the plane is sitting in $\mathbb{R}^3$ (as will be the case later on) this choice is also equivalent to choosing a perpendicular direction to the plane, and using the right hand rule. No matter which approach we take, there are always two possible ways of making the choice. Such a choice is called an **orientation** of the plane.

Once we have made a choice of orientation for a given plane in $\mathbb{R}^n$, we can then compute planar curvature of the curve in the plane just as for curves in $\mathbb{R}^2$. If we were to choose the opposite orientation it is not hard to see that the planar curvature of the curve would change its sign. Thus planar curvature in arbitrary planes is well-defined only with respect to a chosen orientation of the plane.

Exercises

4.7.1. Find $\bar{T}(t)$ and $\bar{N}(t)$ and $\bar{\kappa}(t)$ for the following curves.

(i) $c\colon (-\infty, \infty) \to \mathbb{R}^2$ given by $c(t) = \begin{pmatrix} t \\ t^2 \end{pmatrix}$;

(ii) $d\colon (0, \infty) \to \mathbb{R}^3$ given by $d(t) = \begin{pmatrix} \ln t \\ e^t \end{pmatrix}$.

4.7.2. Find the planar curvature of the logarithmic spiral in Exercise 4.3.2.

4.7.3*. Find a formula for planar curvature analogous to the formula for curvature given in Lemma 4.5.7 (iv). In particular, if a curve $c\colon (a, b) \to \mathbb{R}^2$ is given by $c(t) = \begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix}$, where $c_1, c_2\colon (a, b) \to \mathbb{R}$ are smooth functions, express the planar curvature in terms of c_1 and c_2.

4.7.4*. Find a formula for the planar curvature of the graph of a function of the form $y = f(x)$.

4.7.5. Let $c\colon (a, b) \to \mathbb{R}^2$ be a smooth curve such that the image of c is entirely contained in the closed ball in $\mathbb{R}^2$ of radius R centered at the origin. If $\|c(q)\| = R$ for some $q \in (a, b)$, show that $\kappa(q) \geq \frac{1}{R}$.

Endnotes

Notes for Section 4.2

(A) The openness of the domains of smooth functions is crucial if we are to use the standard definition of derivatives. It is possible to extend the definition of what it means to be differentiable to non-open subsets of Euclidean space, but we will avoid doing so to help clarify the nature of smoothness, and to point the way more clearly to smooth manifolds.

(B) Though it may seem like a stringent requirement that all smooth functions used are infinitely differentiable, that is, all partial derivatives of all orders exist and are continuous, such functions are actually quite plentiful. It would be possible to deal with functions that are only twice or thrice differentiable, but the gain in doing so is negligible, and is outweighed by the nuisance of having to pay closer attention in all statements of theorems and proofs to the exact level of differentiability.

(C) See [MU1, Chapter I] for a clarification of the relation between functions of various degrees of differentiability.

Notes for Section 4.3

(A) Some books use the terminology "parametrized by arc-length" to mean what we call "unit speed."

(B) See [JU] for a literary look at the smoothness of curves.

Notes for Section 4.4

In single variable Calculus, the curves used are the graphs of functions of the form $y = f(x)$. Such functions have the form $f\colon (a, b) \to \mathbb{R}$. Graphs in the x-y plane of such functions have one axis representing the independent variable (namely x) and one axis representing the dependent variable (namely y). By contrast, when we view the "graph" of a function of the form $c\colon (a, b) \to \mathbb{R}^3$ we are actually looking at the image of the function, since $\mathbb{R}^3$ only has room for the dependent variables (namely x, y and z). Hence our definition of the tangent vector looks slightly different than that seen in the Calculus of a single variable.

CHAPTER V

Smooth Surfaces

5.1 Introduction

The outline for our study of smooth surfaces is somewhat like our study of smooth curves, and we will be making use of the material concerning smooth functions discussed in Section 4.2. In both cases we define what it means to be "smooth" via parametrizations, define tangent and normal vectors, and then search for geometric quantities such as curvature. There are, however, two fundamental technical differences between surfaces and curves: A surface usually cannot be presented via a single parametrization, and there is no analog for surfaces to unit speed parametrizations. These complications lead to some rather dry technical discussions in the present chapter. As for curves, we will restrict our attention to smooth surfaces in $\mathbb{R}^3$.

5.2 Coordinate Patches and Smooth Surfaces

A topological surface, which is by definition a subset of $\mathbb{R}^n$, need not have a parametrization given as part of its definition. A parametrization for a surface would minimally mean a map from an open subset of $\mathbb{R}^2$ onto the surface (or a piece of one, since even simple surfaces such as the sphere will need to be parametrized in pieces). To make precise this notion, which will be needed to define smooth surfaces, recall the definition of a topological surface: A subset $Q \subset \mathbb{R}^n$ is a topological surface if for each point $p \in Q$ there is an open subset $W \subset Q$ containing p such that W is homeomorphic to the open disk $\text{int } D^2 \subset \mathbb{R}^2$. It would make no difference if the open disk $\text{int } D^2$ were replaced by an arbitrary open set $U \subset \mathbb{R}^2$, and we will do so for convenience. Now, fix the point $p \in Q$. Let us denote by x a choice of homeomorphism $U \to W$. Such a map is the foundation for our parametrization of surfaces. Since an arbitrary continuous map can have a very crinkled image, we add the requirement that x be a smooth function. Even a smooth map can degenerate, however, and in order to avoid this problem we need more conditions on the map x.

If $x: U \to \mathbb{R}^3$ is a smooth map, we can write x in coordinates as

$$x(v) = \begin{pmatrix} f_1(v) \\ f_2(v) \\ f_3(v) \end{pmatrix}$$

for all $v \in U$, where $f_1, f_2, f_3: U \to \mathbb{R}$ are smooth real-valued functions. The function x thus has two partial derivatives, namely

$$x_1 = \begin{pmatrix} \frac{\partial f_1}{\partial s} \\ \frac{\partial f_2}{\partial s} \\ \frac{\partial f_3}{\partial s} \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} \frac{\partial f_1}{\partial t} \\ \frac{\partial f_2}{\partial t} \\ \frac{\partial f_3}{\partial t} \end{pmatrix}.$$

Observe that x_1 and x_2 are the columns of the Jacobian matrix Dx. To insure that the map x is not degenerate we make the following definition.

Definition. Let $U \subset \mathbb{R}^2$ be an open set. A smooth map $x: U \to \mathbb{R}^3$ is a **coordinate patch** if it is injective and if $x_1 \times x_2 \neq 0$ at all points of U. $\Diamond$

Equivalent conditions to the above definition are that x_1 and x_2 are linearly independent, or that the Jacobian matrix Dx has rank 2 at each point of U.

Example 5.2.1. (1) Let $x: \mathbb{R}^2 \to \mathbb{R}^3$ be given by

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ s^2 + t^2 \end{pmatrix}.$$

To see that x is injective, observe that $x(\binom{s}{t}) = x(\binom{u}{v})$ implies $s = u$ and $t = v$. The partial derivatives of x are

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 2s \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ 2t \end{pmatrix}.$$

Hence

$$x_1 \times x_2 = \begin{pmatrix} -2s \\ -2t \\ 1 \end{pmatrix},$$

which is never zero. Therefore x is a coordinate patch.

(2) Let $y: \mathbb{R}^2 \to \mathbb{R}^3$ be given by

$$y\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s^3 \\ t^3 \\ 1 \end{pmatrix}.$$

To see that y is injective, observe that $y(\binom{s}{t}) = y(\binom{u}{v})$ implies $s^3 = u^3$ and $t^3 = v^3$, and hence $s = u$ and $t = v$. The partial derivatives of y are

$$y_1 = \begin{pmatrix} 3s^2 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad y_2 = \begin{pmatrix} 0 \\ 3t^2 \\ 0 \end{pmatrix}.$$

Hence

$$y_1 \times y_2 = \begin{pmatrix} 0 \\ 0 \\ 9s^2t^2 \end{pmatrix},$$

which is zero whenever $s = 0$ or $t = 0$. Therefore y is not a coordinate patch. ◊

Coordinate patches now allow us to define smooth surfaces.

Definition. A subset $M \subset \mathbb{R}^3$ is a **smooth surface** if it is a topological surface and if for each point $p \in M$ there is a coordinate patch $x: U \to M \subset \mathbb{R}^3$ such that $p \in x(U)$. ◊

In practice, rather than finding a coordinate patch for each point p in a smooth surface we simply find coordinate patches whose images cover the entire surface. In many cases more than one coordinate patch will be needed. We will not give explicit proofs that the surfaces under consideration are indeed topological surfaces, since it will usually be quite straightforward.

Example 5.2.2. (1) Any open subset $U \subset \mathbb{R}^2$ is a smooth surface covered by the coordinate patch $x: U \to \mathbb{R}^3$ given by

$$x(\begin{pmatrix} s \\ t \end{pmatrix}) = \begin{pmatrix} s \\ t \\ 0 \end{pmatrix}.$$

(2) The unit sphere S^2 is a smooth surface. One method to cover S^2 with coordinate patches is to use the six coordinate patches $x^1, x^2, y^1, y^2, z^1, z^2$: int $D^2 \to$

S^2 given by

$$x^1\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} \sqrt{1-s^2-t^2} \\ s \\ t \end{pmatrix}, \quad x^2\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} -\sqrt{1-s^2-t^2} \\ s \\ t \end{pmatrix},$$

$$y^1\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ \sqrt{1-s^2-t^2} \\ t \end{pmatrix}, \quad y^2\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ -\sqrt{1-s^2-t^2} \\ t \end{pmatrix},$$

$$z^1\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ \sqrt{1-s^2-t^2} \end{pmatrix}, \quad z^2\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ -\sqrt{1-s^2-t^2} \end{pmatrix}.$$

Each of these coordinate patches covers an open hemisphere; see Figure 5.2.1. We leave it to the reader to verify that these six maps are actually coordinate patches. ◊

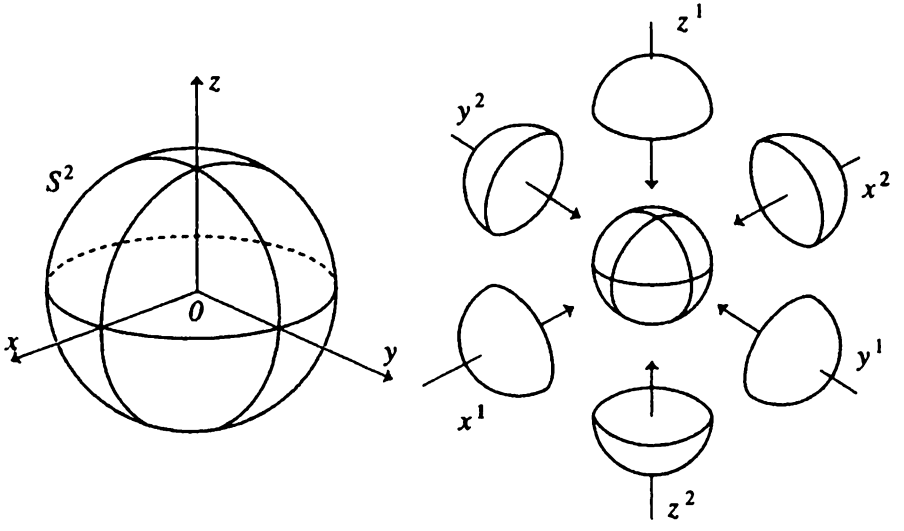


Figure 5.2.1

What is the relation between smooth surfaces, topological surfaces and simplicial surfaces? By definition any smooth surface is a topological surface. It then follows from Theorem 3.4.5 that every compact smooth surface can be triangulated. Is every topological surface also a smooth surface? Whereas not

every topological surface is smooth as given (the surface of a cube, for example), it turns out that every topological surface is homeomorphic to a smooth one. The usual route to proving this fact is to observe that every topological surface is homeomorphic to a simplicial surface, and then to prove that every simplicial surface is homeomorphic to a smooth surface; see [CA1] or [WH3]. Essentially, the classes of all topological surfaces, all simplicial surfaces and all smooth surfaces are equivalent. The analogous result is not true in higher dimensions.

When dealing with various aspects of smooth surfaces, we will have two fundamental ways of proceeding: with reference to a coordinate patch (also known as working in "local coordinates"), or without. This issue is analogous to discussing linear maps with or without reference to a choice of basis. (In fact, we will see that the choice of a coordinate patch yields a choice of basis for the tangent plane, to be defined, at each point in the image of the coordinate patch). Just as in linear algebra we learn what happens under a change of basis, we need to learn how to go from one choice of coordinate patch to another. We start with the following simple lemma.

Lemma 5.2.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. If $V \subset \mathbb{R}^2$ is an open set and $\phi: V \to U$ is a diffeomorphism, then $x \circ \phi: V \to M$ is a coordinate patch.*

Proof. Since x and ϕ are both injective, so is $x \circ \phi$. By the chain rule $D(x \circ \phi) = Dx\, D\phi$. The hypotheses of the lemma imply that both Dx and $D\phi$ are linear maps of rank 2, and hence $D(x \circ \phi)$ has rank 2. It follows that $x \circ \phi$ is a coordinate patch. □

We now turn to the more difficult situation, where two coordinate patches with overlapping images are given.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ and $y: V \to M$ be coordinate patches whose images overlap (see Figure 5.2.2). Let

$$A_{xy} = x^{-1}(x(U) \cap y(V)) \quad \text{and} \quad A_{yx} = y^{-1}(x(U) \cap y(V)). \tag{5.2.1}$$

The composite map

$$y^{-1} \circ x | A_{xy}: A_{xy} \to A_{yx}$$

is called the **change of coordinate function** from x to y and is denoted $\phi_{x,y}$ (entirely non-standard notation). ◊

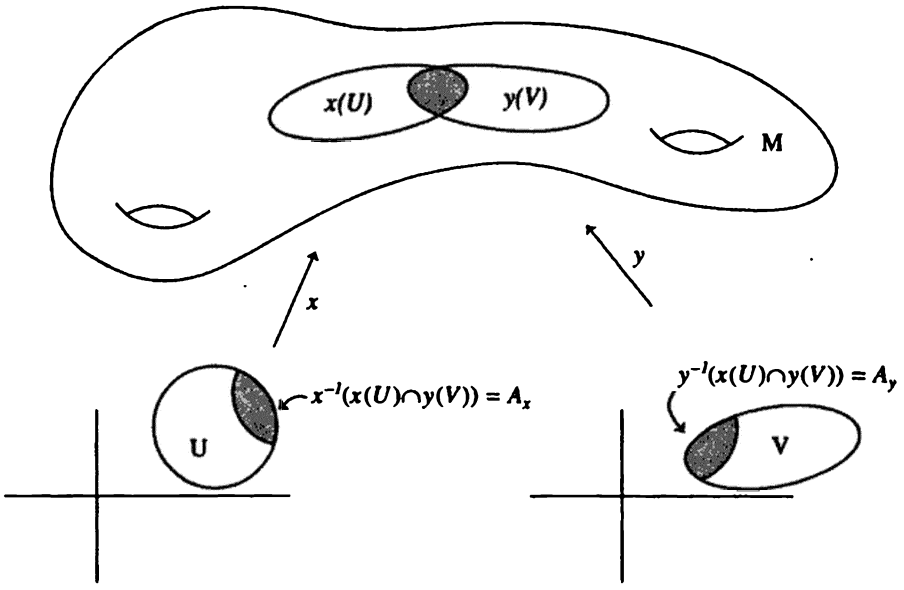


Figure 5.2.2

Since x and y in the above definition are injective, they are bijections onto their images, so we can validly refer to maps x^{-1} and y^{-1} in the above definition (though only as maps of sets, with no mention of differentiability). It is easy to see that $\phi_{x,y}$ is bijective. Also, note that

$$x|A_{xy} = y|A_{yx} \circ \phi_{x,y}. \tag{5.2.2}$$

Example 5.2.4. We continue Example 5.2.2 (2), computing ϕ_{x^1,y^1}. It is seen that $(x^1)^{-1}: x^1(\text{int } D^2) \to \text{int } D^2$ and $(y^1)^{-1}: y^1(\text{int } D^2) \to \text{int } D^2$ are given by

$$(x^1)^{-1}\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \begin{pmatrix} y \\ z \end{pmatrix} \quad \text{and} \quad (y^1)^{-1}\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \begin{pmatrix} x \\ z \end{pmatrix}.$$

Note that

$$x^1(\text{int } D^2) \cap y^1(\text{int } D^2) = \left\{\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in S^2 \mid x > 0, y > 0\right\}.$$

It can then be seen that

$$\begin{aligned} A_{x^1} &= (x^1)^{-1}(x^1(\text{int } D^2) \cap y^1(\text{int } D^2)) = \{\binom{s}{t} \in \text{int } D^2 \mid s > 0\} \\ &= (y^1)^{-1}(x^1(\text{int } D^2) \cap y^1(\text{int } D^2)) = A_{y^1}. \end{aligned}$$

We now have

$$\phi_{x^1,y^1}(\binom{s}{t}) = (y^1)^{-1} \circ (x^1)|A_{x^1}(\binom{s}{t}) = \begin{pmatrix} \sqrt{1-s^2-t^2} \\ t \end{pmatrix},$$

which is smooth on the given domain. ◊

Our main technical result concerning coordinate patches is the following proposition, the proof of which is based on Theorem 4.2.2 (and hence ultimately on the Inverse Function Theorem), as well as Invariance of Domain (Theorem 2.2.1).

Proposition 5.2.5. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point.*

(i) *If $x: U \to M$ is a coordinate patch, then x is a homeomorphism from U onto $x(U)$. Further, the set $x(U)$ is open in M.*

(ii) *Let $T \subset \mathbb{R}^n$ be open, and let $f: T \to x(U) \subset \mathbb{R}^3$ be smooth. Then $x^{-1} \circ f: T \to U$ is smooth.*

(iii) *If $x^1: U_1 \to M$ and $x^2: U_2 \to M$ are coordinate patches with overlapping images, then the change of coordinate function ϕ_{x^1,x^2} is a diffeomorphism.*

Proof. (i). It is a standard result in real analysis that smooth maps are necessarily continuous (see [BT, §19]), so x is continuous. We now show that for each point $p \in U$ there is an open subset $W \subset U$ containing p such that $x(W)$ is open in M and $x|W$ is a homeomorphism from W onto $x(W)$. Since x is a bijection onto its image (by injectivity), it will then follow from Exercise 1.4.4 that x is a homeomorphism onto its image and that $x(U)$ is open in M.

Let $p \in U$ be a point. By assumption the Jacobian matrix of x has rank 2 at each point of U. Hence we can apply Theorem 4.2.2 to deduce that there is an open subset $W \subset U$ containing p, an open subset $V \subset \mathbb{R}^3$ containing $x(p)$, and a smooth map $G: V \to \mathbb{R}^3$ such that $G(V)$ is open in $\mathbb{R}^3$, that G is a diffeomorphism from V onto $G(V)$, that $x(W) \subset V$ and that

$$G \circ x(\binom{s}{t}) = \begin{pmatrix} s \\ t \\ 0 \end{pmatrix},$$

for all $\binom{s}{t} \in W$. Since G is a smooth map, note that it is also continuous.

Let $\pi_{12}: \mathbb{R}^3 \to \mathbb{R}^2$ be projection onto the first two coordinates, that is, $\pi_{12}(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \binom{x}{y}$ for all $\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3$. The map π_{12} is seen to be continuous by an argument similar to the one used in Example 1.3.1 (2). Observe that $\pi_{12} \circ G|x(W) \circ x|W = 1_W$, the identity map on W. Since $x|W: W \to x(W)$ is bijective, it is straightforward to deduce that $(x|W)^{-1} = \pi_{12} \circ G|x(W)$. It now follows that $(x|W)^{-1}$ is continuous, because both π_{12} and $G|x(W)$ are; we therefore deduce that $x|W$ is a homeomorphism from W onto $x(W)$.

It remains to be seen that $x(W)$ is open in M. Since M is a topological surface, there is some open set $Z \subset M$ containing $x(p)$ that is homeomorphic to int D^2; let $h: Z \to$ int D^2 be a homeomorphism. By replacing W with some small enough subset of it, we can assume that $x(W) \subset Z$. It follows that $h \circ x$ is a homeomorphism from W onto the set $h(x(W)) \subset$ int $D^2 \subset \mathbb{R}^2$. By Exercise 2.2.2 (a corollary to Invariance of Domain) we deduce that $h(x(W))$ is open in $\mathbb{R}^2$, and hence in int D^2. Therefore $x(W)$ is open in Z, and hence it is open in M.

(ii). It will suffice to show that for each point $q \in T$ there is an open subset $S \subset T$ containing q such that $x^{-1} \circ f|S$ is smooth. Fix a point $q \in T$, and let $p = x^{-1}(f(q))$. Let G, V, W and π_{12} be as in the proof of part (i). We define the set S to be $S = f^{-1}(x(W))$; this set indeed contains q, and is open in T by the continuity of f. We now see that

$$\begin{aligned} x^{-1} \circ f|S &= (x|W)^{-1} \circ f|S = (\pi_{12} \circ G|x(W)) \circ f|S \\ &= \pi_{12} \circ G \circ f|S. \end{aligned}$$

Hence $x^{-1} \circ f|S$ is smooth, being the composition of smooth maps.

(iii). This follows immediately from part (ii) of the lemma, letting f be x^2 restricted to the domain of ϕ_{x^1,x^2}. □

The following lemma is essentially the converse of part (iii) of the above proposition.

Lemma 5.2.6. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. If $V \subset \mathbb{R}^2$ is an open subset, and $f: V \to U$ is a diffeomorphism, then the composition $x \circ f: V \to M$ is a coordinate patch, and f is the change of coordinate function $\phi_{x \circ f, x}$.*

Proof. Exercise 5.2.5. □

As an application of Proposition 5.2.5, we make the following useful observation about curves in surfaces.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $c: (a, b) \to x(U)$ be a smooth curve. The **pull-back** of c by x is the curve $\bar{c} = x^{-1} \circ c: (a, b) \to U \subset \mathbb{R}^2$. The **coordinate functions** of c with respect to x are the functions $c_1, c_2: (a, b) \to \mathbb{R}$ such that $\bar{c}(t) = \begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix} \in \mathbb{R}^2$ for all $t \in (a, b)$. $\Diamond$

It follows immediately from Proposition 5.2.5 (ii) that the function $\bar{c}$ in the above definition is smooth, and hence so are c_1 and c_2. Observe that $c(t) = x(\bar{c}(t)) = x(\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix})$. For the rest of this book we will use the notation $\bar{c}$ in the above sense. Similarly, for any point $p \in x(U)$ be will let $\bar{p} = x^{-1}(p) \in U$.

We need to define the appropriate type of maps between smooth surfaces. A surface in $\mathbb{R}^3$ is not an open subset, so our usual notion of what it means for a map to be smooth (which we sometimes refer to as "Euclidean smooth" for clarity) cannot be applied directly. The technically convenient approach to take is to pull back a given function to open subsets of $\mathbb{R}^2$ via coordinate patches. We start with a lemma.

Lemma 5.2.7. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $f: M \to \mathbb{R}$ be a map and let $x: U \to M$ and $y: V \to M$ be coordinate patches whose images overlap. Then $f \circ x|A_{xy}$ is Euclidean smooth iff $f \circ y|A_{yx}$ is Euclidean smooth.*

Proof. Exercise 5.2.11. $\square$

We can now safely make the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $f: M \to \mathbb{R}$ be a map. The map f is **surface smooth** (or **smooth** if there is no ambiguity) if for each point $p \in M$ there is a coordinate patch $x: U \to M$ with $p \in x(U)$ such that the composition $f \circ x: U \to \mathbb{R}$ is smooth. $\Diamond$

In practice, there are two common ways of presenting a smooth map on a surface $M \subset \mathbb{R}^3$. First, suppose we have a map $f: W \to \mathbb{R}$ for some open set $W \subset \mathbb{R}^3$ containing M; if f is Euclidean smooth it follows that the restriction of f to M is surface smooth (observe that $f \circ x$ must be Euclidean smooth for any coordinate patch $x: U \to M$, since the composition of two Euclidean smooth maps is Euclidean smooth). Second, suppose we are given a coordinate patch $x: U \to M$; if we then specify a smooth map $\bar{f}: U \to \mathbb{R}$, we can define

a map $f: x(U) \to \mathbb{R}$ by letting $f = \bar{f} \circ x^{-1}$. It is straightforward to show that f is surface smooth. In practice we use this method by giving a formula for $f(x(\binom{s}{t}))$ in terms of s and t.

Example 5.2.8. (1) The function $f: S^2 \to \mathbb{R}$ given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \frac{xyz}{x^2+y^2+z^2}$ is a smooth function, since it is smooth in the standard sense on all of $\mathbb{R}^3 - O_3$, an open subset of $\mathbb{R}^3$ containing S^2.

(2) Let M be the smooth surface that is the image of the coordinate patch given in Example 5.2.1 (1); this surface is the paraboloid $z = x^2 + y^2$. Define a function $f: M \to \mathbb{R}$ by setting $f(x(\binom{s}{t})) = \sin s$. This function is smooth, since it can be expressed as $f = \bar{f} \circ x^{-1}$, where $\bar{f}(\binom{s}{t}) = \sin s$, and this latter function is certainly Euclidean smooth. ◊

The following definition broadens the previous one.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $f: M \to \mathbb{R}^n$ be a map. Let $f_1, \ldots, f_n: M \to \mathbb{R}$ be the component functions of f. The map f is **surface smooth** if each of the maps $f_1, \ldots, f_n$ is surface smooth. ◊

Now suppose that we have two smooth surfaces $M, N \subset \mathbb{R}^3$, and a map $f: M \to N$. There are two approaches to the question of what it would mean for the map f to be smooth. On the one hand, since N is in $\mathbb{R}^3$, we can simply view f as a map $M \to \mathbb{R}^3$, and determine whether f is surface smooth as just defined. On the other hand, we might wish to use coordinate patches for N, just as we did for M in the definition of surface smoothness of maps $M \to \mathbb{R}$. The following lemma shows that both approaches are equivalent.

Lemma 5.2.9. *Let $M, N \subset \mathbb{R}^3$ be smooth surfaces, and let $f: M \to N$ be a map. Then f is surface smooth as a map $M \to \mathbb{R}^3$ iff for each point $p \in M$ there is a coordinate patch $x: U \to M$ with $p \in x(U)$ and a coordinate patch $y: V \to N$ with $f(p) \in y(V)$ such that the composition*

$$y^{-1} \circ f \circ x|x^{-1}(f^{-1}(y(V))): x^{-1}(f^{-1}(y(V))) \to V \subset \mathbb{R}^2$$

is Euclidean smooth.

Proof. See Figure 5.2.3. For convenience, let $A = x^{-1}(f^{-1}(y(V)))$. Suppose first that f is surface smooth as a map $M \to \mathbb{R}^3$. Let $x: U \to M$ and $y: V \to N$ be coordinate patches with $f(x(U)) \cap y(V) \neq \emptyset$. It is straightforward to see

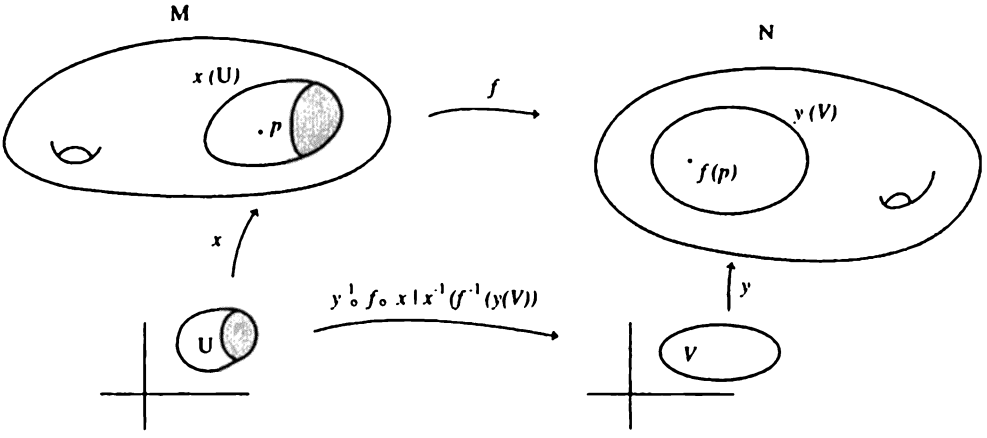


Figure 5.2.3

that $f \circ x: U \to N \subset \mathbb{R}^3$ is Euclidean smooth. Now, using Proposition 5.2.5 (i) we know that $y(V)$ is open in M. The Euclidean smoothness of $f \circ x$ and x imply that these maps are continuous. Using Proposition 5.2.5 (i) it follows that f is continuous. Hence the set A is an open subset of U, and thus of $\mathbb{R}^2$. Therefore $f \circ x|A$ is Euclidean smooth. It now follows from Proposition 5.2.5 (ii) that $y^{-1} \circ f \circ x|A: A \to V$ is Euclidean smooth.

Conversely, suppose that for each point $p \in M$ there is a coordinate patch $x: U \to M$ with $p \in x(U)$ and a coordinate patch $y: V \to N$ with $f(p) \in y(V)$ such that $y^{-1} \circ f \circ x|A: A \to V$ is Euclidean smooth. Since y is Euclidean smooth it follows that $y \circ (y^{-1} \circ f \circ x|A)$ is Euclidean smooth, and this latter map equals $f \circ x|A$. Since $x|A$ is a coordinate patch for M, the image of which contains p, and since p is arbitrary, it follows that f is surface smooth as a map $M \to \mathbb{R}^3$. □

We can now make the following definition.

Definition. Let $M, N \subset \mathbb{R}^3$ be smooth surfaces, and let $f: M \to N$ be a map. The map f is **smooth** if either of the conditions in Lemma 5.2.9 hold. The map f is a **diffeomorphism** if it is bijective and both it and its inverse are smooth. ◊

Exercises

5.2.1*. Show that a compact smooth surface in $\mathbb{R}^3$ cannot be the image of a single coordinate patch.

5.2.2. Are the following functions coordinate patches?

(i) $x\colon \mathbb{R}^2 \to \mathbb{R}^3$ given by $x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ e^{3s-t} \end{pmatrix}$,

(ii) $y\colon (\mathbb{R} - \{0\}) \times (\mathbb{R} - \{0\}) \to \mathbb{R}^3$ given by $y\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s^2 \\ t^2 \\ 1 \end{pmatrix}$,

(iii) $z\colon \mathbb{R}^2 \to \mathbb{R}^3$ given by $z\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} e^{st} \\ t \\ 4 \end{pmatrix}$.

5.2.3. Show that $S^1 \times \mathbb{R} \subset \mathbb{R}^3$ is a smooth surface.

5.2.4. Let $M \subset \mathbb{R}^3$ be a smooth surface. Let A be a non-singular 3×3 matrix and let $q \in \mathbb{R}^3$ be a vector. If $F\colon \mathbb{R}^3 \to \mathbb{R}^3$ is the map given by $F(v) = Av + q$, show that $F(M)$ is a smooth surface. The map F is affine linear, as discussed in the Appendix.

5.2.5*. Prove Lemma 5.2.6.

5.2.6*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x\colon U \to M$ be a coordinate patch and let $c\colon (a, b) \to x(U) \subset \mathbb{R}^3$ be a smooth curve. Show that

$$c'(t) = c_1'(t)x_1(\bar{c}(t)) + c_2'(t)x_2(\bar{c}(t))$$

for all $t \in (a, b)$.

5.2.7. Let $M \subset \mathbb{R}^3$ be a smooth surface, and suppose that $N \subset M$ is open in M. Show that N is a smooth surface.

5.2.8*. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $f\colon M \to \mathbb{R}^n$ be a smooth map. If $T \subset \mathbb{R}^m$ is open and $g\colon T \to M \subset \mathbb{R}^3$ is a Euclidean smooth map, show that $f \circ g\colon T \to \mathbb{R}^n$ is smooth.

5.2.9*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x\colon U \to M$ be a coordinate patch. Show that the map x is a diffeomorphism from U to $x(U)$ (both thought of as smooth surfaces).

5.2.10*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $V \subset \mathbb{R}^2$ be an open set. If $y\colon V \to M$ is a smooth map such that $y(V)$ is open in M and y is a

diffeomorphism from V onto $y(V)$ (both thought of as smooth surfaces), show that y is a coordinate patch.

5.2.11*. Prove Lemma 5.2.7.

5.3 Examples of Smooth Surfaces

We consider here a number of special types of surfaces. We will return to these surfaces repeatedly in examples and exercises.

(1) Monge Patches

These surfaces are the graphs of smooth functions $f: U \to \mathbb{R}$, where $U \subset \mathbb{R}^2$ is an open set. Such a surface M can be covered with one coordinate patch $x: U \to M$ given by

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ f\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) \end{pmatrix}.$$

To see that a monge patch is indeed a smooth surface we need to verify that the function x is a coordinate patch. To see that x is injective, observe that $x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = x\left(\begin{pmatrix} u \\ v \end{pmatrix}\right)$ implies $s = u$ and $t = v$. The partial derivatives of x are

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ f_s \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ f_t \end{pmatrix},$$

where f_s and f_t denote the partial derivatives of f with respect to s and t respectively. Hence

$$x_1 \times x_2 = \begin{pmatrix} -f_s \\ -f_t \\ 1 \end{pmatrix},$$

which is never zero. Therefore x is a coordinate patch. An example of a monge patch is the graph of the function $f\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = st$, which is a saddle surface; see Figure 5.3.1.

(2) Surfaces of Revolution

A surface of revolution is obtained by rotating an injective regular planar curve (called the profile curve) in $\mathbb{R}^3$, where the rotation is about a line that does not intersect the curve and is contained in the plane containing the curve.

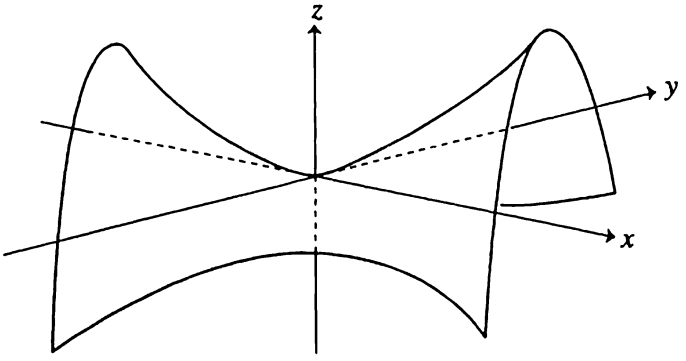


Figure 5.3.1

Without loss of generality we will assume that the profile curve lies in the x-z plane, and that the axis of revolution is the z-axis; we assume that the x-coordinate of each point in the image of the profile curve is positive. See Figure 5.3.2.

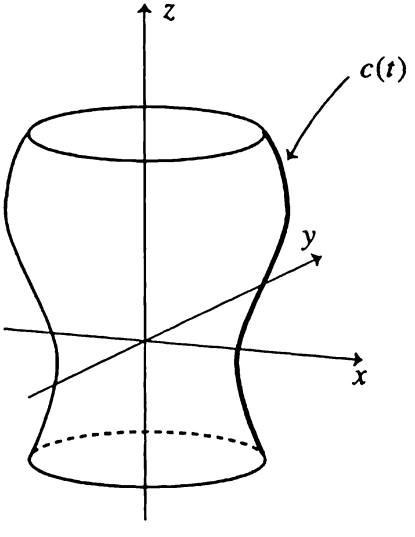


Figure 5.3.2

The most general way of describing a profile curve is by an injective map c from an open interval $J_1 \subset \mathbb{R}$ to the x-z plane taking the form $c(t) = \binom{r(t)}{z(t)}$, where $r(t) > 0$ for all t. The regularity of c is insured by insisting that $r'(t)$

and $z'(t)$ are not both zero for each $t \in J_1$. The surface of revolution can then be covered by coordinate patches of the form

$$x\left(\begin{pmatrix} t \\ \theta \end{pmatrix}\right) = \begin{pmatrix} r(t)\cos\theta \\ r(t)\sin\theta \\ z(t) \end{pmatrix}, \tag{5.3.1}$$

where we use t and θ instead of s and t to conform to standard notation for surfaces of revolution. The domain of x is of the form $J_1 \times J_2$, where $J_2 \subset \mathbb{R}$ is an open interval of length 2π (a closed interval of length 2π would allow the image of x to cover the entire surface of revolution, but then x would not be injective). The reader is asked in Exercise 5.3.1 to verify that maps x of the above form are coordinate patches and that surfaces of revolution are indeed smooth surfaces. The curves on the surface of revolution obtained by holding θ constant and varying t are called meridians (or longitudes), and the curves on the surface obtained by holding t constant and varying θ are called circles of latitude (or parallels). See Figure 5.3.3.

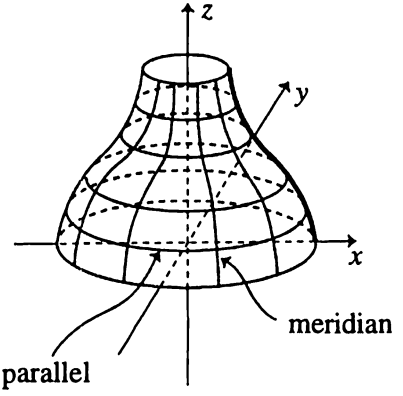


Figure 5.3.3

There are many familiar examples of surfaces of revolution. A sphere of radius R (missing the north and south poles) is obtained by rotating a semi-circle of radius R centered at the origin. A typical coordinate patch is given by

$$x\left(\begin{pmatrix} t \\ \theta \end{pmatrix}\right) = \begin{pmatrix} R\cos t\cos\theta \\ R\cos t\sin\theta \\ R\sin t \end{pmatrix}. \tag{5.3.2}$$

A torus of large radius R and small radius r is obtained by rotating a circle in the x-z plane with radius r and center $\begin{pmatrix} R \\ 0 \\ 0 \end{pmatrix}$), as in Figure 5.3.4. A typical coordinate patch is given by

$$x\left(\begin{pmatrix} t \\ \theta \end{pmatrix}\right) = \begin{pmatrix} (R + r\cos t)\cos\theta \\ (R + r\cos t)\sin\theta \\ r\sin t \end{pmatrix}. \tag{5.3.3}$$

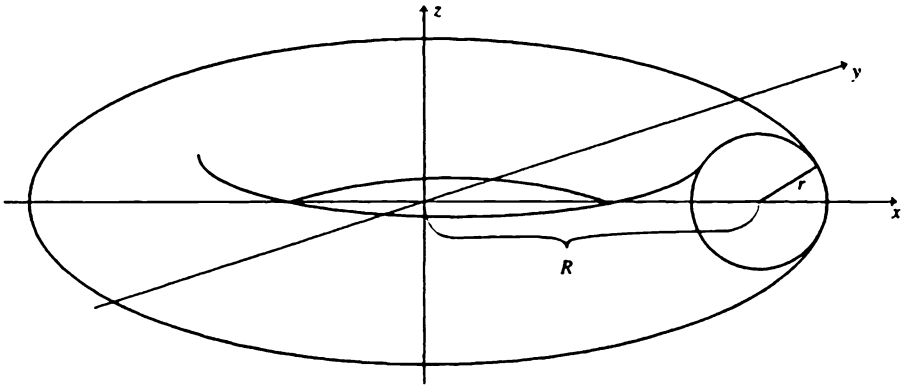


Figure 5.3.4

The coordinate patch of a surface of revolution can be given in a slightly simpler form if the profile curve can be realized as the graph of a function of the form $x = r(z)$, where we assume that $r(z) > 0$ for all z in the domain of the function. We can then parametrize the profile curve as $c(t) = \begin{pmatrix} r(t) \\ t \end{pmatrix}$, and proceed as before.

(3) Ruled Surfaces

Intuitively, ruled surfaces are obtained by moving a straight line through $\mathbb{R}^3$. If we fix a point on this line and trace its path as the line is moved we obtain a smooth curve in $\mathbb{R}^3$; for each point on this curve we can describe a corresponding line on the surface. See Figure 5.3.5. More precisely, we start by specifying two smooth functions $c\colon J_1 \to \mathbb{R}^3$ and $\delta\colon J_2 \to \mathbb{R}^3$, where $J_1, J_2 \subset \mathbb{R}$ are open intervals. We think of c as the smooth curve described above, and we assume that it is injective and regular. For each point $c(s)$ in the image of c, we think of $\delta(s)$ as giving the direction of the line that is in the ruled surface and

that contains $c(s)$. We can therefore parametrize a ruled surface by a coordinate patch $x: J_1 \times J_2 \to \mathbb{R}^3$ of the form

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = c(s) + t\delta(s).$$

Note that for fixed s we do obtain a straight line as we vary t; these lines are called **rulings**. If we impose no restrictions on c and δ then x need not satisfy either of the conditions of a coordinate patch. However, we compute that $x_1 = c'(s) + t\delta'(s)$ and $x_2 = \delta(s)$, and we observe that these vectors are linearly independent for small t if $c'(s)$ and $\delta(s)$ are themselves linearly independent. Further, if t is taken to be small enough then it can be seen that x is injective. We will assume that these criteria hold.

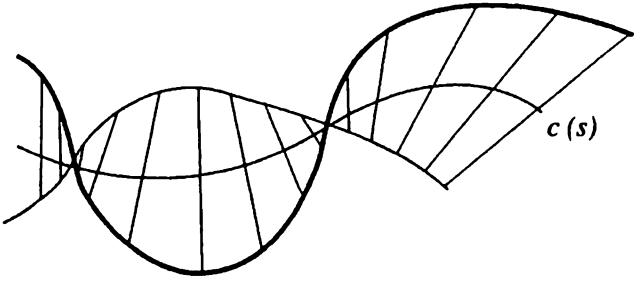


Figure 5.3.5

A nice example of a ruled surface is the parametrization of the Möbius strip (from which one line segment has been removed) given by

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} \cos s \\ \sin s \\ 0 \end{pmatrix} + t \begin{pmatrix} \cos \frac{s}{2} \cos s \\ \cos \frac{s}{2} \sin s \\ \sin \frac{s}{2} \end{pmatrix},$$

where $s \in (-\pi, \pi)$ and $t \in (-\frac{1}{2}, \frac{1}{2})$. As we go around the unit circle in the x-y plane once, the rulings make a 180° turn. See Figure 5.3.6.

Another ruled surface is the right helicoid, which is a smoothed out version of a spiral staircase. See Figure 5.3.7. This surface can be parametrized by

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} 0 \\ 0 \\ bs \end{pmatrix} + t \begin{pmatrix} \cos s \\ \sin s \\ 0 \end{pmatrix},$$

where $b \neq 0$.

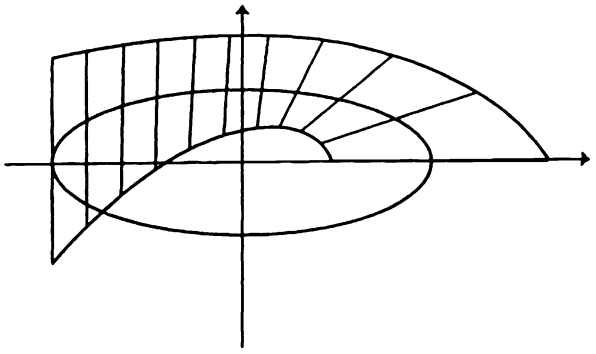


Figure 5.3.6

Although in general the function $\delta(s)$ is chosen independently of the curve $c(s)$, there is a useful case where $\delta(s)$ does depend upon $c(s)$. The **rectifying developable surface** generated by an injective unit-speed curve c: $(a, b) \to \mathbf{R}^3$ is the ruled surface parametrized by

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = c(s) + tB(s),$$

where $B(s)$ is the unit binormal to the curve $c(s)$. For example, if c is a planar curve then $B(s)$ is constant by Proposition 4.5.6, and the rectifying developable surface generated by such c is a right cylinder with cross section the image of the curve c.

A somewhat surprising example of a ruled surface, seen in Figure 5.3.8, is the elliptic hyperboloid of one sheet, given by the quadratic equation

$$\frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2} = 1.$$

That this surface is ruled can be seen by the parametrization

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} a\cos s \\ b\sin s \\ 0 \end{pmatrix} + t\begin{pmatrix} -a\sin s \\ b\cos s \\ c \end{pmatrix}.$$

The reader may verify that this parametrization does indeed yield the elliptic paraboloid. Figure 5.3.9 shows how to visualize this ruling of the surface in the case of a circular cross section. But not only is this surface ruled, in fact there is a second way to rule it! Simply twist the string construction in Figure 5.3.9 the other way. The other ruling is left to the reader to construct.

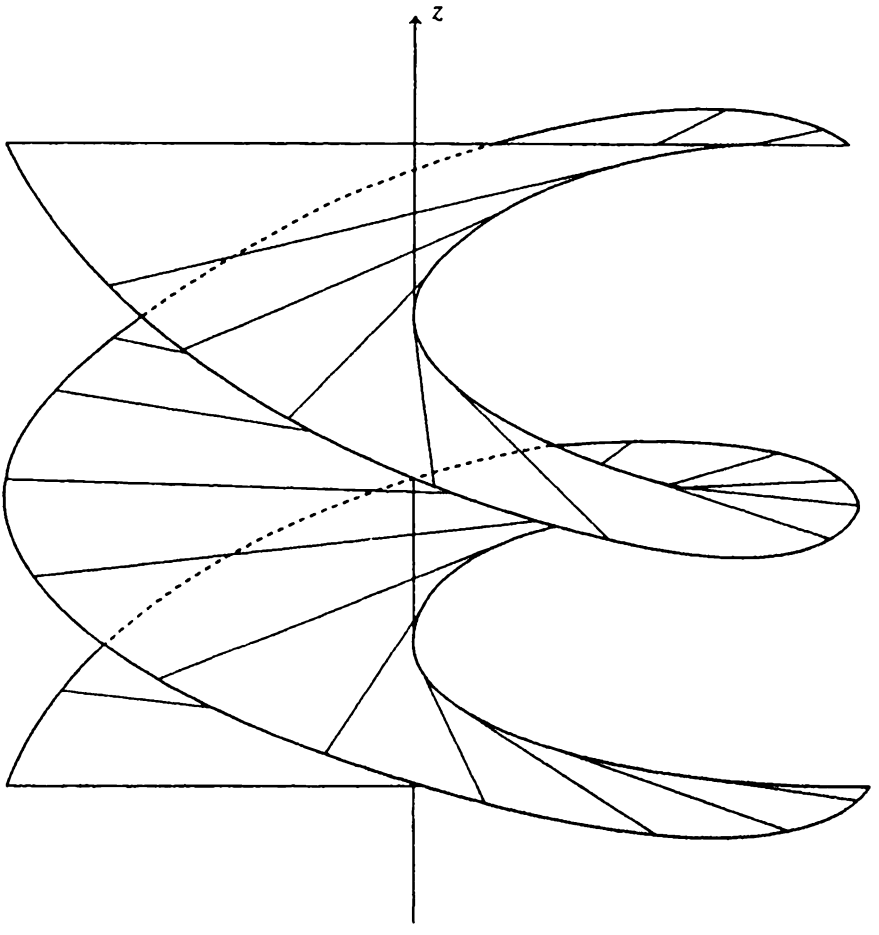


Figure 5.3.7

(4) Level Surfaces

These are smooth surfaces of the form $F^{-1}(a) = \{p \in \mathbb{R}^3 \mid F(p) = a\}$ for some smooth function $F: V \to \mathbb{R}$, where $V \subset \mathbb{R}^3$ is an open set and $a \in \mathbb{R}$ is a number. An example of such a surface is the elliptic hyperboloid of one sheet mentioned in the discussion of ruled surfaces, which could be written as $F^{-1}(1)$ for $F\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \frac{x^2}{a^2} + \frac{y^2}{b^2} - \frac{z^2}{c^2}$. If we arbitrarily write down a smooth function $F: V \to \mathbb{R}$ and arbitrarily choose a number $a \in \mathbb{R}$, then the set $F_{-1}(a)$

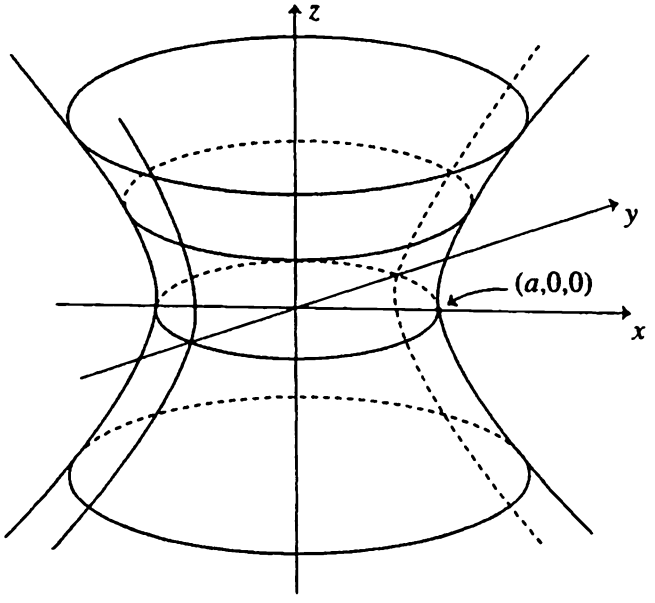


Figure 5.3.8

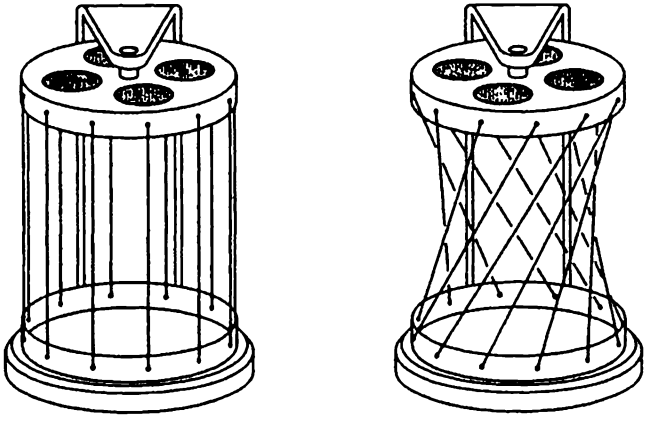


Figure 5.3.9

might not be a surface, though we do obtain a surface in many instances. The following definition and proposition allow us to verify that certain sets of the form $F^{-1}(a)$ are indeed smooth surfaces; the proof of the proposition, which uses the Inverse Function Theorem, is in Appendix A5.1.

Definition. Let $V \subset \mathbb{R}^3$ be an open set and let $F: V \to \mathbb{R}$ be a smooth function. A number $a \in \mathbb{R}$ is called a **regular value** of F if $DF(p)$ has rank 1 for all $p \in F^{-1}(a)$; if the set $F^{-1}(a)$ is empty then a is automatically considered regular. $\Diamond$

The condition that $DF(p)$ has rank 1 (the maximal possible rank) in the above definition is equivalent to the condition that at least one of $\frac{\partial F}{\partial u_1}(p)$, $\frac{\partial F}{\partial u_2}(p)$ and $\frac{\partial F}{\partial u_3}(p)$ is not zero.

Proposition 5.3.1. *Let $V \subset \mathbb{R}^3$ be an open set, let $F: V \to \mathbb{R}$ be a smooth function, and let $a \in \mathbb{R}$ be a regular value of F. Then if the set $F^{-1}(a)$ is non-empty it is a smooth surface.*

Example 5.3.2. The hyperbolic paraboloid is the quadric surface given by the equation

$$z = \frac{x^2}{a^2} - \frac{y^2}{b^2}.$$

To prove that the set of all points in $\mathbb{R}^3$ that satisfy this equation is truly a smooth surface, let $F\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \frac{x^2}{a^2} - \frac{y^2}{b^2} - z$. The hyperbolic paraboloid is then $F^{-1}(0)$. We compute that $\frac{\partial F}{\partial x} = \frac{2x}{a^2}$, $\frac{\partial F}{\partial y} = -\frac{2y}{b^2}$ and $\frac{\partial F}{\partial z} = -1$. Since the last of these partial derivatives is never zero, then every point in $\mathbb{R}$ is a regular value of F. Hence Proposition 5.3.1 implies that the hyperbolic paraboloid is a smooth surface. $\Diamond$

There is, unfortunately, no simple way to write explicit coordinate charts for level surfaces, so these surfaces will at times be hard to deal with computationally, even though many familiar surfaces, such as spheres and ellipsoids, are level surfaces.

Exercises

5.3.1*. Verify that maps of the form given in Equation 5.3.1 are coordinate patches, and that surfaces of revolution are smooth surfaces.

5.3.2. Show that the ellipsoid $\frac{x^2}{a^2} + \frac{y^2}{b^2} + \frac{z^2}{c^2} = 1$ is a smooth surface.

5.3.3. Explicitly parametrize the hyperboloid of one sheet $\frac{x^2}{4} + \frac{y^2}{4} - \frac{z^2}{9} = 1$.

5.3.4. Show that a circular cone in $\mathbb{R}^3$ with its vertex removed is each of the four types of surfaces considered in this section. Without loss of generality assume that the cone has its vertex at the origin, has angle α at the vertex (where $0 < \alpha < \pi$), and has as its axis of symmetry the positive z-axis.

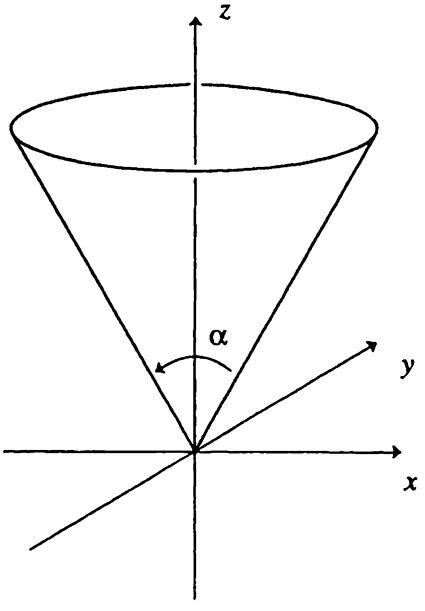


Figure 5.3.10

5.3.5. Show that the hyperbolic paraboloid (in Example 5.3.2) can be ruled in two different ways.

5.3.6. Find an example of a smooth function $F: V \to \mathbb{R}$ and a number $a \in \mathbb{R}$ such that a is not a regular value of F but such that the set $F^{-1}(a)$ is nonetheless a smooth surface. Hence we cannot improve Proposition 5.3.1 to be "if and only if."

5.4 Tangent and Normal Vectors

For a smooth surface, the tangent plane at a point on the surface is the plane that best "fits" the surface at the point of tangency. Smoothness is crucial here, since

at a non-smooth point on a surface there need not be a well-defined plane of best fit, as in Figure 5.4.1. We will start with a definition of individual tangent vectors at a point on a surface, and then prove that this collection of vectors form a plane. Although we might think intuitively of tangent vectors at a point p on a surface as starting at the point p, for convenience we will consider all vectors as being translated so that they to start at the origin. The intuitive idea behind the following definition is the observation that any vector tangent to a smooth surface is tangent to some smooth curve in the surface.

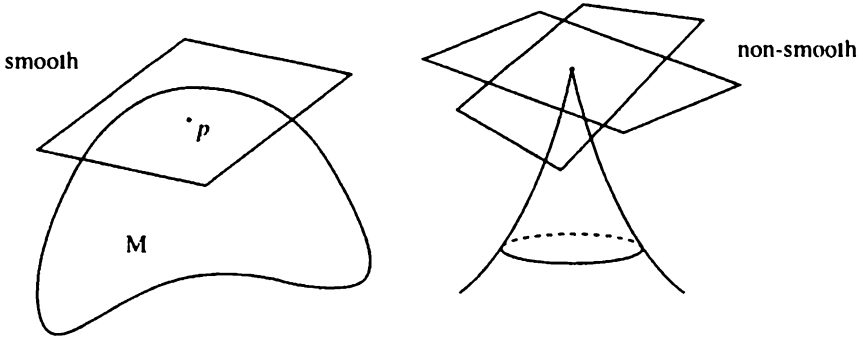


Figure 5.4.1

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. A vector v in $\mathbb{R}^3$ is a **tangent vector** to M at p if there is a curve $c\colon (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ such that $c(0) = p$ and $c'(0) = v$. The collection of all tangent vectors to M at p is denoted T_pM, and it is called the **tangent plane** to M at p (see Figure 5.4.2). $\Diamond$

For each tangent vector v as in the above definition there will be many corresponding curves c.

Example 5.4.1. (1) Consider $\mathbb{R}^2$ as a smooth surface in $\mathbb{R}^3$. Since the tangent vector to any curve in $\mathbb{R}^2$ is itself in $\mathbb{R}^2$, it is not hard to see that for each point $p \in \mathbb{R}^2$ we have $T_p\mathbb{R}^2 = \mathbb{R}^2$.

(2) Let $p = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in S^2$, and let $v \in \mathbb{R}^3$ be any vector of the form $v = \begin{pmatrix} 0 \\ a \\ b \end{pmatrix}$.

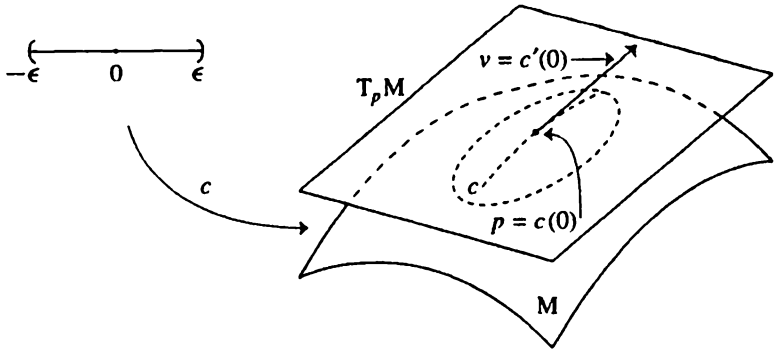


Figure 5.4.2

Letting $D = \sqrt{a^2 + b^2}$, then the curve $c: (-\pi, \pi) \to S^2$ given by

$$c(t) = \begin{pmatrix} \cos Dt \\ (a/D) \sin Dt \\ (b/D) \sin Dt \end{pmatrix}$$

has the property that $c(0) = p$ and $c'(0) = v$. We thus see that T_pS^2 contains the y-z plane. That T_pS^2 in fact equals the y-z plane will follow from Lemma 5.4.2 below. ◊

The following lemma not only shows that the collection of tangent vectors at a point forms a plane, but it also gives an easy way of finding tangent planes.

Lemma 5.4.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Then T_pM is a two-dimensional vector space. Moreover, if $x: U \to M$ is a coordinate patch such that $p \in x(U)$, then T_pM is the subspace of $\mathbb{R}^3$ spanned by the vectors $\{x_1, x_2\}$ evaluated at $x^{-1}(p)$.*

Proof. It will suffice to prove the second part of the lemma. Let $\text{span}\{x_1, x_2\}$ denote the vector subspace of $\mathbb{R}^3$ spanned by the vectors $\{x_1, x_2\}$, where x_1 and x_2 are evaluated at $x^{-1}(p)$; we will prove that $T_pM = \text{span}\{x_1, x_2\}$. First, suppose $v \in T_pM$, so that $v = c'(0)$ for some curve $c: (-\epsilon, \epsilon) \to M$ such that $c(0) = p$. We can assume without loss of generality that the image of c is entirely contained in $x(U)$. Let $c_1, c_2: (-\epsilon, \epsilon) \to \mathbb{R}$ be the coordinate functions of the curve c with respect to x. Using Exercise 5.2.6 we see that $v = c'(0) = c_1'(0)x_1 + c_2'(0)x_2 \in \text{span}\{x_1, x_2\}$.

Next, suppose $v \in \text{span}\{x_1, x_2\}$, so that $v = ax_1 + bx_2$ for some real numbers a and b. Without loss of generality we may assume that the coordinate patch x has been chosen so that $O_2 \in \mathbb{R}^2$ is contained in the set U, and that $x(O_2) = p$ (the reader should show why it is safe to make this assumption). Since U is open, all points in $\mathbb{R}^2$ close enough to the origin are contained in U. We now define a curve c: $(-\epsilon, \epsilon) \to x(U) \subset M$ by the formula $c(t) = x(\begin{pmatrix} at \\ bt \end{pmatrix})$ for some small enough number ϵ (to insure that the points $\begin{pmatrix} at \\ bt \end{pmatrix}$ are contained in U for all $t \in (-\epsilon, \epsilon)$). It is now straightforward to verify that $c(0) = p$ and $c'(0) = ax_1 + bx_2 = v$. Hence $v \in T_pM$. □

The proof of this lemma follows a very typical pattern. The set T_pM was defined without reference to a coordinate patch, but we needed coordinate patches to prove something about T_pM. We could have taken the alternative route of defining T_pM as the vector space spanned by $\{x_1, x_2\}$ evaluated at $x^{-1}(p)$, making the lemma unnecessary, but we would then have had to have proved that the definition did not depend upon the choice of coordinate patch.

Example 5.4.3. We use the parametrization of S^2 given in Equation 5.3.2, with $R = 1$. Let p be a point on the equator of S^2, so that $p = x(\begin{pmatrix} 0 \\ \theta \end{pmatrix})$ for some θ. We compute that

$$x_1(\begin{pmatrix} t \\ \theta \end{pmatrix}) = \begin{pmatrix} -\sin t \cos\theta \\ -\sin t \sin\theta \\ \cos t \end{pmatrix} \quad \text{and} \quad x_2(t, \theta) = \begin{pmatrix} -\cos t \sin\theta \\ \cos t \cos\theta \\ 0 \end{pmatrix}. \quad (5.4.1)$$

Hence the tangent plane at a typical point on the equator of S^2 is the vector space spanned by the two vectors

$$x_1(\begin{pmatrix} 0 \\ \theta \end{pmatrix}) = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{and} \quad x_2(\begin{pmatrix} 0 \\ \theta \end{pmatrix}) = \begin{pmatrix} -\sin\theta \\ \cos\theta \\ 0 \end{pmatrix}. \quad \diamond$$

What is the analog for surfaces in $\mathbb{R}^3$ of the Frenet frame for curves $\{T, N, B\}$? Without a choice of a coordinate patch there are no distinguished tangent vectors at any point on a surface, only a tangent plane. With a chosen coordinate patch, however, we can select x_1 and x_2 as two of our three distinguished vectors; different choices of x yield different of x_1 and x_2. The vectors x_1 and x_2 are in general neither unit vectors nor orthogonal. We could replace x_1 and x_2 by an orthonormal basis for T_pM, but we would then lose useful

information about the behavior of x. We can, however, find a unit vector that is orthogonal to both x_1 and x_2 and hence orthogonal to T_pM.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ be a coordinate patch such that $p \in x(U)$. The **normal vector** function $n: U \to S^2$ with respect to x is defined by

$$n(\bar{p}) = \frac{x_1(\bar{p}) \times x_2(\bar{p})}{\|x_1(\bar{p}) \times x_2(\bar{p})\|},$$

for each $\bar{p} \in U$. $\Diamond$

We can think of n is a smooth function $U \to \mathbb{R}^3$. Also, we note that up to sign, n is independent of the choice of coordinate patch.

Example 5.4.4. (1) From Example 5.4.1 (1) it follows that the normal vector at each point of the plane $\mathbb{R}^2$ is $\begin{pmatrix} 0 \\ 0 \\ \pm 1 \end{pmatrix}$.

(2) Continuing Example 5.4.3, we see that at a typical point on the equator of S^2 we have

$$n\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right) = \frac{x_1\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right) \times x_2\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right)}{\left\|x_1\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right) \times x_2\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right)\right\|} = \begin{pmatrix} -\cos\theta \\ -\sin\theta \\ 0 \end{pmatrix} = -x\left(\begin{pmatrix} 0 \\ \theta \end{pmatrix}\right).$$

By the symmetry of S^2 we see that the normal vector at any point $p \in S^2$ is thus $-p$. Some other choices of coordinate patches would yield p as the normal vector rather than $-p$. $\Diamond$

Exercises

5.4.1. Describe the tangent plane and normal vector to the following surfaces at the specified points; does your answer make sense intuitively?

(i) M is the monge patch for the function $f(s, t) = st$, and p is the origin in $\mathbb{R}^3$.

(ii) M is the right helicoid parametrized in Section 5.3, and p is on the z-axis.

5.4.2*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, let $v, w \in T_pM$ be vectors, and let a, b be real numbers. Suppose that $x: U \to M$ is a coordinate

patch such that U contains the origin in $\mathbb{R}^2$ and $x(O_2) = p$ (we can always find such a coordinate patch). Suppose further that $c_v, c_w: (-\epsilon, \epsilon) \to x(U)$ are curves in M such that $c_v(0) = c_w(0) = p$, that $c_v'(0) = v$ and $c_w'(0) = w$. Show that the curve $c: (-\delta, \delta) \to M$ given by

$$c(t) = x(a(x^{-1} \circ c_v)(t) + b(x^{-1} \circ c_w)(t))$$

is well-defined for some small enough number $\delta > 0$, that $c(0) = p$ and $c'(0) = av + bw$.

5.4.3. Find the normal vector at any point on a monge patch.

5.4.4. Find the normal vector at any point on a surface of revolution.

5.4.5. Show that S^2 is orientable using coordinate patches.

5.4.6*. Let $M = F^{-1}(a)$ be a level surface for some smooth function $F: V \to \mathbb{R}$ (where $V \subset \mathbb{R}^3$ is an open set) and some regular value a of F. Show that the function $n: M \to \mathbb{R}^3$ defined by

$$n = \frac{(DF)^t}{\|(DF)^t\|}$$

is a normal vector field defined on all of M (where t denotes transpose).

5.4.7*. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ be a coordinate patch, with U assumed to be connected. Suppose there is a number k and a vector w in $\mathbb{R}^3$ such that the normal vector n satisfies the equation $n(\bar{p}) = kx(\bar{p}) + w$ for all $\bar{p} \in U$. Show that if $k = 0$ then $x(U)$ is contained in a plane, and if $k \neq 0$ then $x(U)$ is contained in a sphere of radius $\frac{1}{|k|}$.

5.5 First Fundamental Form

In classical Euclidean geometry it is necessary to measure things such as lengths and angles. From linear algebra we know that these two quantities can be computed using the standard inner product in Euclidean space. One of the truly important ideas in differential geometry is to use the inner product of vectors in the tangent plane at each point on a surface as the basic tool for geometric measurements in the surface. This idea is found at least as far back as the work of Gauss, and was brought to prominence by Riemann in his amazing lecture of

1854 entitled "Über die Hypothesen, welche der Geometrie zu Grunde liegen." See [SK3 vol. II] for a translation of Riemann's text, as well as very useful commentary on the work of both Gauss and Riemann.

The inner product is an example of a bilinear form. Details and examples concerning bilinear forms may be found in many linear algebra texts, for example [LA1, Chapter VIII]. We will restrict our attention to vector spaces over the real numbers.

Definition. Let V be a vector space over the real numbers. A **bilinear form** on V is a map $B: V \times V \to \mathbb{R}$ that is linear in each variable; that is, for all vectors $v, w, z \in V$ and all real number a and b we have

(1) $B(av + bw, z) = aB(v, z) + bB(w, z)$,

(2) $B(v, aw + bz) = aB(v, w) + bB(v, z)$.

A bilinear form B is **symmetric** if $B(v, w) = B(w, v)$ for all $v, w \in V$. $\Diamond$

The most familiar example of a bilinear form on a vector space is an inner product. As stated in the following definition, a bilinear form on a finite dimensional vector space gives rise to a matrix once a basis for the vector space is chosen.

Definition. Let V be a finite vector space over the real numbers, and let B be a bilinear form on V. Suppose $\{v_1, \ldots, v_n\}$ is a basis for V. The matrix for B with respect to this basis is defined to be

$$M = \begin{pmatrix} B(v_1, v_1) & \cdots & B(v_1, v_n) \\ \vdots & \ddots & \vdots \\ B(v_n, v_1) & \cdots & B(v_n, v_n) \end{pmatrix}. \quad \Diamond$$

Suppose that V, B and M are as in the above definition. If $x, y \in V$ are vectors, and X and Y are column vectors representing x and y with respect to the basis B, then

$$B(x, y) = X^t M Y,$$

where X^t is the transpose of X.

We now return to smooth surfaces, starting with the following simple but important definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $p \in M$ be a point. The **first fundamental form** of M at p is the function $I_p: T_pM \times T_pM \to \mathbb{R}$ defined

by

$$I_p(v, w) = \langle v, w \rangle,$$

where $\langle v, w \rangle$ is the standard inner product in $\mathbb{R}^3$ of the vectors $v, w \in T_pM \subset \mathbb{R}^3$. The **first fundamental form** of M is the collection, denoted I, of all functions I_p at all points $p \in M$. $\Diamond$

Though it may appear as if we are doing nothing other than renaming an already familiar concept, the use of the above definition will become more apparent in later sections. The first fundamental form is a geometric quantity, not a topological one, since it very much depends upon the way M sits in $\mathbb{R}^3$. As we will discuss more thoroughly in Section 5.9, if we were to deform M in $\mathbb{R}^3$ then the first fundamental form would, in general, change. On the other hand, the first fundamental form was defined with no reference to a particular choice of coordinate patch. Also, by using the properties of the standard inner product in Euclidean space, we see that I_p is a symmetric bilinear form for each $p \in M$.

To carry out computations it will be useful to express I in terms of a given coordinate patch. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. We now compute the matrix for the bilinear form I_p with respect to the basis $\{x_1, x_2\}$ of T_pM; we denote this matrix $\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$. Hence

$$g_{ij} = \langle x_i, x_j \rangle \tag{5.5.1}$$

for $i, j = 1, 2$, where both sides of this equation are being evaluated at $x^{-1}(p)$. We can think of the g_{ij} as smooth functions $U \to \mathbf{R}$, since the vectors x_1, x_2 are smooth functions with domain U. Due to their absolute centrality in the study of surfaces, we give the quantities g_{ij} a name.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. The functions $g_{ij}: U \to \mathbf{R}$ for $i, j = 1, 2$ defined by Equation 5.5.1 are called the **metric coefficients** of M with respect to the coordinate patch x. For convenience we use (g_{ij}) to denote the matrix $\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}$. $\Diamond$

The metric coefficients definitely depend upon the choice of coordinate patch used. Also, note that $g_{12} = g_{21}$ by the symmetry of the standard inner product on $\mathbb{R}^3$; thus (g_{ij}) is a symmetric matrix. Gauss used the symbols E, F and G to denote what we (and most modern books) call g_{11}, g_{12} and g_{22}

respectively; at times we will use Gauss' notation in certain formulas where the subscripts of the g_{ij} notation become too cumbersome.

Example 5.5.1. (1) The plane $\mathbb{R}^2 \subset \mathbb{R}^3$ can be parametrized as a monge patch by the function

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ 0 \end{pmatrix}.$$

Thus

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \quad \text{and} \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}.$$

We then see that

$$g_{11} = \left\langle \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \right\rangle = 1, \quad g_{22} = \left\langle \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle = 1$$

$$g_{12} = g_{21} = \left\langle \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \right\rangle = 0.$$

Thus we have

$$(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

(2) The saddle surface is parametrized by the monge patch

$$x\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} s \\ t \\ st \end{pmatrix}.$$

Thus

$$x_1 = \begin{pmatrix} 1 \\ 0 \\ t \end{pmatrix}, \quad \text{and} \quad x_2 = \begin{pmatrix} 0 \\ 1 \\ s \end{pmatrix},$$

and hence

$$(g_{ij}) = \begin{pmatrix} 1+t^2 & st \\ st & 1+s^2 \end{pmatrix}.$$

The following lemma will be useful for later calculations.

Lemma 5.5.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ be a coordinate patch. Then*

(i) $\det(g_{ij}) = \|x_1 \times x_2\|^2$;

(ii) $n = \dfrac{x_1 \times x_2}{\sqrt{\det(g_{ij})}}$.

Proof. (i). We compute

$$\begin{aligned}\det(g_{ij}) &= g_{11}\, g_{22} - (g_{12})^2 = \langle x_1, x_1\rangle\langle x_2, x_2\rangle - \langle x_1, x_2\rangle^2 \\ &= \|x_1\|^2\, \|x_2\|^2 - \|x_1\|^2\, \|x_2\|^2 \cos^2\varphi \\ &= \|x_1\|^2\, \|x_2\|^2 \sin^2\varphi \\ &= \|x_1 \times x_2\|^2,\end{aligned}$$

where φ is the angle between x_1 and x_2.

(ii). This follows immediately from the definition of n and part (i) of the lemma. □

It will be important to know how the matrix (g_{ij}) changes under a change of coordinate patches. Observe that a change of coordinate patches essentially yields a change of basis in each tangent plane.

Lemma 5.5.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ and $y: V \to M$ be coordinate patches with overlapping images. Let (g_{ij}) denote the metric coefficients with respect to x, and let $(\bar{g}_{ij})$ denote the metric coefficients with respect to y. If J denotes the Jacobian matrix of the change of coordinate function $\phi_{x,y}$, then at all points in the domain of $\phi_{x,y}$ we have*

$$(\bar{g}_{ij} \circ \phi_{x,y}) = (J^{-1})^t\, (g_{ij})\, J^{-1} \quad \textit{and} \quad \det(\bar{g}_{ij} \circ \phi_{x,y}) = \frac{\det(g_{ij})}{(\det J)^2}. \tag{5.5.2}$$

Proof. Recall that x_1 and x_2 are the column vectors of the matrix Dx, and similarly for y. It follows that

$$(g_{ij}) = (Dx)^t\, Dx \quad \text{and} \quad (\bar{g}_{ij}) = (Dy)^t\, Dy.$$

Using the chain rule and Equation 5.2.2, though dropping the notation for the restrictions to the sets $x^{-1}(x(U) \cap y(V))$ and $y^{-1}(x(U) \cap y(V))$), we deduce that $Dx = (Dy \circ \phi_{x,y})\, J$. We now have

$$\begin{aligned}(g_{ij}) &= \big[((Dy \circ \phi_{x,y})J\big]^t (Dy \circ \phi_{x,y})J = J^t (Dy \circ \phi_{x,y})^t (Dy \circ \phi_{x,y})J \\ &= J^t \big(\bar{g}_{ij} \circ \phi_{x,y}\big)J.\end{aligned}$$

The result now follows using standard properties of matrices and determinants. □

An example of the above lemma is given in Exercise 5.5.7.

Exercises

5.5.1. For a general monge patch, as parametrized in Section 5.3, show that

$$(g_{ij}) = \begin{pmatrix} 1+(f_1)^2 & f_1 f_2 \\ f_1 f_2 & 1+(f_2)^2 \end{pmatrix},$$

where f_1 and f_2 are the partial derivatives of f.

5.5.2. Find (g_{ij}) for the torus as parametrized in Section 5.3.

5.5.3*. The **catenoid** is the surface of revolution obtained by revolving the graph of $x = \cosh z$ (called a catenary) about the z-axis. An alternate parametrization of the catenary is given by the curve

$$c(t) = \begin{pmatrix} \sqrt{1+t^2} \\ \sinh^{-1}(t) \end{pmatrix},$$

(use hyperbolic trigonometry identities to verify that this parametrization works); we use this parametrization to construct our surface of revolution. Show that the metric coefficients for the catenoid are

$$(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1+t^2 \end{pmatrix}.$$

5.5.4*. For the surface of revolution parametrized by

$$y\left(\begin{pmatrix} t \\ \theta \end{pmatrix}\right) = \begin{pmatrix} t\sin\theta \\ t\cos\theta \\ \ln t \end{pmatrix},$$

show that

$$(g_{ij}) = \begin{pmatrix} \frac{1+t^2}{t^2)} & 0 \\ 0 & t^2 \end{pmatrix}.$$

5.5.5*. For a general surface of revolution, as parametrized in Section 5.3, show that

$$(g_{ij}) = \begin{pmatrix} (r')^2+(z')^2 & 0 \\ 0 & r^2 \end{pmatrix}.$$

5.5.6*. For a general rectifying developable surface, as parametrized in Section 5.3, show that

$$(g_{ij}) = \begin{pmatrix} 1 & 0 \\ 0 & 1 + s^2\tau^2(s) \end{pmatrix},$$

where $\tau(s)$ is the torsion of the curve c.

5.5.7. Consider the two coordinate charts for the sphere S^2 given by

$$x\left(\begin{pmatrix} t \\ \theta \end{pmatrix}\right) = \begin{pmatrix} R\cos t\cos\theta \\ R\cos t\sin\theta \\ R\sin t \end{pmatrix} \quad \text{and} \quad y\left(\begin{pmatrix} u \\ v \end{pmatrix}\right) = \begin{pmatrix} \sqrt{1-u^2-v^2} \\ u \\ v \end{pmatrix}.$$

(1) Find $\phi_{x,y}$ and its Jacobian matrix.

(2) Using the notation of Lemma 5.5.3, compute (g_{ij}), $\det(g_{ij})$, $(\bar{g}_{ij})$, $(\bar{g}_{ij} \circ \phi_{x,y})$, and $\det(\bar{g}_{ij} \circ \phi_{x,y})$.

(3) Verify that the conclusion of Lemma 5.5.3 holds for this example.

5.5.8*. For the parametrization of the right helicoid given in Section 5.3, show that

$$(g_{ij}) = \begin{pmatrix} t^2 + b^2 & 0 \\ 0 & 1 \end{pmatrix}.$$

5.5.9*. This exercise has three steps.

(1) Show that

$$\frac{\partial g_{ij}}{\partial u_k} = \frac{\partial}{\partial u_k}\langle x_i, x_j\rangle = \langle x_{ik}, x_j\rangle + \langle x_{jk}, x_i\rangle$$

for $i, j, k = 1, 2$.

(2) Show that

$$\langle x_{ij}, x_k\rangle = \frac{1}{2}\left\{\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right\}$$

for $i, j, k = 1, 2$.

(3) Show that

$$\begin{aligned} \langle x_{11}, x_1\rangle &= \frac{1}{2}\frac{\partial g_{11}}{\partial u_1}, \quad \langle x_{12}, x_1\rangle = \frac{1}{2}\frac{\partial g_{11}}{\partial u_2}, \\ \langle x_{22}, x_1\rangle &= \frac{\partial g_{12}}{\partial u_2} - \frac{1}{2}\frac{\partial g_{22}}{\partial u_1}, \quad \langle x_{11}, x_2\rangle = \frac{\partial g_{12}}{\partial u_1} - \frac{1}{2}\frac{\partial g_{11}}{\partial u_2} \\ \langle x_{12}, x_2\rangle &= \frac{1}{2}\frac{\partial g_{jk}}{\partial u_i}\frac{\partial g_{22}}{\partial u_1}, \quad \langle x_{22}, x_2\rangle = \frac{1}{2}\frac{\partial g_{22}}{\partial u_2}. \end{aligned} \tag{5.5.3}$$

5.6 Directional Derivatives — Coordinate-Free

A directional derivative on a surface is, intuitively, very much like the directional derivatives studied in multivariable Calculus, namely a derivative of a function in the direction of a given vector (not necessarily of unit length). This idea will be developed in stages in this section and the next, following the treatment in [BO, §VII].

Let $M \subset \mathbb{R}^3$ be a smooth surface and $p \in M$ be a point. Suppose we have a smooth function $f: M \to \mathbb{R}$ and a vector $v \in T_pM$. To find the derivative of f in the direction v, we use our definition of tangent vectors on surfaces. Let $c: (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ be a smooth curve such that $c(0) = p$ and $c'(0) = v$. The function $f \circ c: (-\epsilon, \epsilon) \to \mathbb{R}$ is smooth by Exercise 5.2.8, so we can then compute $(f \circ c)'(0)$, which would be a good candidate for the derivative of f in the direction of v, except that there are many possible curves c. The following somewhat surprising lemma shows that there is in fact no ambiguity here.

Lemma 5.6.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $f: M \to \mathbb{R}$ be a smooth function, let $p \in M$ be a point and let $v \in T_pM$ be a vector. If $c^1, c^2: (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ are curves such that $c^1(0) = c^2(0) = p$ and $(c^1)'(0) = (c^2)'(0) = v$, then*

$$(f \circ c^1)'(0) = (f \circ c^2)'(0).$$

Proof. Let $x: U \to M$ be a coordinate patch such that $p \in x(U)$. We may assume without loss of generality that ϵ is chosen small enough so that the images of c^1 and c^2 are contained in $x(U)$. For each $i = 1, 2$ let $c_1^i, c_2^i: (a, b) \to \mathbb{R}$ be the coordinate functions of c^i with respect to x. Using Exercise 5.2.6 we have $(c^i)'(t) = (c_1^i)'(t)x_1 + (c_2^i)'(t)x_2$ for $i = 1, 2$, where x_1 and x_2 are evaluated at $x^{-1} \circ c(t)$. Since $(c^1)'(0) = (c^2)'(0)$, and since the vectors $\{x_1, x_2\}$ form a basis for the tangent plane, it follows that $(c_k^1)'(0) = (c_k^2)'(0)$ for $k = 1, 2$. Note further that

$$x^{-1} \circ c^i(t) = \begin{pmatrix} c_1^i(t) \\ c_2^i(t) \end{pmatrix}$$

for each $i = 1, 2$; taking the derivative at $t = 0$ of both sides of this equation and using our previous observations implies $(x^{-1} \circ c^1)'(0) = (x^{-1} \circ c^2)'(0)$.

Next, observe that $f \circ x$ is Euclidean smooth, and by Proposition 5.2.5 (ii) we know that $x^{-1} \circ c^i$ is Euclidean smooth as well. Using the chain rule we

now have

$$\begin{aligned}(f \circ c^1)'(0) &= [(f \circ x) \circ (x^{-1} \circ c^1)]'(0) \\ &= D(f \circ x)(x^{-1} \circ c^1(0)) \cdot (x^{-1} \circ c^1)'(0) \\ &= D(f \circ x)(x^{-1} \circ c^2(0)) \cdot (x^{-1} \circ c^2)'(0) = (f \circ c^2)'(0),\end{aligned}$$

where $D(f \circ x)$ denotes the Jacobian matrix, and $\cdot$ is matrix multiplication. □

We can now make the following definition safely.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, $f: M \to \mathbb{R}$ be a smooth function, let $p \in M$ be a point and let $v \in T_pM$ be a vector. If $c: (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ is a curve such that $c(0) = p$ and $c'(0) = v$, the **directional derivative** of f at p in the direction v is the number $\widetilde{\nabla}_v f$ defined by

$$\widetilde{\nabla}_v f = (f \circ c)'(0). \quad \Diamond$$

The directional derivative satisfies a number of standard properties.

Lemma 5.6.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $f, g: M \to \mathbb{R}$ be smooth functions, let $p \in M$ be a point, let $v, w \in T_pM$ be vectors, and let a, b be real numbers. Then*

(i) $\widetilde{\nabla}_{av+bw} f = a\widetilde{\nabla}_v f + b\widetilde{\nabla}_w f.$
(ii) $\widetilde{\nabla}_v(f+g) = \widetilde{\nabla}_v f + \widetilde{\nabla}_v g.$
(iii) $\widetilde{\nabla}_v fg = \big(\widetilde{\nabla}_v f\big)g(p) + f(p)\big(\widetilde{\nabla}_v g\big).$

Proof. Let $c_v, c_w: (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ be curves such that $c_v(0) = c_w(0) = p$, that $c_v'(0) = v$ and $c_w'(0) = w$.

(i). Let $x: U \to M$ be a coordinate patch such that U contains O_2 and $x(O_2) = p$ (such a coordinate patch can always be found). We may assume that ϵ is small enough so that the images of c_v and c_w are in $x(U)$. Let the curve $c: (-\delta, \delta) \to M$ be as in Exercise 5.4.2, so that $c(0) = p$ and $c'(0) = av + bw$. Observe that $x^{-1} \circ c_v(0) = x^{-1} \circ c_w(0) = x^{-1} \circ c(0) = O_2$. Then, proceeding

similarly to the argument in the proof of Lemma 5.6.1, we have

$$\begin{aligned}
\widetilde{\nabla}_{av+bw} f &= (f \circ c)'(0) = [(f \circ x) \circ (x^{-1} \circ c)]'(0) \\
&= D(f \circ x)(x^{-1} \circ c(0)) \cdot (x^{-1} \circ c)'(0) \\
&= D(f \circ x)(O_2) \cdot (x^{-1} \circ x(a(x^{-1} \circ c_v) + b(x^{-1} \circ c_w)))'(0) \\
&= D(f \circ x)(O_2)\ (a(x^{-1} \circ c_v) + b(x^{-1} \circ c_w))'(0) \\
&= aD(f \circ x)(O_2) \cdot (x^{-1} \circ c_v)'(0) \ + \ bD(f \circ x)(O_2) \cdot (x^{-1} \circ c_w)'(0) \\
&= aD(f \circ x)(x^{-1} \circ c_v(0)) \cdot (x^{-1} \circ c_v)'(0) \\
&\qquad + \ bD(f \circ x)(x^{-1} \circ c_w(0)) \cdot (x^{-1} \circ c_w)'(0) \\
&= \cdots = a\widetilde{\nabla}_v f + b\widetilde{\nabla}_w f.
\end{aligned}$$

(ii) & (iii). Exercise 5.6.2. □

Example 5.6.3. Let $f: S^2 \to \mathbb{R}$ be given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = xy + xz$, let $p = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \in S^2$ and let $v = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in T_p S^2$. To find $\widetilde{\nabla}_v f$ we choose the curve $c: (-\pi, \pi) \to S^2$ given by $c(t) = \begin{pmatrix} \sin t \\ 0 \\ \cos t \end{pmatrix}$, which has $c(0) = p$ and $c'(0) = v$. Thus

$$\widetilde{\nabla}_v f = (f \circ c)'(0) = (\sin t \cos t)'(0) = 1. \quad \Diamond$$

Next we consider vector fields on surfaces (not necessarily tangent vector fields). As before, we will think of all vectors in $\mathbb{R}^3$ as being translated so that they start at the origin.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. A **smooth vector field** on M is a smooth function $Z: M \to \mathbb{R}^3$. A smooth vector field Z is a **tangent vector field** if $Z(p) \in T_p M$ for each $p \in M$. $\Diamond$

The function Z in the above definition can be written in terms of components as

$$Z(p) = \begin{pmatrix} z_1(p) \\ z_2(p) \\ z_3(p) \end{pmatrix},$$

where $z_1, z_2, z_3: M \to \mathbb{R}$ are smooth functions.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $Z: M \to \mathbb{R}^3$ be a smooth vector field, let $p \in M$ be a point and let $v \in T_pM$ be a vector. The **directional derivative** of Z in the direction v is the vector $\widetilde{\nabla}_v Z$ defined by

$$\widetilde{\nabla}_v Z = \begin{pmatrix} \widetilde{\nabla}_v z_1 \\ \widetilde{\nabla}_v z_2 \\ \widetilde{\nabla}_v z_3 \end{pmatrix}. \quad \Diamond$$

Using the notation in the above definition, we note that if $c: (-\epsilon, \epsilon) \to M$ is a curve such that $c(0) = p$ and $c'(0) = v$, then

$$\widetilde{\nabla}_v Z = (Z \circ c)'(0).$$

Once again, the directional derivative for vector fields satisfies a number of standard properties.

Lemma 5.6.4. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $Z, Y: M \to \mathbb{R}^3$ be smooth vector fields, let $f: M \to \mathbb{R}$ be a smooth function, let $p \in M$ be a point, let $v, w \in T_pM$ be vectors, and let a, b be real numbers. Then*

(i) $\widetilde{\nabla}_{av+bw} Z = a\widetilde{\nabla}_v Z + b\widetilde{\nabla}_w Z.$
(ii) $\widetilde{\nabla}_v (Z + Y) = \widetilde{\nabla}_v Z + \widetilde{\nabla}_v Y.$
(iii) $\widetilde{\nabla}_v fZ = \big(\widetilde{\nabla}_v f\big) Z(p) + f(p)\big(\widetilde{\nabla}_v Z\big).$
(iv) $\widetilde{\nabla}_v \langle Z, Y\rangle = \langle \widetilde{\nabla}_v Z, Y(p)\rangle + \langle Z(p), \widetilde{\nabla}_v Y\rangle.$

Proof. Exercise 5.6.3. □

Example 5.6.5. Let $Z: S^1 \times \mathbb{R} \to \mathbb{R}^3$ be given by $Z\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \begin{pmatrix} -y \\ x \\ 0 \end{pmatrix}$, let $p = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in S^1 \times \mathbb{R}$ and let $v = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in T_pS^1 \times \mathbb{R}$. To find $\widetilde{\nabla}_v Z$ we choose the curve $c: (-\pi, \pi) \to S^1 \times \mathbb{R}$ given by $c(t) = \begin{pmatrix} \cos t \\ \sin t \\ 0 \end{pmatrix}$, which has $c(0) = p$ and $c'(0) = v$. Thus

$$\widetilde{\nabla}_v Z = (Z \circ c)'(0) = \begin{pmatrix} -\sin t \\ \cos t \\ 0 \end{pmatrix}'(0) = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}. \quad \Diamond$$

For the next-to-last stage in this line of development, suppose that a vector field Z on M is actually a tangent vector field. If $p \in M$ and $v \in T_pM$ we can compute $\widetilde{\nabla}_v Z$ considering Z as a vector field (tangent or not). There is

no reason to expect that $\widetilde{\nabla}_v Z \in T_pM$, even though both v and Z are tangent to M. For example, the vector field Z in Example 5.6.5 is actually a tangent vector field, whereas $\widetilde{\nabla}_v Z = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}$ is not a tangent vector at $p = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Since tangent vectors are particularly useful, we remedy this situation essentially by brute force (the need for this remedy will become apparent in Section 6.2). We use the following notation. Let H be a plane in $\mathbb{R}^3$ containing the origin; the map $\Pi_H: \mathbb{R}^3 \to H$ will denote orthogonal projection of $\mathbb{R}^3$ onto H.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $Z: M \to \mathbb{R}^3$ be a tangent vector field, let $p \in M$ be a point and let $v \in T_pM$ be a vector. The **covariant derivative** of Z with respect to v is the vector $\nabla_v Z$ defined by

$$\nabla_v Z = \Pi_{T_pM}\left(\widetilde{\nabla}_v Z\right). \quad \Diamond$$

By definition $\nabla_v Z \in T_pM$. All the expected properties hold.

Lemma 5.6.6. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $Z, Y: M \to \mathbb{R}^3$ be smooth tangent vector fields, let $f: M \to \mathbb{R}$ be a smooth function, let $p \in M$ be a point, let $v, w \in T_pM$ be vectors, and let a, b be real numbers. Then*

(i) *The function $T_pM \to T_pM$ given by $v \mapsto \nabla_v Z$ is a linear map.*
(ii) $\nabla_v(Z + Y) = \nabla_v Z + \nabla_v Y$.
(iii) $\nabla_v fZ = (\nabla_v f)Z(p) + f(p)(\nabla_v Z)$.
(iv) $\widetilde{\nabla}_v\langle Z, Y\rangle = \langle \nabla_v Z, Y(p)\rangle + \langle Z(p), \nabla_v Y\rangle$.

Proof. (i). Using Lemma 5.6.4 (i), as well as the linearity of Π_{T_pM}, we have

$$\begin{aligned}\nabla_{av+bw} Z &= \Pi_{T_pM}\left(\widetilde{\nabla}_{av+bw} Z\right) = \Pi_{T_pM}\left(a\widetilde{\nabla}_v Z + b\widetilde{\nabla}_w Z\right) \\ &= a\Pi_{T_pM}\left(\widetilde{\nabla}_v Z\right) + b\Pi_{T_pM}\left(\widetilde{\nabla}_w Z\right) = a\nabla_v Z + b\nabla_w Z.\end{aligned}$$

This proves what we are trying to show.

(ii) and (iii). These are similar to part (i).

(iv). By Lemma 5.6.4 (iv) we have $\widetilde{\nabla}_v\langle Z, Y\rangle = \langle \widetilde{\nabla}_v Z, Y(p)\rangle + \langle Z(p), \widetilde{\nabla}_v Y\rangle$. By writing the vectors $\widetilde{\nabla}_v Z$ and Y in terms of the basis $\{x_1, x_2, n\}$ of $\mathbb{R}^3$, and by observing that Π_{T_pM} has the effect of removing from any vector its component in the n direction, it can be shown that $\langle \widetilde{\nabla}_v Z, Y(p)\rangle = \langle \nabla_v Z, Y(p)\rangle$, and similarly with the roles of Z and Y reversed. $\square$

Example 5.6.7. Continuing Example 5.6.5, we see that $\widetilde{\nabla}_v Z = \begin{pmatrix} -1 \\ 0 \\ 0 \end{pmatrix}$ is perpendicular to $T_p S^1 \times \mathbb{R}$ at the given point p. Hence $\nabla_v Z$ is the zero vector. ◊

We now turn to directional derivatives of vector fields defined along curves in a surface. Since the image of a curve in a surface may intersect itself, we take the domain of the vector field to be the domain of the curve, namely the interval (a, b), rather than on the image of the curve.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $c: (a, b) \to M \subset \mathbb{R}^3$ be a smooth curve. A smooth **vector field** along the curve c is a smooth function $Z: (a, b) \to \mathbb{R}^3$. The vector field Z is **tangent** to M along c if $Z(t) \in T_{c(t)}M$ for all $t \in (a, b)$. ◊

The definition of a vector field along a curve looks just like the definition of a curve itself; it's simply a matter of how we view things — as vectors or as endpoints of vectors. We can define the derivative $\frac{dZ}{dt}$ in the usual way, taking the derivative of each component of Z. All the usual rules for derivatives hold for $\frac{dZ}{dt}$. As before, even if Z is a tangent vector field along a curve c in M, there is no reason to expect that $\frac{dZ}{dt}$ will be tangent to M. We remedy this problem as before.

Definition. Let $c: (a, b) \to M \subset \mathbb{R}^3$ be a smooth curve, and let $Z: (a, b) \to \mathbb{R}^3$ be a smooth vector field along c that is tangent to M along c. The **covariant derivative** of Z along c is the vector field $\frac{DZ}{dt}$ along c defined by

$$\frac{DZ}{dt} = \Pi_{T_{c(t)}M}\left(\frac{dZ}{dt}\right). \quad \diamond$$

By definition $\frac{DZ}{dt} \in T_{c(t)}M$ for all $t \in (a, b)$. All the expected properties hold.

Lemma 5.6.8. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $c: (a, b) \to M$ be a smooth curve, let $Z, Y: (a, b) \to \mathbb{R}^3$ be smooth vector fields along c that are tangent to M along c, and let $f: (a, b) \to \mathbb{R}$ be a smooth function. Then*

(i) $\dfrac{D(Z+Y)}{dt} = \dfrac{DZ}{dt} + \dfrac{DY}{dt}$.

(ii) $\dfrac{D(fZ)}{dt} = \dfrac{df}{dt}Z(t) + f(t)\dfrac{DZ}{dt}$.

(iii) $\dfrac{d}{dt}\langle Z, Y\rangle = \langle \dfrac{DZ}{dt}, Y\rangle + \langle Z, \dfrac{DY}{dt}\rangle$.

Proof. Exercise 5.6.5. □

What is the relation of the constructions $\nabla_v Z$ and $\frac{DZ}{dt}$? We can ask this question in two ways. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $p \in M$. First, suppose we are given a smooth tangent vector field $Z: M \to \mathbb{R}^3$, and suppose we are given $v \in T_pM$. Can we express $\nabla_v Z$ as $\frac{DY}{dt}$ for some smooth tangent vector field $Y: (a, b) \to \mathbb{R}^3$ along some smooth curve c in M? Conversely, suppose we are given a smooth tangent vector field $Y: (a, b) \to \mathbb{R}^3$ along some smooth curve c in M. Can we express $\frac{DY}{dt}$ as $\nabla_v Z$ for some appropriate Z and v? The following proposition, which will be of use later on, resolves these questions. Note that for a given smooth tangent vector field $Y: (a, b) \to \mathbb{R}^3$ along some smooth curve c in M, we cannot always extend Y to a smooth tangent vector field on M, and if we can extend Y, there might be more than one way of doing so. In the proposition, it will suffice to have an extension of Y to some open subset of M (rather than all M), since we can take directional derivatives on any open subset of M.

Proposition 5.6.9. *Let $M \subset \mathbb{R}^3$ be a smooth surface.*

(i) *Let $Z: M \to \mathbb{R}^3$ be a smooth tangent vector field on M, let $p \in M$ be a point, and let $v \in T_pM$ be a vector. If $c: (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ is a curve such that $c(0) = p$ and $c'(0) = v$, then*

$$\nabla_v Z = \frac{D(Z \circ c)}{dt}(0).$$

(ii) *Let $c: (a, b) \to M$ be a smooth curve and let $Y: (a, b) \to \mathbb{R}^3$ be a smooth vector field along c that is tangent to M along c. If $\widetilde{Y}: V \to \mathbb{R}^3$ is a smooth tangent vector field for some open subset $V \subset M$ containing the image of c, such that $\widetilde{Y} \circ c(t) = Y(t)$ for all $t \in (a, b)$, then*

$$\frac{DY}{dt} = \nabla_{c'(t)} \widetilde{Y}$$

at each $t \in (a, b)$. In particular $\nabla_{c'(t)} \widetilde{Y}$ does not depend upon the choice of extension $\widetilde{Y}$ of Y.

Proof. (i). This simply requires tracing through the definitions.

(ii). Note that $\widetilde{Y}$ is a smooth vector field on an open subset of M. For each $t \in (a, b)$ the point $c(t)$ is in the domain of $\widetilde{Y}$, and the vector $c'(t)$ is in $T_{c(t)}M$ for

$t \in (a, b)$. The curve c is thus a curve such that $c(t) = c(t)$ and $c'(t) = c'(t)$. Now use part (i). □

Exercises

5.6.1. Let $f: S^1 \times \mathbb{R} \to \mathbb{R}$ be given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = xz^2 + 3y$, let $p = \begin{pmatrix} 1 \\ 1 \\ 0 \end{pmatrix} \in S^2$, let $v = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix} \in T_p S^1 \times \mathbb{R}$ and let $w = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \in T_p S^1 \times \mathbb{R}$. Find $\widetilde{\nabla}_v f$ and $\widetilde{\nabla}_w f$.

5.6.2*. Prove Lemma 5.6.2 (ii), (iii).

5.6.3*. Prove Lemma 5.6.4.

5.6.4*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $W \subset \mathbb{R}^3$ be an open set containing M, let $p \in M$ be a point and let $v \in T_p M$ be a vector. Suppose that v can be written as $v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \in \mathbb{R}^3$.

(i) If $f: W \to \mathbb{R}$ is a smooth function, show that $\widetilde{\nabla}_v f = Df(p)\, v$.

(ii) If $Z: W \to \mathbb{R}^3$ is a smooth vector field, show that $\widetilde{\nabla}_v Z = DZ(p)\, v$.

5.6.5*. Prove Lemma 5.6.8.

5.6.6. Let T^2 be the torus as discussed in Section 5.3. Let $Z: T^2 \to \mathbb{R}^3$ be given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = \begin{pmatrix} z \\ y \\ x+y \end{pmatrix}$, let $p = \begin{pmatrix} R+r \\ 0 \\ 0 \end{pmatrix} \in T^2$ and let $v = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \in T_p T^2$. Find $\widetilde{\nabla}_v Z$ and $\nabla_v Z$.

5.7 Directional Derivatives — Coordinates

We now develop an expression for the covariant derivative (both $\nabla_v Z$ and $\frac{DZ}{dt}$) in terms of a coordinate patch. Although in the previous section we discussed functions and vector fields defined on all of M, we could apply the directional derivative to functions, vector fields and curves defined only on $x(U)$, which is an open subset of M. For the rest of this section let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, let $x: U \to M$ be a coordinate patch such that $p \in x(U)$, let $v \in T_p M$ be a vector, let $c: (a, b) \to x(U)$ be a smooth

curve, let $Z: x(U) \to \mathbb{R}^3$ be a smooth tangent vector field on $x(U)$, and let $Y: (a, b) \to \mathbb{R}^3$ be a smooth vector field along c that is tangent to M along c. We let $\bar{p} = x^{-1}(p)$, let $\bar{c} = x^{-1} \circ c$, and let c_1, c_2 be the coordinate functions of c with respect to x. It is evident that there are numbers v^1 and v^2 such that $v = v^1 x_1(\bar{p}) + v^2 x_2(\bar{p})$. It is shown in Exercise 5.7.1 (i) that there are unique smooth coordinate functions $Z^1, Z^2: U \to \mathbb{R}$ such that $Z \circ x(\bar{q}) = Z^1(\bar{q}) x_1(\bar{q}) + Z^2(\bar{q}) x_2(\bar{q})$ for $\bar{q} \in U$. Similarly, it is shown in Exercise 5.7.1 (ii) that there are unique smooth coordinate functions $Y^1, Y^2: (a, b) \to \mathbb{R}$ such that $Y(t) = Y^1(t) x_1(\bar{c}(t)) + Y^2(t) x_2(\bar{c}(t))$ for all $t \in (a, b)$. Finally, to make effective use of the summation notation we will at times denote the variables in the function x by u_1 and u_2 instead of s and t.

Our goal is to express $\nabla_v Z$ in terms of Z^1, Z^2, v^1, v^2 and x, and similarly for $\frac{DY}{dt}$. We start with the following technicality, since the vector field Z is defined on $x(U)$, whereas when things are expressed in terms of a coordinate patch (for example, the metric coefficients) they are defined on the set U.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. A **tangent vector field** on M with respect to x is a smooth function $X: U \to \mathbb{R}^3$ such that $X(\bar{p}) \in T_{x(\bar{p})}M$ for all $\bar{p} \in U$. If $v \in T_{x(\bar{p})}M$ is a vector, then the **covariant derivative** of X in the direction v is the vector $\overline{\nabla}_v X$ defined by $\overline{\nabla}_v X = \nabla_v (X \circ x^{-1})$. $\Diamond$

It is straightforward to see that $\overline{\nabla}_v$ has all the usual properties, analogous to Lemma 5.6.6. If the above definition is viewed backward then we see that for the tangent vector field Z as above, we have $\nabla_v Z = \overline{\nabla}_v (Z \circ x)$.

We can think of the partial derivatives x_1 and x_2 as tangent vector fields on M with respect to x. Our first step is to compute the covariant derivative $\overline{\nabla}_{x_i(\bar{p})} x_j$ for each $\bar{p} \in U$, where $i, j = 1, 2$. This covariant derivative is itself a tangent vector in $T_{x(\bar{p})}M$, so that it is uniquely expressible in terms of the basis $\{x_1(\bar{p}), x_2(\bar{p})\}$ of $T_{x(\bar{p})}M$. Hence there are unique numbers $\Gamma_{ij}^1(\bar{p})$ and $\Gamma_{ij}^2(\bar{p})$ such that

$$\overline{\nabla}_{x_i(\bar{p})} x_j = \Gamma_{ij}^1(\bar{p})\, x_1(\bar{p}) + \Gamma_{ij}^2(\bar{p})\, x_2(\bar{p}). \tag{5.7.1}$$

Because the above equation works at each $\bar{p} \in U$, we obtain a real-valued function $\Gamma_{ij}^k: U \to \mathbb{R}$ for each $i, j, k = 1, 2$, eight functions in all. For ease of notation we will often drop the arguments in the above equation and write

$$\overline{\nabla}_{x_i} x_j = \Gamma_{ij}^1\, x_1 + \Gamma_{ij}^2\, x_2 = \sum_{k=1}^{2} \Gamma_{ij}^k\, x_k. \tag{5.7.2}$$

Though we have dropped the arguments in expressions of the form $\overline{\nabla}_{x_i} x_j$, it is important to keep in mind that we are really using the symbols x_i in two different ways: The x_i in the subscript of $\overline{\nabla}$ is shorthand for a single vector $x_i(\tilde{p})$, whereas the x_j being operated on by the $\overline{\nabla}$ is a vector field.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. The functions $\Gamma_{ij}^k: U \to \mathbb{R}$ for $i, j, k = 1, 2$ defined by Equation 5.7.1 are called the **Christoffel symbols** for the coordinate patch x. $\Diamond$

The following lemma simplifies things a bit.

Lemma 5.7.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch.*

(i) *For all $i, j = 1, 2$ and at all points $\tilde{p} \in U$ we have*

$$\overline{\nabla}_{x_i(\tilde{p})} x_j = \overline{\nabla}_{x_j(\tilde{p})} x_i.$$

(ii) *For all $i, j, k = 1, 2$ we have*

$$\Gamma_{ij}^k = \Gamma_{ji}^k.$$

Proof. (i). Let $p \in x(U)$ be a point. Using the definition of $\overline{\nabla}$, Exercise 5.7.2 and the equality of mixed partial derivatives for smooth functions, we compute

$$\begin{aligned}
\overline{\nabla}_{x_i(\tilde{p})} x_j &= \nabla_{x_i(\tilde{p})}(x_j \circ x^{-1}) = \Pi_{T_pM}\left(\widetilde{\nabla}_{x_i(\tilde{p})}(x_j \circ x^{-1})\right) \\
&= \Pi_{T_pM}\left(\frac{\partial x_j}{\partial u_i}\right) = \Pi_{T_pM}\left(\frac{\partial^2 x}{\partial u_i\, \partial u_j}\right) \\
&= \Pi_{T_pM}\left(\frac{\partial^2 x}{\partial u_j\, \partial u_i}\right) = \cdots = \overline{\nabla}_{x_j(\tilde{p})} x_i.
\end{aligned}$$

(ii). By the definition of the Christoffel symbols and part (i) of this lemma we have

$$\sum_{k=1}^{2} \Gamma_{ij}^k x_k = \overline{\nabla}_{x_i} x_j = \overline{\nabla}_{x_j} x_i = \sum_{k=1}^{2} \Gamma_{ji}^k x_k,$$

where for convenience we drop the arguments. Since the vectors $\{x_1, x_2\}$ form a basis for the tangent plane their coefficients must be equal in the first and last terms of this equation. $\square$

The next lemma tells us three things about Christoffel symbols: it shows that they only depend upon the metric coefficients, and not on the choice of

coordinate patch (and are thus considered "intrinsic," to be discussed in the smooth case in §5.9); it shows that they are smooth functions of u_1 and u_2 (since the g_{ij} are smooth); it gives us a convenient way to compute them.

Lemma 5.7.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. For all $i, j = 1, 2$, we have*

$$\begin{pmatrix} \Gamma_{ij}^1 \\ \Gamma_{ij}^2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial g_{j1}}{\partial u_i} + \frac{\partial g_{i1}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_1} \\ \frac{\partial g_{j2}}{\partial u_i} + \frac{\partial g_{i2}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_2} \end{pmatrix}.$$

Proof. Let $p \in x(U)$ be a point. Using Exercise 5.7.2 and Lemma 5.6.6 we compute

$$\begin{aligned} \frac{\partial g_{ij}}{\partial u_k}(\bar{p}) &= \frac{\partial g_{ij} \circ x^{-1} \circ x}{\partial u_k}(\bar{p}) = \tilde{\nabla}_{x_k(\bar{p})}(g_{ij} \circ x^{-1}) \\ &= \tilde{\nabla}_{x_k(\bar{p})}\langle x_i \circ x^{-1}, x_j \circ x^{-1}\rangle \\ &= \langle \nabla_{x_k(\bar{p})} x_i \circ x^{-1}, x_j \circ x^{-1}(p)\rangle + \langle x_i \circ x^{-1}(p), \nabla_{x_k(\bar{p})} x_j \circ x^{-1}\rangle \\ &= \langle \overline{\nabla}_{x_k(\bar{p})} x_i, x_j(\bar{p})\rangle + \langle x_i(\bar{p}), \bar{\nabla}_{x_k(\bar{p})} x_j\rangle. \end{aligned}$$

By permuting the three subscripts and dropping the arguments in the functions we obtain two other equations:

$$\begin{aligned} \frac{\partial g_{jk}}{\partial u_i} &= \langle \overline{\nabla}_{x_i} x_j, x_k\rangle + \langle x_j, \overline{\nabla}_{x_i} x_k\rangle, \\ \frac{\partial g_{ik}}{\partial u_j} &= \langle \overline{\nabla}_{x_j} x_i, x_k\rangle + \langle x_i, \overline{\nabla}_{x_j} x_k\rangle. \end{aligned}$$

Using Lemma 5.7.1 (i) we now solve for $\langle \overline{\nabla}_{x_i} x_j, x_k\rangle$ by adding the second and third equations and subtracting the first, obtaining

$$\langle \overline{\nabla}_{x_i} x_j, x_k\rangle = \frac{1}{2}\left(\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right).$$

On the other hand, by definition $\overline{\nabla}_{x_i} x_j = \sum_{l=1}^{2} \Gamma_{ij}^l x_l$. Taking the inner product with x_k yields

$$\langle \overline{\nabla}_{x_i} x_j, x_k\rangle = \sum_{l=1}^{2} \Gamma_{ij}^l \langle x_l, x_k\rangle = \sum_{l=1}^{2} \Gamma_{ij}^l g_{lk}.$$

Combining these last two equations, we obtain

$$\sum_{l=1}^{2} \Gamma_{ij}^{l} g_{lk} = \frac{1}{2}\left(\frac{\partial g_{jk}}{\partial u_i} + \frac{\partial g_{ik}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_k}\right). \tag{5.7.3}$$

For each $i, j = 1, 2$ Equation 5.7.3 can be rewritten in matrix form as

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} \Gamma_{ij}^{1} \\ \Gamma_{ij}^{2} \end{pmatrix} = \frac{1}{2} \begin{pmatrix} A_{ij}^{1} \\ A_{ij}^{2} \end{pmatrix}. \tag{5.7.4}$$

Multiplying on the left by the matrix $(g_{ij})^{-1}$ yields the desired result. □

For ease of computation we write out the conclusion of the above lemma for each possible combination of i and j, noting that $\Gamma_{ij}^{k} = \Gamma_{ji}^{k}$.

$$\begin{aligned}
\begin{pmatrix} \Gamma_{11}^{1} \\ \Gamma_{11}^{2} \end{pmatrix} &= \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2}\frac{\partial g_{11}}{\partial u_1} \\ \frac{\partial g_{12}}{\partial u_1} - \frac{1}{2}\frac{\partial g_{11}}{\partial u_2} \end{pmatrix} \\
\begin{pmatrix} \Gamma_{12}^{1} \\ \Gamma_{12}^{2} \end{pmatrix} &= \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} \frac{1}{2}\frac{\partial g_{11}}{\partial u_2} \\ \frac{1}{2}\frac{\partial g_{22}}{\partial u_1} \end{pmatrix} \\
\begin{pmatrix} \Gamma_{22}^{1} \\ \Gamma_{22}^{2} \end{pmatrix} &= \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial g_{12}}{\partial u_2} - \frac{1}{2}\frac{\partial g_{22}}{\partial u_1} \\ \frac{1}{2}\frac{\partial g_{22}}{\partial u_2} \end{pmatrix}.
\end{aligned} \tag{5.7.5}$$

Example 5.7.3. (1) We continue Example 5.5.1 (1). Since the g_{ij} are all constant, their partial derivatives are all zero. Hence Lemma 5.7.2 implies that $\Gamma_{ij}^{k} = 0$ for all $i, j, k = 1, 2$.

(2) We continue Exercise 5.5.8, though we now substitute u_1 and u_2 for s and t respectively. We see that $\frac{\partial g_{11}}{\partial u_2} = 2u_2$, and all the other partials of the g_{ij}'s are zero. Plugging these values into Equation 5.7.5, we obtain

$$\begin{aligned}
\begin{pmatrix} \Gamma_{11}^{1} \\ \Gamma_{11}^{2} \end{pmatrix} &= \frac{1}{(u_2)^2 + b^2} \begin{pmatrix} 1 & 0 \\ 0 & (u_2)^2 + b^2 \end{pmatrix} \begin{pmatrix} 0 \\ -\frac{1}{2}2u_2 \end{pmatrix} = \begin{pmatrix} 0 \\ -u_2 \end{pmatrix} \\
\begin{pmatrix} \Gamma_{12}^{1} \\ \Gamma_{12}^{2} \end{pmatrix} &= \frac{1}{(u_2)^2 + b^2} \begin{pmatrix} 1 & 0 \\ 0 & (u_2)^2 + b^2 \end{pmatrix} \begin{pmatrix} \frac{1}{2}2u_2 \\ 0 \end{pmatrix} = \begin{pmatrix} \frac{u_2}{(u_2)^2 + b^2} \\ 0 \end{pmatrix} \\
\begin{pmatrix} \Gamma_{22}^{1} \\ \Gamma_{22}^{2} \end{pmatrix} &= \frac{1}{(u_2)^2 + b^2} \begin{pmatrix} 1 & 0 \\ 0 & (u_2)^2 + b^2 \end{pmatrix} \begin{pmatrix} 0 \\ 0 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \end{pmatrix}.
\end{aligned}$$ ◊

We now take our first step toward computing directional derivatives. We make use of the notation mentioned at the start of this section.

Lemma 5.7.4. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $c: (a, b) \to x(U)$ be a smooth curve. Then*

$$\frac{D(x_i \circ \bar{c})}{dt} = \sum_{k=1}^{2} \sum_{j=1}^{2} \Gamma_{ji}^{k}(\bar{c}(t))\, c_j'(t)\, x_k(\bar{c}(t)).$$

Proof. Since the tangent vector field $x_i \circ x^{-1}$ is defined on the open subset $x(U)$ of M, and since $(x_i \circ x^{-1}) \circ c = x_i \circ \bar{c}$, we can use Proposition 5.6.9 (ii) and Exercise 5.2.6 to see that

$$\begin{aligned}\frac{D(x_i \circ \bar{c})}{dt} &= \nabla_{c'(t)}(x_i \circ x^{-1}) = \overline{\nabla}_{c'(t)} x_i = \overline{\nabla}_{[c_1'(t)x_1(\bar{c}(t)) + c_2'(t)x_2(\bar{c}(t))]} x_i \\ &= c_1'(t) \overline{\nabla}_{x_1(\bar{c}(t))} x_i + c_2'(t) \overline{\nabla}_{x_2(\bar{c}(t))} x_i \\ &= c_1'(t) \sum_{k=1}^{2} \Gamma_{1i}^{k}(\bar{c}(t))\, x_k(\bar{c}(t)) + c_2'(t) \sum_{k=1}^{2} \Gamma_{2i}^{k}(\bar{c}(t))\, x_k(\bar{c}(t)).\end{aligned}$$

The desired result is obtained by rearranging the terms and using summation notation. □

The following proposition is what we have been after for tangent vector fields along curves.

Proposition 5.7.5. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $c: (a, b) \to x(U)$ be a smooth curve, and let $Y: (a, b) \to \mathbb{R}^3$ be a smooth vector field along c that is tangent to M along c. Then*

$$\frac{DY}{dt} = \sum_{k=1}^{2} \Big(\frac{dY^k}{dt} + \sum_{i=1}^{2} \sum_{j=1}^{2} \Gamma_{ij}^{k}(\bar{c}(t))\, c_j'(t)\, Y^i(t) \Big)\, x_k(\bar{c}(t)).$$

Proof. Using Lemmas 5.6.8 and 5.7.4, we see that

$$\begin{aligned}\frac{DY}{dt} &= \sum_{i=1}^{2} \frac{D}{dt} \left[Y^i(t) x_i(\bar{c}(t)) \right] = \sum_{i=1}^{2} \Big(\frac{dY^i}{dt} x_i(\bar{c}(t)) + Y^i(t) \frac{D(x_i \circ \bar{c})}{dt} \Big) \\ &= \sum_{i=1}^{2} \Big(\frac{dY^i}{dt} x_i(\bar{c}(t)) + Y^i(t) \sum_{k=1}^{2} \sum_{j=1}^{2} \Gamma_{ji}^{k}(\bar{c}(t))\, c_j'(t)\, x_k(\bar{c}(t)) \Big).\end{aligned}$$

This last expression is seen to equal the desired result after some rearranging and changing of indices, and using Lemma 5.7.1. □

Finally, we turn tangent vector fields on surfaces.

Proposition 5.7.6. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $Z: x(U) \to \mathbb{R}^3$ be a smooth tangent vector field, let $p \in x(U)$ be a point, and let $v \in T_pM$ be a vector. Then*

$$\nabla_v Z = \sum_{k=1}^{2} \sum_{j=1}^{2} \Big(v^j \frac{\partial Z^k}{\partial u_j}(\bar{p}) + \sum_{i=1}^{2} \Gamma_{ij}^k(\bar{p})\, v^j\, Z^i(\bar{p}) \Big)\, x_k(\bar{p}).$$

Proof. Let $c: (-\epsilon, \epsilon) \to x(U)$, for some number $\epsilon > 0$, be a curve such that $c(0) = p$ and $c'(0) = v$. Observe that $\bar{c}(0) = \bar{p}$, using our usual notation. We know by Exercise 5.2.6 that $c_1'(0)x_1(\bar{p}) + c_2'(0)x_2(\bar{p}) = c'(0) = v = v^1 x_1(\bar{p}) + v^2 x_2(\bar{p})$. Equating coefficients, we deduce that $c_1'(0) = v^1$ and $c_2'(0) = v^2$.

The function $Z \circ c: (-\epsilon, \epsilon) \to \mathbb{R}^3$ satisfies the hypotheses of Proposition 5.7.5. Further, it is straightforward to see that $Z \circ c(t) = Z^1(\bar{c}(t))x_1(\bar{c}(t)) + Z^2(\bar{c}(t))x_2(\bar{c}(t))$, so that $Z^1 \circ \bar{c}$ and $Z^2 \circ \bar{c}$ must be the unique coordinate functions for $Z \circ c$. By the chain rule we have

$$\frac{d(Z^k \circ \bar{c})}{dt} = \sum_{j=1}^{2} \frac{\partial Z^k}{\partial u_j}(\bar{c}(t))\, c_j'(t)$$

for each $k = 1, 2$. At $t = 0$ this last equation yields

$$\frac{d(Z^k \circ \bar{c})}{dt}(0) = \sum_{j=1}^{2} \frac{\partial Z^k}{\partial u_j}(\bar{p}) v^j.$$

Propositions 5.6.9 (i) and 5.7.5 now tell us that

$$\begin{aligned}
\nabla_v Z &= \frac{D(Z \circ c)}{dt}(0) \\
&= \sum_{k=1}^{2} \Big(\frac{d(Z^k \circ \bar{c})}{dt}(0) + \sum_{i=1}^{2} \sum_{j=1}^{2} \Gamma_{ij}^k(\bar{c}(0))\, c_j'(0)\, Z^i(\bar{c}(0)) \Big)\, x_k(\bar{c}(0)) \\
&= \sum_{k=1}^{2} \Big(\sum_{j=1}^{2} \frac{\partial Z^k}{\partial u_j}(\bar{p}) v^j + \sum_{i=1}^{2} \sum_{j=1}^{2} \Gamma_{ij}^k(\bar{p})\, v^j\, Z^i(\bar{p}) \Big)\, x_k(\bar{p}) \\
&= \sum_{k=1}^{2} \sum_{j=1}^{2} \Big(v^j \frac{\partial Z^k}{\partial u_j}(\bar{p}) + \sum_{i=1}^{2} \Gamma_{ij}^k(\bar{p})\, v^j\, Z^i(\bar{p}) \Big)\, x_k(\bar{p}). \quad \square
\end{aligned}$$

Example 5.7.7. We view the sphere S^2 as a surface of revolution, using the coordinate patch x given by Equation 5.3.2, with $R = 1$. Note that $r(t) = \cos t$ and $z(t) = \sin t$. Let $p \in S^2$ be $p = \begin{pmatrix} \sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix}$, let $v \in T_pS^2$ be $v = \begin{pmatrix} -1 \\ 1 \\ 1 \end{pmatrix}$ and let Z be the tangent vector field on the image of x given by $Z \circ x = \begin{pmatrix} -\cos t \sin\theta \\ \cos t \cos\theta \\ 0 \end{pmatrix}$. (Although in our discussion above we had Z given directly, there is no reason not to give Z by giving $Z \circ x$, since we could in theory always compute $Z = (Z \circ x) \circ x^{-1}$.) We want to find $\nabla_v Z$ using Proposition 5.7.6.

First, observe that $\bar{p} = \begin{pmatrix} \pi/4 \\ 0 \end{pmatrix}$. Next, we have

$$x_1 = \begin{pmatrix} -\sin t \cos\theta \\ -\sin t \sin\theta \\ \cos t \end{pmatrix}, \quad \text{and} \quad x_2 = \begin{pmatrix} -\cos t \sin\theta \\ \cos t \cos\theta \\ 0 \end{pmatrix},$$

and hence

$$x_1(\bar{p}) = \begin{pmatrix} -\sqrt{2}/2 \\ 0 \\ \sqrt{2}/2 \end{pmatrix}, \quad \text{and} \quad x_2(\bar{p}) = \begin{pmatrix} 0 \\ \sqrt{2}/2 \\ 0 \end{pmatrix}.$$

Note that $Z \circ x = 0 \cdot x_1 + 1 \cdot x_2$, so that the coordinate functions Z^1 and Z^2 of Z are the constant functions $Z^1 = 0$ and $Z^2 = 1$. Hence the partial derivatives of the Z^i are all zero. Also, the vector v can be written as $v = (\sqrt{2}/2)x_1(\bar{p}) + (\sqrt{2}/2)x_2(\bar{p})$; hence $v^1 = v^2 = \sqrt{2}/2$. Finally, we need the Christoffel symbols for the coordinate patch we are using, and these can be obtained by substituting our particular functions $r(t)$ and $z(t)$ into the result of Exercise 5.7.3. We obtain $\Gamma^1_{22} = \sin t \cos t$, $\Gamma^2_{12} = \Gamma^2_{21} = -\tan t$, and all the other Christoffel symbols are zero. Hence $\Gamma^1_{22}(\bar{p}) = 1/2$, $\Gamma^2_{12}(\bar{p}) = \Gamma^2_{21}(\bar{p}) = -1$, and all the other

Christoffel symbols at $\bar{p}$ are zero. We then compute

$$\begin{aligned}\nabla_v Z &= \sum_{k=1}^{2}\sum_{j=1}^{2}\Big(\frac{\sqrt{2}}{2}\cdot 0+\sum_{i=1}^{2}\Gamma_{ij}^{k}(\bar{p})\,\frac{\sqrt{2}}{2}\,Z^{i}(\bar{p})\Big)\,x_k(\bar{p})\\ &= \Gamma_{22}^{1}(\bar{p})\,\frac{\sqrt{2}}{2}\,Z^{2}(\bar{p})\,x_1(\bar{p})+\Gamma_{12}^{2}(\bar{p})\,\frac{\sqrt{2}}{2}\,Z^{1}(\bar{p})\,x_2(\bar{p})\\ &\quad+\Gamma_{21}^{2}(\bar{p})\,\frac{\sqrt{2}}{2}\,Z^{2}(\bar{p})\,x_2(\bar{p})\\ &= \frac{1}{2}\,\frac{\sqrt{2}}{2}\cdot 1\cdot\begin{pmatrix}-\sqrt{2}/2\\0\\\sqrt{2}/2\end{pmatrix}+(-1)\,\frac{\sqrt{2}}{2}\cdot 0\cdot\begin{pmatrix}0\\\sqrt{2}/2\\0\end{pmatrix}\\ &\quad+(-1)\,\frac{\sqrt{2}}{2}\cdot 1\cdot\begin{pmatrix}0\\\sqrt{2}/2\\0\end{pmatrix}=\begin{pmatrix}-1/4\\1/2\\1/4\end{pmatrix}. \quad \Diamond\end{aligned}$$

We conclude this section with the following lemma, to be used in Section 7.3. Let $(a, b) \times (d, e)$ be a subset of $\mathbb{R}^2$, and let $h\colon (a, b) \times (d, e) \to M$ be a smooth function. We then have two obvious tangent vector fields on the image of h in M, namely $\frac{\partial h}{\partial s}$ and $\frac{\partial h}{\partial t}$. We can restrict each of these vector fields to curves of the form $t = k$ for some constant $k \in (d, e)$, and we can then calculate $\frac{D}{\partial s}$ of these vector fields along each of these curves; similarly, we can restrict each of these vector fields to curves of the form $s = m$ for some constant $m \in (a, b)$, and we can then calculate $\frac{D}{\partial t}$ of these vector fields along each of these curves. See Figure 5.7.1. In particular, we can compute each of $\frac{D}{\partial s}\frac{\partial h}{\partial t}$ and $\frac{D}{\partial t}\frac{\partial h}{\partial s}$.

Lemma 5.7.8. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $(a, b) \times (d, e)$ be a subset of $\mathbb{R}^2$, and let $h\colon (a, b) \times (d, e) \to M$ be a smooth function. Then*

$$\frac{D}{\partial s}\frac{\partial h}{\partial t} = \frac{D}{\partial t}\frac{\partial h}{\partial s}$$

at all points in $(a, b) \times (d, e)$.

Proof. Exercise 5.7.8. □

Exercises

5.7.1*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x\colon U \to M$ be a coordinate patch.

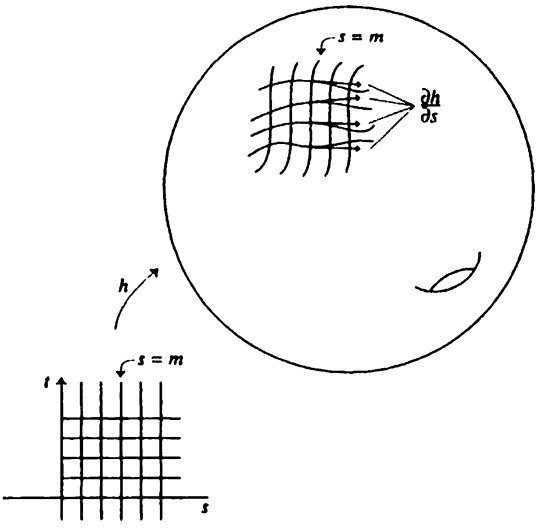


Figure 5.7.1

(i) Let $Z: x(U) \to \mathbb{R}^3$ be a smooth tangent vector field on $x(U)$. Show that there are unique smooth coordinate functions $Z^1, Z^2: U \to \mathbb{R}$ such that $Z \circ x(\bar{q}) = Z^1(\bar{q})x_1(\bar{q}) + Z^2(\bar{q})x_2(\bar{q})$ for $\bar{q} \in U$.

(ii) Let $Y: (a, b) \to \mathbb{R}^3$ be a smooth vector field along c which is tangent to M along c. Show that there are unique smooth coordinate functions $Y^1, Y^2: (a, b) \to \mathbb{R}$ such that $Y(t) = Y^1(t)x_1(\bar{c}(t)) + Y^2(t)x_2(\bar{c}(t))$ for all $t \in (a, b)$.

5.7.2*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. If $Z: x(U) \to \mathbb{R}^3$ is a smooth vector field (not necessarily tangent), show that

$$\tilde{\nabla}_{x_i(\bar{p})} Z = \frac{\partial Z \circ x}{\partial u_i}(\bar{p})$$

at each point in $\bar{p} \in U$.

5.7.3*. Show that the Christoffel symbols for a general surface of revolution, as parametrized in Section 5.3, are

$$\Gamma^1_{11} = \frac{r'r'' + z'z''}{(r')^2 + (z')^2}, \qquad \Gamma^1_{22} = -\frac{rr'}{(r')^2 + (z')^2}, \qquad \Gamma^2_{12} = \Gamma^2_{21} = \frac{r'}{r}$$

and all other Γ_{ij}^k are 0. If the profile curve of the surface of revolution is unit speed, verify that $\Gamma_{11}^1 = 0$ and $\Gamma_{22}^1 = -rr'$.

5.7.4*. Suppose that a coordinate patch has $g_{11} = 1$, $g_{12} = 0$ and $g_{22} = G$ for some smooth function G. Show that the Christoffel symbols for this coordinate patch are

$$\Gamma_{22}^1 = -\frac{1}{2}\frac{\partial G}{\partial u_1}, \qquad \Gamma_{12}^2 = \Gamma_{21}^2 = \frac{1}{2G}\frac{\partial G}{\partial u_1}, \qquad \Gamma_{22}^2 = \frac{1}{2G}\frac{\partial G}{\partial u_2}$$

and all other Γ_{ij}^k are 0.

5.7.5. Let $c\colon (a, b) \to \mathbb{R}^3$ be an injective, unit-speed curve. Show that the Christoffel symbols for the rectifying developable surface generated by the curve are

$$\Gamma_{22}^1 = -t\tau^2(s), \qquad \Gamma_{12}^2 = \Gamma_{21}^2 = \frac{t\tau^2(s)}{1 + t^2\tau^2(s)}, \qquad \Gamma_{22}^2 = \frac{t^2\tau(s)\tau'(s)}{1 + t^2\tau^2(s)}$$

and all other Γ_{ij}^k are 0 (where $\tau(s)$ is the torsion of the curve c).

5.7.6. Show that for any coordinate patch

$$\frac{1}{2}\frac{\partial}{\partial u_1}\ln\det(g_{ij}) = \Gamma_{11}^1 + \Gamma_{12}^2.$$

5.7.7. Let M be the right helicoid, as parametrized in Section 5.3 (though once again using u_1 and u_2 instead of s and t), let $p \in M$ be $p = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$, let $v \in T_pM$ be $v = \begin{pmatrix} 2 \\ 1 \\ b \end{pmatrix}$ and let Z be the tangent vector field on the image of x given by $Z \circ x = \begin{pmatrix} u_2 \cos u_1 \\ u_2 \sin u_1 \\ 0 \end{pmatrix}$. Find $\nabla_v Z$.

5.7.8*. Prove Lemma 5.7.8.

5.8 Length and Area

We wish to find the lengths of smooth curves contained in smooth surfaces in $\mathbb{R}^3$, and the areas of regions contained in such surfaces. We begin with the lengths

of curves. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. If $c: (a, b) \to x(U)$ is a smooth curve, we can certainly compute its length using Equation 4.3.1, simply ignoring the fact that the image of c is in M. However, if we let $c_1, c_2: (a, b) \to \mathbb{R}$ denote the coordinate functions of c with respect to x, the following lemma gives a formula for the length of c expressed in terms of c_1, c_2 and x.

Lemma 5.8.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, and let $c: (a, b) \to x(U)$ be a smooth curve. Then the length of c is given by*

$$\textit{Length}(c) = \int_a^b \sqrt{(c_1'(t))^2 g_{11}(\bar{c}(t)) + 2c_1'(t)\, c_2'(t)\, g_{12}(\bar{c}(t)) + (c_2'(t))^2 g_{22}(\bar{c}(t))}\, dt.$$

Proof. Exercise 5.8.1. □

Example 5.8.2. Let $M \subset \mathbb{R}^3$ be the right helicoid with the parametrization given in Section 5.3. We wish to find the length of the curve $c: [0, 1] \to M$ given by

$$c(t) = \begin{pmatrix} (bt^2/2)\cos t \\ (bt^2/2)\sin t \\ bt \end{pmatrix}.$$

It is seen that $c(t) = x\left(\begin{pmatrix} t \\ bt^2/2 \end{pmatrix}\right)$, so that $\bar{c}(t) = \begin{pmatrix} t \\ bt^2/2 \end{pmatrix}$, and thus $c_1(t) = t$ and $c_2(t) = \frac{bt^2}{2}$. Hence $c_1'(t) = 1$ and $c_2'(t) = bt$. Using the values for the metric coefficients of the right helicoid computed in Exercise 5.5.8 we see that

$$g_{11}(\bar{c}(t)) = t^2 + b^2, \qquad g_{12}(\bar{c}(t)) = 0, \qquad g_{22}(\bar{c}(t)) = 1.$$

Hence

$$\text{Length}(c) = \int_0^1 \sqrt{1^2 \cdot (b^2t^4/4 + b^2) + 2 \cdot 1 \cdot bt \cdot 0 + b^2t^2 \cdot 1}\; dt = \frac{7b}{6}. \quad \Diamond$$

We now find the area of a region in a smooth surface in $\mathbb{R}^3$, a problem not as simple as it may appear at first. Recall from Calculus how to find the area of a region of $\mathbb{R}^2$ by double integration: Divide the region into small pieces (most of which are rectangles), find the areas of all the rectangles, add these areas up, and take the limit as the little rectangles have smaller and smaller areas. Although a similar construction for a non-planar surface might be attempted, examples show that such an attempt is doomed to failure. See [SK1, pp. 128–130].

An alternative approach is as follows. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ be a coordinate patch. Suppose $S \subset M$ is a region contained in $x(U)$. (For regions not contained in the image of a single coordinate patch we can break up the region into pieces, each of which is contained in the image of a coordinate patch.) Intuitively, we cover S by a lot of small parallelograms, which are spanned by the vectors $\{x_1, x_2\}$ at various points in U. The area of the parallelograms spanned by these vectors is simply $\|x_1 \times x_2\|$, which by Lemma 5.5.2 (i) is equal to $\sqrt{\det(g_{ij})}$. If we add the areas of these parallelograms, and take the limit as the parallelograms get smaller and smaller, we are led to the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $S \subset x(U)$ be a set. The **area** of S, denoted Area(S), is the number

$$\text{Area}(S) = \iint_{x^{-1}(S)} \sqrt{\det(g_{ij})}\, ds\, dt, \tag{5.8.1}$$

provided the integral exists. ◊

We cannot "prove" that Area(S) as we have defined it corresponds to our intuition, though a more detailed explanation that this definition is plausible can be found in [DO1, §2-8]. The following lemma shows that the area computed using a coordinate patch is independent of the choice of coordinate patch.

Lemma 5.8.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ and $y: V \to M$ be coordinate patches, and let (g_{ij}) and $(\bar{g}_{ij})$ denote the metric coefficients for x and y respectively. If $S \subset x(U) \cap y(V)$ is a set, then*

$$\iint_{x^{-1}(S)} \sqrt{det(g_{ij})}\, ds\, dt = \iint_{y^{-1}(S)} \sqrt{det(\bar{g}_{ij})}\, dq\, dr.$$

Proof. The essence of the proof is the change of variable formula for double integrals (see [SK1, p. 67]). Let J denote the Jacobian matrix of the change of coordinate function $\phi_{x,y}$. Then by Lemma 5.5.3 we have

$$\det(g_{ij}) = \det(\bar{g}_{ij} \circ \phi_{x,y})\, (\det J)^2.$$

Recalling that $y \circ \phi_{x,y} = x$ for suitable restrictions of the domains of x and y, we compute

$$\iint_{x^{-1}(S)} \sqrt{\det(g_{ij})}\, ds\, dt = \iint_{\phi_{x,y}^{-1}(y^{-1}(S))} \sqrt{\det(\bar{g}_{ij}(\phi_{x,y}))}\, |\det J|\, ds\, dt$$
$$= \iint_{y^{-1}(S)} \sqrt{\det(\bar{g}_{ij})}\, dq\, dr,$$

where the last equality is precisely the change of variable formula. □

Example 5.8.4. Let us compute the area of the sphere S_R of radius R centered at the origin. The coordinate patch of S_R given by Equation 5.3.2, taking $U = (-\pi/2, \pi/2) \times (-\pi, \pi)$ as the domain, covers all of the surface except for an arc, so it suffices to find the area of $x(U)$. Using Exercise 5.5.4 in the special case of S_R it can be computed that $\det(g_{ij}) = R^4 \cos^2 t$. Hence

$$\text{Area}(S_R) = \text{Area}(x(U)) = \iint_{x^{-1}(x(U))} \sqrt{\det(g_{ij})}\, dt\, d\theta$$
$$= \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} R^2 \cos t\, dt\, d\theta = 4\pi R^2. \quad \Diamond$$

We are now in a position to compute the integrals of real-valued functions defined on a surface, once again assuming that the region over which we are integrating is contained in the image of a coordinate patch.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $S \subset x(U)$ be a set and let $f: S \to \mathbb{R}$ be a function. The **integral** of f over S, denoted $\int_S f dA$, is the number

$$\int_S f dA = \iint_{x^{-1}(S)} f(x(\begin{pmatrix} s \\ t \end{pmatrix}))\sqrt{\det(g_{ij})}\, ds\, dt,$$

provided the integral exists. ◊

The symbol "dA" is analogous to the use of "dx" in single variable integrals. The nice feature of the above definition is that it takes an integral on a surface and converts it into an integral over a region in $\mathbb{R}^2$. We now verify that this definition does not depend upon the choice of coordinate patch used.

Lemma 5.8.5. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ and $y: V \to M$ be coordinate patches, and let (g_{ij}) and $(\bar{g}_{ij})$ denote the metric coefficients for*

x and y respectively. If $S \subset x(U) \cap y(V)$ is a set, and if $f: S \to \mathbb{R}$ is a function for which the following two integrals exist, then

$$\iint_{x^{-1}(S)} f(x(\begin{pmatrix} s \\ t \end{pmatrix}))\sqrt{det(g_{ij})}\, ds\, dt = \iint_{y^{-1}(S)} f(y(\begin{pmatrix} q \\ r \end{pmatrix}))\sqrt{det(\bar{g}_{ij})}\, dq\, dr.$$

Proof. Exercise 5.8.6. □

Example 5.8.6. Continuing Example 5.8.4, let us integrate over all of S_R the function $f: S_R \to \mathbb{R}$ given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = z^2\sqrt{x^2+y^2}$. We compute that $f(x(\begin{pmatrix} t \\ \theta \end{pmatrix})) = R^3 \sin^2 t \cos t$. Hence

$$\int_{S_R} f dA = \int_{-\pi}^{\pi} \int_{-\pi/2}^{\pi/2} R^3 \sin^2 t \cos t\, R^2 \cos t\, dt\, d\theta = \frac{\pi^2 R^5}{4}. \quad \Diamond$$

Exercises

5.8.1*. Prove Lemma 5.8.1.

5.8.2. Let $M \subset \mathbb{R}^3$ be the saddle surface with the parametrization given in Example 5.5.1 (2). Find the length of the curve $c: [-1, 1] \to M$ given by

$$c(t) = \begin{pmatrix} e^t \\ e^t \\ e^{2t} \end{pmatrix}.$$

5.8.3. Find the area of the torus of large radius R and small radius r.

5.8.4. Find the area of one complete turn of the right helicoid for $t \in (-1, 1)$.

5.8.5. Find the area of a general monge patch restricted to a region R in the x-y plane. Compare what you get to the formula for the surface area of the graph of a function of two variables found in any Calculus text.

5.8.6*. Prove Lemma 5.8.5.

5.8.7. Integrate the function given by $f(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = x^2 + y^2 + z^2$ over the torus of large radius R and small radius r.

5.9 Isometries

An arbitrary smooth map $f: M \to N$ of smooth surfaces might or might not intuitively "deform" M by stretching, shrinking, etc. Think of the fact that it is easy to wrap a cylinder with a sheet of wrapping paper, but difficult to wrap a ball. We construct here a rigorous definition that corresponds to the intuitive notion of preserving the geometry of a surface. Since we can compute quantities such as length and area by using the first fundamental form, we will define a class of maps that preserve the first fundamental form. We start with a useful technicality.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $f: M \to \mathbb{R}^3$ be a smooth map and let $p \in M$ be a point. The **differential** of f at p is the map $df_p: T_pM \to \mathbb{R}^3$ given by

$$df_p(v) = \widetilde{\nabla}_v f$$

for all $v \in T_pM$. $\Diamond$

Though this definition appears to be no more than renaming the directional derivative, it has the effect of changing our point of view. When we write $\widetilde{\nabla}_v f$ we are thinking of v as fixed, with $\widetilde{\nabla}_v$ acting on the set of smooth functions (or vector fields) on M by taking their directional derivatives. By contrast, when we write df_p we are thinking of f as being fixed, with df_p acting on the vectors in T_pM. What makes the differential useful is the following result.

Lemma 5.9.1. *Let $M, N \subset \mathbb{R}^3$ be smooth surfaces, let $f: M \to N$ be a smooth map, and let $p \in M$ be a point. Then the map df_p is a linear map from T_pM into $T_{f(p)}N$.*

Proof. The linearity follows immediately from Lemma 5.6.4 (i). Let $v \in T_pM$ be a vector. Then there is some curve $c: (-\epsilon, \epsilon) \to M$ such that $c(0) = p$ and $c'(0) = v$. Hence

$$df_p(v) = \widetilde{\nabla}_v f = (f \circ c)'(0).$$

Observe that $f \circ c: (-\epsilon, \epsilon) \to N$ is a curve in N such that $f \circ c(0) = f(p)$. By the definition of tangent vectors we have $(f \circ c)'(0) \in T_{f(p)}N$, and thus $df_p(v) \in T_{f(p)}N$. $\square$

Example 5.9.2. Consider the map $f: \mathbb{R}^2 \to S^1 \times \mathbb{R}$ given by

$$f\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = \begin{pmatrix} \cos s \\ \sin s \\ t \end{pmatrix}.$$

Let $p \in \mathbb{R}^2$ be the origin, so that $f(p) = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. It is not hard to see that $T_p\mathbb{R}^2$ is the x-y plane, and $T_{f(p)}(S^1 \times \mathbb{R})$ is the y-z plane. Let $v = \begin{pmatrix} a \\ b \end{pmatrix} \in T_p\mathbb{R}^2$ be a vector. The function $c: (-\infty, \infty) \to \mathbb{R}^2$ given by $c(t) = \begin{pmatrix} at \\ bt \end{pmatrix}$ has $c(0) = p$ and $c'(0) = v$. Then $f \circ c(t) = \begin{pmatrix} \cos at \\ \sin at \\ bt \end{pmatrix}$. Hence

$$df_p(v) = \tilde{\nabla}_v f = (f \circ c)'(0) = \begin{pmatrix} 0 \\ a \\ b \end{pmatrix}. \quad \Diamond$$

The notion of the differential can be used to formulate the analog for smooth surfaces of the Inverse Function Theorem.

Proposition 5.9.3. *Let $M \subset \mathbb{R}^3$ and $N \subset \mathbb{R}^3$ be smooth surfaces, let $f: M \to N$ be a smooth map and let $p \in M$ be a point. If df_p is a non-singular linear map then there is an open subset $W \subset M$ containing p such that $f(W)$ is open in N and $f|W: W \to F(W)$ is a diffeomorphism.*

Proof. Since f is a smooth map it follows from Lemma 5.2.9 that there is a coordinate patch $x: U \to M$ with $p \in x(U)$ and a coordinate patch $y: V \to N$ with $f(p) \in y(V)$ such that the composition

$$y^{-1} \circ f \circ x|x^{-1}(f^{-1}(y(V))): x^{-1}(f^{-1}(y(V))) \to V \subset \mathbb{R}^2$$

is Euclidean smooth. For convenience we let $A = x^{-1}(f^{-1}(y(V)))$ and $g = y^{-1} \circ f \circ x|A$. Observe that $f \circ x|A = y \circ g$. As usual let $\bar{p} = x^{-1}(p)$. It now follows from Exercise 5.9.8 that $df_p \circ dx_{\bar{p}} = dy_{g(\bar{p})} \circ dg_{\bar{p}}$. Since df_p, $dx_{\bar{p}}$ and $dy_{g(\bar{p})}$ are all non-singular linear maps (using the hypothesis of the proposition and Exercise 5.9.7), it follows that $dg_{\bar{p}}$ is also non-singular.

The map $dg_{\bar{p}}$ is a linear map from $T_{\bar{p}}A$ to $T_{g(\bar{p})}V$; both these tangent planes are just $\mathbb{R}^2$. We wish to compute the matrix of $dg_{\bar{p}}$ with respect to the standard basis $\{e_1, e_2\}$ of $\mathbb{R}^2$. For each $i = 1, 2$, let $c_i: (-\infty, \infty) \to \mathbb{R}^2$ be the curve given by $c_i(t) = \bar{p} + te_i$. Thus $c_i(0) = \bar{p}$ and $c_i'(0) = e_i$. Using the definition

of the differential of a map we now compute

$$dg_{\bar{p}}(e_i) = \tilde{\nabla}_{e_i} g = (g \circ c_i)'(0) = \frac{\partial g}{\partial u_i}|_{\bar{p}},$$

where the last equality holds by the definition of partial derivatives. It follows that the matrix of $dg_{\bar{p}}$ with respect to the standard basis of $\mathbb{R}^2$ is precisely the Jacobian matrix $Dg(\bar{p})$. Hence $Dg(\bar{p})$ is non-singular. Applying the Inverse Function Theorem to g we deduce that there is an open subset $T \subset A$ containing p such that $g(T)$ is open in V and $g|T$ is a diffeomorphism from T onto $g(T)$. It is now straightforward to verify that the set $W = f(T) \subset M$ has the desired properties. □

We can now use the concept of the differential of a map to define what we mean by a map that preserves the first fundamental form.

Definition. Let $M, N \subset \mathbb{R}^3$ be smooth surfaces and let $f: M \to N$ be a smooth map. The map f is an **isometry** if it is a diffeomorphism and if

$$I_{f(p)}(df_p(v), df_p(w)) = I_p(v, w) \tag{5.9.1}$$

for all $p \in M$ and all $v, w \in T_pM$. The map f is a **local isometry** if for each point $p \in M$ there is an open subset $V \subset M$ containing p such that $f(V)$ is open in N and $f|V: V \to f(V)$ is an isometry. ◊

Note that we can write Equation 5.9.1 as

$$\langle df_p(v), df_p(w) \rangle = \langle v, w \rangle. \tag{5.9.2}$$

All isometries are local isometries, but as we will see in Example 5.9.5 the converse is not true. Isometries are relatively rare, and it is local isometries that will be of most use. The following proposition helps determine in practice which maps are local isometries.

Proposition 5.9.4. *Let $M, N \subset \mathbb{R}^3$ be smooth surfaces and let $f: M \to N$ be a smooth map. The following are equivalent:*

(1) *The map f is a local isometry.*
(2) *Equation 5.9.1 holds for all points $p \in M$ and all vectors $v, w \in T_pM$.*
(3) *For each point $p \in M$ and each coordinate patch $x: U \to M$ with $p \in x(U)$, there is an open subset $V \subset U$ such that $p \in x(V)$ and that $f \circ x|V: V \to N$ is a coordinate patch with the same metric coefficients as $x|V$.*

(4) *For each point $p \in M$ there is a coordinate patch $x: U \to M$ such that $p \in x(U)$ and $f \circ x: U \to N$ is a coordinate patch with the same metric coefficients as x.*

(5) *For each point $p \in M$ there is an open set $A \subset M$ containing p such that if $c: (a, b) \to A$ is a smooth curve then Length(c) = Length$(f \circ c)$.*

Proof. We prove (2) $\Rightarrow$ (3), leaving the other parts to the reader in Exercise 5.9.2. Let $p \in M$ be a point, and let $x: U \to M$ be a coordinate patch with $p \in x(U)$; we will show that there is some open subset $V \subset U$ containing p such that $y = f \circ x|V: V \to N$ is a coordinate patch, and that it has the same metric coefficients as x. By hypothesis on f we know that $f \circ x$ is a smooth map. Let $e_1 = \binom{1}{0}$ and $e_2 = \binom{0}{1}$ in $\mathbb{R}^2$. Let $q \in x(U)$ be a point, and let $\bar{q} = x^{-1}(q)$. For $i = 1, 2$ we define a curve $c_{q,i}: (-\epsilon, \epsilon) \to x(U)$ by

$$c_{q,i}(t) = x(\bar{q} + te_i),$$

where $\epsilon > 0$ is some small enough number so that the $c_{q,i}$ are well-defined. We then see from the definition of partial derivatives that

$$c_{q,i}(0) = q \quad \text{and} \quad c'_{q,i}(0) = x_i,$$

and

$$f \circ c_{q,i}(0) = f(q) \quad \text{and} \quad (f \circ c_{q,i})'(0) = (f \circ x)_i,$$

where the functions x_i and $(f \circ x)_i$ are evaluated at $\bar{q} \in U$. It now follows from the definition of df_q that

$$df_q(x_i) = \widetilde{\nabla}_{x_i} f = (f \circ c_{q,i})'(0) = (f \circ x)_i. \tag{5.9.3}$$

Since Equation 5.9.2 holds for all points in M it follows from Equation 5.9.3 that

$$\langle (f \circ x)_i, (f \circ x)_j \rangle = \langle df_q(x_i), df_q(x_j) \rangle = \langle x_i, x_j \rangle \tag{5.9.4}$$

for all $i, j = 1, 2$. We deduce from this equation that $\|(f \circ x)_i\| = \|x_i\|$ for $i = 1, 2$, and that the angle between $(f \circ x)_1$ and $(f \circ x)_2$ equals the angle between x_1 and x_2. Since x_1 and x_2 are linearly independent so are $(f \circ x)_1$ and $(f \circ x)_2$ at $\bar{q}$. It follows that $f \circ x$ has rank 2 at all points of its domain, and by making use of Corollary 4.2.3 we deduce that there is some open subset $V \subset U$ containing $x^{-1}(p)$ such that $f \circ x|V$ is injective. It follows that $f \circ x|V$ is a coordinate patch. The equality of the metric coefficients of $x|V$ and $f \circ x|V$ follows immediately from Equation 5.9.4. $\square$

Note that Condition (3) of Proposition 5.9.4 does not say that any coordinate patch for M and any coordinate patch for N have the same metric coefficients; it is only stated that there are coordinate patches for the two surfaces with the same metric coefficients.

Example 5.9.5. We are now in a position to verify that rolling a piece of paper into a cylinder is a local isometry; more precisely, we will show that the map $f: \mathbb{R}^2 \to S^1 \times \mathbb{R}$ given in Example 5.9.2 is a local isometry. Let $p = \binom{p_1}{p_2} \in \mathbb{R}^2$ be a point, and let $U = (p_1 - \pi, p_1 + \pi) \times \mathbb{R}$. Then the inclusion map $x: U \to \mathbb{R}^2$ given by $x(q) = q$ is a coordinate patch for $\mathbb{R}^2$ with $p \in x(U)$. It is not hard to see that the map $f \circ x: U \to S^1 \times \mathbb{R}$ is injective. Further, the partial derivatives of $f \circ x$ are

$$(f \circ x)_1 = \begin{pmatrix} -\sin s \\ \cos s \\ 0 \end{pmatrix} \quad \text{and} \quad (f \circ x)_2 = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}.$$

Hence

$$x_1 \times x_2 = \begin{pmatrix} \cos s \\ \sin s \\ 0 \end{pmatrix},$$

which is never zero. Therefore $f \circ x$ is a coordinate patch. It is straightforward to compute the metric coefficients for both x and $f \circ x$, and they are both $(g_{ij}) = \left(\begin{smallmatrix} 1 & 0 \\ 0 & 1 \end{smallmatrix}\right)$. Thus Condition (4) of Proposition 5.9.4 is satisfied for each $p \in \mathbb{R}^2$. However, there cannot be an isometry between the two surfaces, since any such isometry would be a diffeomorphism, and hence a homeomorphism, and it was shown in Exercise 3.8.5 that $\mathbb{R}^2 \not\approx S^1 \times \mathbb{R}$. $\Diamond$

Local isometries help us express rigorously what it means for a quantity measured on a smooth surface to be intrinsic. We have mentioned the concept of intrinsicness in Section 3.7 when discussing simplicial surfaces, and intuitively the idea is the same in the present case. Since the first fundamental form is the source of geometric measurements such as length and area, it is reasonable to say that a quantity is intrinsic if is can be measured strictly in terms of the first fundamental form. More practically, we say that a quantity is intrinsic if it can be measured entirely in terms of the metric coefficients when expressed in terms of a coordinate patch. For example, the Christoffel symbols are intrinsic, using Lemma 5.7.2. It is seen from Proposition 5.9.4 that if $f: M \to N$ is a

local isometry, then any intrinsic quantity will be the same at p and $f(p)$ for each $p \in M$.

Exercises

5.9.1. Let the map $f: S^1 \times \mathbb{R} \to \mathbb{R}^2$ be given by

$$f\left(\begin{pmatrix} x \\ y \\ z \end{pmatrix}\right) = \begin{pmatrix} x \\ y \\ 0 \end{pmatrix}.$$

Let $p \in S^1 \times \mathbb{R}$ be $p = \begin{pmatrix} 1 \\ 0 \\ 0 \end{pmatrix}$. Describe the map df_p.

5.9.2*. Prove the remaining parts of Proposition 5.9.4.

5.9.3. Show that $\mathbb{R}^2$ and a cone (without its vertex) are locally isometric.

5.9.4. The catenoid is defined in Exercise 5.5.3. See Figure 5.9.1. Show that the catenoid and the right helicoid with $b = 1$ are locally isometric. It turns out that these surfaces can be continuously deformed one into the other without changing the metric coefficients during the deformation; see [SK3 vol. III, pp. 248–249] or [SR, p. 121] for nice illustrations of this deformation.

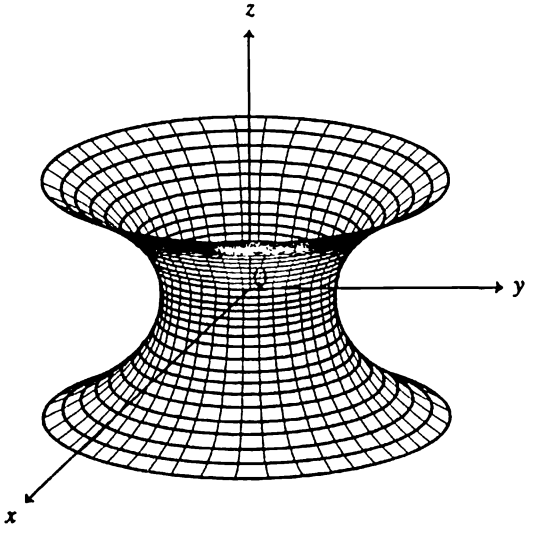


Figure 5.9.1

5.9.5*. Let $M \subset \mathbb{R}^3$ be a surface and let $x: U \to M$ be a coordinate patch. Define the set $Tx(U) \subset \mathbb{R}^3 \times \mathbb{R}^3 = \mathbb{R}^6$ to be

$$Tx(U) = \{(q, v) \in \mathbb{R}^3 \times \mathbb{R}^3 \mid q \in x(U) \text{ and } v \in T_q M\}.$$

Let

$$\Psi: U \times \mathbb{R}^2 \to Tx(U) \subset \mathbb{R}^6$$

be defined by

$$(q, v) \overset{\Psi}{\longmapsto} (x(q), dx_q(v)).$$

Show that Ψ is a homeomorphism. (Although we have not defined what it would mean for a map defined on a set such as $Tx(U)$ to be smooth, such a definition is possible, and the map Ψ is in fact a diffeomorphism.)

5.9.6*. Let $M \subset \mathbb{R}^3$ and $N \subset \mathbb{R}^3$ be smooth surfaces, and let $f: M \to N$ be a diffeomorphism. Show that df_p is a linear isomorphism from $T_p M$ to $T_{f(p)} N$ for each point $p \in M$.

5.9.7*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in x(U)$ be a point and let $v \in T_p M$ be a vector. Let $\bar{p} = x^{-1}(p)$, and suppose v is written in coordinates as $v = v^1 x_1(\bar{p}) + v^2 x_2(\bar{p})$. Combining Exercises 5.2.9 and 5.9.6, it follows that $dx_{\bar{p}}$ is a linear isomorphism from $T_{\bar{p}} U = \mathbb{R}^2$ to $T_p M$. Show that

$$(dx_{\bar{p}})^{-1}(v) = \begin{pmatrix} v^1 \\ v^2 \end{pmatrix}.$$

In particular, observe that $dx_{\bar{p}}(\begin{pmatrix} 1 \\ 0 \end{pmatrix}) = x_1(\bar{p})$ and similarly for x_2.

5.9.8*. Let $M, N, Q \subset \mathbb{R}^3$ be smooth surfaces, let $f: M \to N$ and $g: N \to Q$ be smooth maps and let $p \in M$ be a point. Show that $d(g \circ f)_p = dg_{f(p)} \circ df_p$.

5.9.9*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. If $v, w \in T_p M$ are any two linearly independent vectors, show that there is a coordinate patch $x: U \to M$ such that $p \in x(U)$ and $x_1(\bar{p}) = v$ and $x_2(\bar{p}) = w$, where $\bar{p} = x^{-1}(p)$.

5.9.10*. Let $M \subset \mathbb{R}^3$ be a smooth surface. Let R be a non-singular orthogonal 3×3 and let $q \in \mathbb{R}^3$ be a vector. If $F: \mathbb{R}^3 \to \mathbb{R}^3$ is the map given by

$F(v) = Rv+q$, show that $F: M \to F(M)$ is an isometry. (The set $F(M) \subset \mathbb{R}^3$ is a smooth surface by Exercise 5.2.4.)

Appendix A5.1 Proof of Proposition 5.3.1

Proof of Proposition 5.3.1 We use the Inverse Function Theorem, following [DO1]. Let $p = \begin{pmatrix} p_1 \\ p_2 \\ p_3 \end{pmatrix} \in F^{-1}(a)$ be a point. We will construct a coordinate patch $x: U \to F^{-1}(a)$ such that $p \in x(U)$, where U is some open subset of $\mathbb{R}^2$ to be determined, such that $x(U)$ is open in $F^{-1}(a)$ and x is a homeomorphism onto $x(U)$. It will then follow that $F^{-1}(a)$ is a topological surface and that it is smooth.

We start by noting that since $DF(p)$ has maximal rank, at least one of $\frac{\partial F}{\partial u_1}(p)$, $\frac{\partial F}{\partial u_2}(p)$ and $\frac{\partial F}{\partial u_3}(p)$ is not zero. Let us assume that $\frac{\partial F}{\partial u_3}(p) \neq 0$. We now define a function $G: V \to \mathbb{R}^3$ by

$$G(u) = \begin{pmatrix} u_1 \\ u_2 \\ F(u) \end{pmatrix},$$

where $u = \begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix}$. Clearly G is a smooth function. It is straightforward to verify that $\det DG(p) \neq 0$. Hence we can apply the Inverse Function Theorem (Theorem 4.2.1) to G at the point p, to deduce that there is a subset $W \subset V$ that contains p and that is open in $\mathbb{R}^3$, such that $G(W)$ is open in $\mathbb{R}^3$ and G is a diffeomorphism from W onto $G(W)$. Observe also that

$$G(p) = \begin{pmatrix} p_1 \\ p_2 \\ F(p) \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ a \end{pmatrix},$$

so $\begin{pmatrix} p_1 \\ p_2 \\ a \end{pmatrix} \in G(W)$.

By Lemma 1.2.9 (ii) we can find an open subset $U \subset \mathbb{R}^2$ containing $\begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$ and a number $\epsilon > 0$ such that

$$U \times (a-\epsilon, a+\epsilon) \subset G(W).$$

The function $G^{-1}: G(W) \to W$ is a smooth map and can be written in components as

$$G^{-1}(u) = \begin{pmatrix} g_1(u) \\ g_2(u) \\ g_3(u) \end{pmatrix}$$

for some smooth functions $g_1, g_2, g_3: G(W) \to \mathbb{R}$. By the definition of inverse functions we have

$$\begin{pmatrix} u_1 \\ u_2 \\ u_3 \end{pmatrix} = u = G \circ G^{-1}(u) = \begin{pmatrix} g_1(u) \\ g_2(u) \\ F\begin{pmatrix} g_1(u) \\ g_2(u) \\ g_3(u) \end{pmatrix} \end{pmatrix}.$$

Therefore

$$g_1(u) = u_1, \quad g_2(u) = u_2 \quad \text{and} \quad F(\begin{pmatrix} u_1 \\ u_2 \\ g_3(u) \end{pmatrix}) = u_3. \tag{A5.1.1}$$

Define a map $h: U \to \mathbb{R}$ by

$$h(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}) = g_3(\begin{pmatrix} u_1 \\ u_2 \\ a \end{pmatrix}).$$

It follows from Equation A5.1.1 that

$$F(\begin{pmatrix} u_1 \\ u_2 \\ h(\bar{u}) \end{pmatrix}) = a \tag{A5.1.2}$$

for all $\bar{u} = \binom{u_1}{u_2} \in U$.

Let $x: U \to F^{-1}(a)$ be given by

$$x(\bar{u}) = \begin{pmatrix} u_1 \\ u_2 \\ h(\bar{u}) \end{pmatrix}.$$

Observe that $p \in x(U)$. The desired properties of x will follow from Equation A5.1.3 below; note that the left hand side of Equation A5.1.3 is an open subset

of $F^{-1}(a)$ containing the point p, and the right hand side of Equation A5.1.3 is the graph of the function h, which is simply a monge patch.

The equation we need to prove is

$$F^{-1}(a)\cap G^{-1}(U\times(a-\epsilon,a+\epsilon)) = \left\{ \begin{pmatrix} u_1 \\ u_2 \\ h(\bar{u}) \end{pmatrix} \mid \bar{u} = \begin{pmatrix} u_1 \\ u_2 \end{pmatrix} \in U \right\}. \quad \text{(A5.1.3)}$$

First, suppose that $z = \begin{pmatrix} z_1 \\ z_2 \\ z_3 \end{pmatrix} \in F^{-1}(a) \cap G^{-1}(U \times (a-\epsilon, a+\epsilon))$. Since $z \in G^{-1}(U \times (a-\epsilon, a+\epsilon))$ it follows immediately from Equation A5.1.1 that $\bar{z} = \begin{pmatrix} z_1 \\ z_2 \end{pmatrix} \in U$. Using Equation A5.1.2 we deduce that

$$G\left(\begin{pmatrix} z_1 \\ z_2 \\ h(\bar{z}) \end{pmatrix}\right) = \begin{pmatrix} z_1 \\ z_2 \\ F\left(\begin{pmatrix} z_1 \\ z_2 \\ h(\bar{z}) \end{pmatrix}\right) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ a \end{pmatrix}. \quad \text{(A5.1.4)}$$

For convenience, let $z_a = \begin{pmatrix} z_1 \\ z_2 \\ a \end{pmatrix}$. Now, on the other hand, it follows from the hypothesis on z and from the definition of G that

$$G(z) = \begin{pmatrix} z_1 \\ z_2 \\ F(z) \end{pmatrix} = z_a. \quad \text{(A5.1.5)}$$

Combining Equations A5.1.4 and A5.1.5 with the fact that G is injective on W we deduce that $z_3 = h(\bar{z})$. Thus z is in the right hand side of Equation A5.1.3, and we have therefore proved the inclusion $\subset$ of that equation.

Next, let $\begin{pmatrix} z_1 \\ z_2 \\ h(\bar{z}) \end{pmatrix}$ be an element of the right hand side of Equation A5.1.3. It follows immediately from Equation A5.1.2 that this point is in $F^{-1}(a)$. Further, from the definition of U we know that $z_a \in G(W)$. Hence, using the definition of h, we have

$$\begin{pmatrix} z_1 \\ z_2 \\ h(\bar{z}) \end{pmatrix} = \begin{pmatrix} z_1 \\ z_2 \\ g_3(z_a) \end{pmatrix} = G^{-1}(z_a) \in G^{-1}(U \times (a-\epsilon, a+\epsilon)).$$

We have therefore proved the inclusion $\supset$ of Equation A5.1.3. □

Endnotes

Notes for Section 5.2

(A) In contrast to surfaces, where the the topological, simplicial and smooth categories are essentially equivalent, in higher dimensions the topological properties of smooth manifolds are quite different from the topological properties of topological and piecewise linear manifolds (which generalize simplicial surfaces); for example, there are topological manifolds that are not homeomorphic to any smooth manifold. The relations between the three categories of manifolds has been a very active area of research for the past thirty years. For results in dimensions higher than four see [K-S]; for a summary of spectacular recent work in four dimensions see [F-Q] and [F-L].

(B) Our use of Theorem 3.4.5 to demonstrate that every smooth surface can be triangulated makes for an indirect, and needlessly hard, way of proving this fact. See [WH2] or [MU1] for a more direct proof.

(C) Many differential geometry texts avoid using Invariance of Domain (Theorem 2.2.1) in the proof of Proposition 5.2.5 by restricting to the use of coordinate patches that have continuous inverses. The advantage of such a restriction is that it is not necessary to assume a priori that smooth surfaces are topological surfaces; on the other hand, when verifying that a given map is indeed a coordinate patch, the verification of continuity of the inverse map is often neglected (and tedious if carried out), giving our approach an advantage.

Notes for Section 5.3

In the discussion of level surfaces, we make use of the concept of regular values of a smooth function. By Sard's Theorem (see [MI3] or [SK1]), it follows that the regular values of any given smooth function $F: V \to \mathbb{R}$ consist of "most" of the numbers in $\mathbb{R}$, where the word "most" can be given a precise meaning.

Notes for Section 5.4

Recall the notion of orientability of topological surfaces in Section 2.5. Since a smooth surface is also a topological surface the notion of orientability is applicable to smooth surfaces as well. It is also possible to give a description of orientability of smooth surfaces in terms of coordinate patches. In Section

2.4 the Möbius strip was distinguished from the right circular cylinder using the idea of coloring the two sides of the cylinder. Alternately, note that we can choose a normal vector at each point of the cylinder so that the choice of normal vectors is a continuous function on the whole cylinder; on the Möbius strip, on the other hand, no such choice of normal vectors can be made. In other words, we can cover the cylinder with coordinate patches such that when any two of these coordinate patches overlap, they determine the same normal vectors at all points of the overlap; on the Möbius strip no such choice of coordinate patches can be made. See Figure 5.E.1. This last observation holds in general: A smooth surface in $\mathbb{R}^3$ is orientable iff it can be covered with coordinate patches such that when any two of these coordinate patches overlap, they determine the same normal vectors at all points in the overlap.

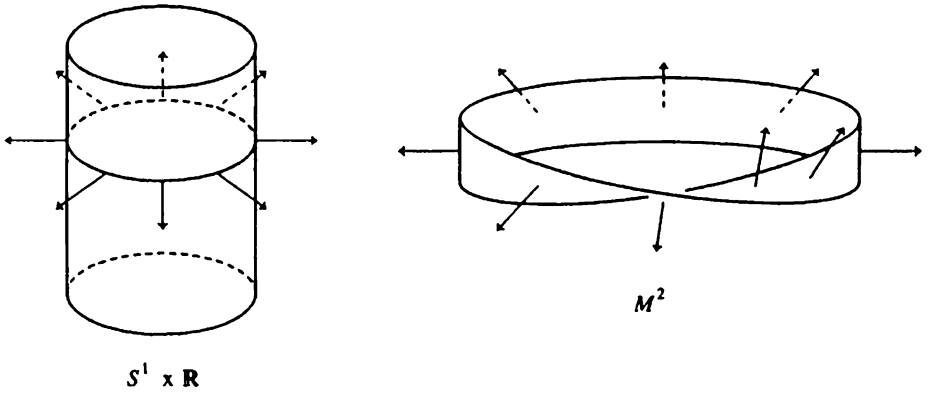


Figure 5.E.1

Notes for Section 5.5

In very classical books on differential geometry, expressions of the form $ds^2 = E\,dx^2 + 2F\,dxdy + G\,dy^2$ are often found. Such expressions were the method of writing the metric coefficients before the modern formulation came into use, and they can be given meaning in terms of differential forms. See [SK1].

Notes for Section 5.6

The notation $\widetilde{\nabla}_v f$ is non-standard, but I do not like the more standard notations I have encountered, and $\widetilde{\nabla}_v$ is very much analogous to the (completely standard) notation ∇_v that we use.

Notes for Section 5.7

In some texts (particularly older ones) the Christoffel symbols are called the "Christoffel symbols of the second kind," indicating, as you might expect, that there are also "Christoffel symbols of the first kind" — though fortunately people seem to make do without them these days.

CHAPTER VI

Curvature of Smooth Surfaces

6.1 Introduction

Just as we defined curvature for simplicial surfaces in Section 3.9 and for smooth curves in Section 4.5, in the present chapter we define curvature for smooth surfaces — technically more difficult than the previous cases, but more rewarding as well. Throughout this section, let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point.

Computing the curvature of a smooth surface means assigning a number to each point of the surface, where this number quantifies the amount of curving at the point. Thus, the curvature of our smooth surface M should be a function $k: M \to \mathbb{R}$ that satisfies a number of properties. Among those properties are the following:

(1) The function $k: M \to \mathbb{R}$ is a smooth map;

(2) a point with an open neighborhood contained in a plane has zero curvature;

(3) if $p, q \in M$ are points such that p has a neighborhood that is more of a sharp peak than a neighborhood of q, then $k(p) > k(q)$. (See Figure 6.1.1.)

(4) the quantity $k(p)$ is intrinsic, as discussed in Section 5.9.

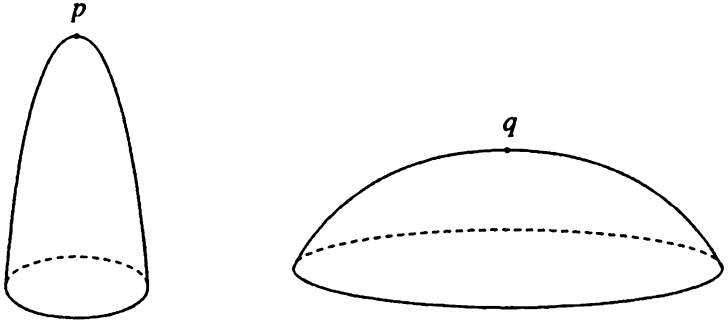


Figure 6.1.1

The first successful definition of curvature for smooth surfaces in $\mathbb{R}^3$ is due to Gauss. Prior to Gauss the approach was to view a surface as a collection of curves, in the hope that understanding the curves could yield information about the surface. More precisely, let $\Omega \subset \mathbb{R}^3$ be a plane that contains p and is parallel to the normal vector to M at p (recall that we translate all normal vectors so that they start at the origin). Then there is an open subset of $\Omega \cap M$ that contains p and that can be parametrized as the image of a smooth curve $c: (a, b) \to M$ (see Exercise 6.1.1 for details). We can compute the planar curvature of this curve. If we consider all possible planes Ω then we obtain a collection of curves through p, and the planar curvatures of these curves ought to tell us how the surface as a whole is curving at p. Unfortunately, it is far from obvious how to assemble the information from the curvatures of all these curves into one number. This approach will turn out to work with hindsight, and we will return to it in Section 6.3 after developing other methods.

Gauss' approach to surfaces was more subtle, involving the surface as a whole rather than considering a surface as a collection of curves. The essence of Gauss' approach is to see how the normal vector to a surface varies as a function of points in the surface. As we have defined it, the normal vector n has as its domain the set U rather than the surface itself (or at least a piece of the surface). To remedy this problem, we use the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. The **Gauss map** of x is the map $\hat{n}: x(U) \to S^2$ given by $\hat{n} = n \circ x^{-1}$. ◊

If we change coordinate patches then $\hat{n}$ might change at most by a minus sign, and it will turn out that our calculations will not be affected. The Gauss map is smooth, as can be seen immediately from the definition of smooth maps on smooth surfaces. The Gauss map certainly depends upon the way in which the surface is sitting in $\mathbb{R}^3$, and might change if the surface is deformed.

Gauss' idea was to choose a small region $T \subset x(U)$ containing p in its interior, and then to compare $\text{Area}(T)$ with something like $\text{Area}(\hat{n}(T))$. In simple cases, such as a sphere, the quantity $\text{Area}(\hat{n}(T))$ works fine; in more complicated situations, however, the map $\hat{n}$ is not necessarily injective, and we need to define something like the area of $\hat{n}(T)$, but something that takes into account whether $\hat{n}$ "flips T over" or not by giving a positive result if it does not flip T over and a negative result if it does. If $\hat{n}$ folds over only part of T, the positive and negative areas would cancel.

Example 6.1.1. (1) Consider S_R, the sphere of radius R in $\mathbb{R}^3$ centered at the origin. If we choose the function n to be the outward pointing unit normal to S_R then $\hat{n}(p) = p/\|p\|$, and this function is defined on all of S_R (so we need not restrict attention to the image of a single coordinate patch). As seen in Figure 6.1.2 the Gauss map simply has the effect of shrinking any region $T \subset S_R$ by a factor of R^2, and there is no problem with using Area($\hat{n}(T)$) in this case.

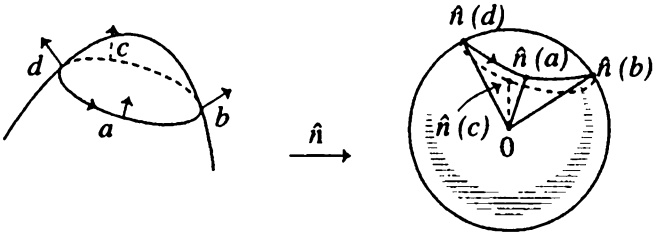


Figure 6.1.2

(2) Consider the saddle surface shown in Figure 6.1.3, with p the origin. We choose the upward pointing normal vectors. Draw a diamond with vertices a, b, c and d surrounding the point p, as in the figure. The image of the diamond under the Gauss map is still a diamond, but it is flipped over; going around the boundary of the image of the diamond in the alphabetical order of the vertices, we see that the direction is reversed in comparison to the boundary of the original diamond. If T were the region bounded by the diamond, we would want to take the area of $\hat{n}(T)$ with a negative sign. $\Diamond$

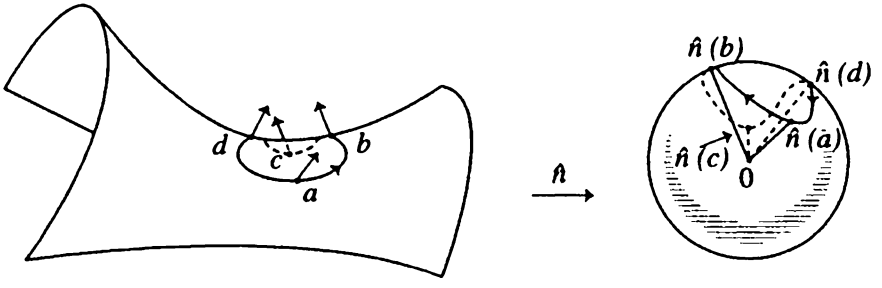


Figure 6.1.3

Instead of figuring out some detailed geometric way of defining what we mean by "area" that takes flipping, and worse, into account, we use integration to take care of the problem. First, note that it follows from the definition of n and Lemma 5.5.2 (i) that

$$\langle x_1 \times x_2, n\rangle = \langle \|x_1 \times x_2\| n, n\rangle = \|x_1 \times x_2\| = \sqrt{\det(g_{ij})}. \tag{6.1.1}$$

Using Equation 5.8.1 we then deduce that

$$\text{Area}(T) = \iint_{x^{-1}(T)} \langle x_1 \times x_2, n\rangle \, ds\, dt. \tag{6.1.2}$$

Now we can think of $n|x^{-1}(T)$ as if it were a coordinate patch for $\hat{n}(T)$ (though strictly speaking this isn't necessarily true). The map n would then be its own normal up to $\pm$, since the image of $\hat{n}$ is in S^2, and vectors from the origin to S^2 are normal to S^2. Hence, by analogy to Equation 6.1.2, it is not unreasonable to define the oriented area of $\hat{n}(T)$ to be the number $\text{Area}_o(\hat{n}(T))$ given by

$$\text{Area}_o(\hat{n}(T)) = \iint_{x^{-1}(T)} \langle n_1 \times n_2, n\rangle \, ds\, dt, \tag{6.1.3}$$

where n_1 and n_2 are the partial derivatives of n.

Gauss considered the ratio $\frac{\text{Area}_o(\hat{n}(T))}{\text{Area}(T)}$. This ratio has some of the properties that we would expect of curvature, for example, property (3) listed above, but unfortunately it depends upon the choice of T. To remedy this situation, Gauss defined his curvature to be

$$K(p) = \lim_{T \to \{p\}} \frac{\text{Area}_o(\hat{n}(T))}{\text{Area}(T)},$$

where the limit is over all regions T that shrink down to the point p. We would have to go to some effort to define rigorously what is meant by this limit, and to show that the limit always exists, and even if we did all this, the result would still end up rather technically unwieldy. We will therefore turn to a somewhat more modern (and very standard) approach, which turns out to be equivalent to Gauss' definition (to be seen in Section 6.4).

Example 6.1.2. Let $M \subset \mathbb{R}^3$ be a surface contained in a plane in $\mathbb{R}^3$. Then the Gauss map for any coordinate patch for M will be constant, so $\text{Area}_o(\hat{n}(T))$ will be zero for any subset $T \subset M$. Hence $K(p) = 0$ for all $p \in M$. Thus Gauss' definition of curvature satisfies property (2) of curvature mentioned earlier. $\Diamond$

Exercises

6.1.1*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point and let $\Omega \subset \mathbb{R}^3$ be a plane that contain p and is parallel to the normal vector to M at p. Our goal is to show that there is an open subset of $\Omega \cap M$ that contains p and that can be parametrized as the image of a smooth curve; more precisely, there is some injective smooth curve $c\colon (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ such that $c((-\epsilon, \epsilon)) \subset \Omega \cap M$ and $c(0) = p$. By a rotation and translation of $\mathbb{R}^3$ we can assume without loss of generality that p is the origin of $\mathbb{R}^3$ and the normal vector to M at p is $\begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}$. The proof is broken into three steps.

(1) Show that there is an open neighborhood $U \subset M$ containing p such that orthogonal projection $\Pi_{T_pM}\colon U \to T_pM$ is an injective smooth map.

(2) Show that the subset U (or some open subset of it) is the image of a monge patch.

(3) Prove the desired result.

6.1.2. Calculate the curvature at all points of each of the following surfaces using Gauss' definition of curvature, though dealing with the limit intuitively. (By symmetry it suffices in each case to find the curvature at one point in each surface.)

(i) $S^1 \times \mathbb{R}$;

(ii) S_R.

6.2 The Weingarten Map and the Second Fundamental Form

We develop in this section the technical tools needed for defining Gaussian curvature. We start with a brief discussion of bilinear forms induced by linear maps; see Section 5.5 for basic definitions concerning bilinear forms.

Definition. Let V be a vector space, and let $\langle\ ,\ \rangle$ be an inner product on V. If $F\colon V \to V$ is a linear map, the **induced** bilinear form, denoted B_F, is the bilinear form on V given by

$$B_F(v, w) = \langle F(v), w \rangle$$

for all $v, w \in V$. $\Diamond$

Now suppose that V is finite-dimensional and that a basis has been chosen for V. If $F: V \to V$ is a linear map, then we can form two matrices using this basis: the matrix for the linear map F and the matrix for the bilinear form B_F. The following two lemmas express the relation between these matrices. We omit the proofs of these lemmas, which use standard ideas from linear algebra (see [FR] for an outline of the proofs). Note that an inner product is itself a type of bilinear form, and it thus has a matrix with respect to the given basis.

Lemma 6.2.1. *Let V be a finite-dimensional vector space for which an inner product and a basis have been chosen; let G denote the matrix of the inner product with respect to the basis. Let $F: V \to V$ be a linear map, let B_F be the induced bilinear form and let A and M denote the matrices for F and B_F respectively with respect to the basis. Then $M = A^t G$.*

Lemma 6.2.2. *Let V be a finite-dimensional vector space.*

(i) *If B is a bilinear form on V, then B is symmetric iff the matrix for B with respect to any basis of V is a symmetric matrix.*

(ii) *Let $F: V \to V$ be a linear map. The following are equivalent:*

(a) *F is self-adjoint;*
(b) *the matrix for F with respect to any orthonormal basis is symmetric;*
(c) *the induced bilinear form B_F is symmetric.*

We now turn to surfaces. Throughout this section let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. We wish to see how the normal vector changes at a point on a smooth surface. We proceed technically by viewing the Gauss map $\hat{n}: x(U) \to S^2 \subset \mathbb{R}^3$ as a vector field on $x(U)$, though it is certainly not a tangent vector field, and taking the directional derivatives $\widetilde{\nabla}_v \hat{n}$ for all tangent vectors $v \in T_p M$.

Lemma 6.2.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in x(U)$ be a point and let $v \in T_p M$ be a vector. Then $\widetilde{\nabla}_v \hat{n} \in T_p M$.*

Proof. The vector field $\hat{n}$ is a unit vector field, so that $\langle \hat{n}, \hat{n} \rangle = 1$ at all points in $x(U)$. Hence

$$0 = \widetilde{\nabla}_v \langle \hat{n}, \hat{n} \rangle = 2 \langle \widetilde{\nabla}_v \hat{n}, \hat{n} \rangle.$$

Thus $\widetilde{\nabla}_v \hat{n}$ is perpendicular to $\hat{n}$, so it must be in $T_p M$. □

The above lemma justifies the following definition; the minus sign in the definition, found in most books, is for later convenience.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. The **Weingarten map** of M at p is the map $L: T_pM \to T_pM$ given by $L(v) = -\widetilde{\nabla}_v \hat{n}$ for all $v \in T_pM$. $\Diamond$

There is a Weingarten map for each point p in the surface, but since there will never be any ambiguity about the point p under consideration, we use the letter L to denote the Weingarten map at each point. Although we chose a coordinate patch in order to have a Gauss map, if we had chosen a different coordinate patch the only possible change in the Weingarten map could be its sign, since the vector field $\hat{n}$ could at most change sign. Hence L is well-defined up to $\pm$; the sign cannot be chosen in any intrinsic way.

The following lemma, which gives a crucial property of the Weingarten map, is derived straightforwardly from Lemma 5.6.4 (i), and we omit the proof.

Lemma 6.2.4. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, and let $p \in x(U)$ be a point. The map $L: T_pM \to T_pM$ is a linear map.*

Example 6.2.5. (1) Let p be a point in a surface M such that p has a flat open neighborhood. In the flat open neighborhood the normal map is constant, and thus the directional derivative of the normal map in any direction is zero. The Weingarten map is thus the zero map.

(2) We continue Example 6.1.1 (1). By the symmetry of the sphere it is clear that the Weingarten map is the same at all points of the sphere, and that at any point $p \in S_R$ the Weingarten map has the same effect on all vectors in T_pS_R. Hence, we only need to compute the Weingarten map acting on one tangent vector at one point. Let $p = \begin{pmatrix} R \\ 0 \\ 0 \end{pmatrix}$, and let $v = \begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} \in T_pS_R$. The curve $c: (-\pi/2, \pi/2) \to S_R$ given by

$$c(t) = \begin{pmatrix} R\cos\frac{t}{R} \\ R\sin\frac{t}{R} \\ 0 \end{pmatrix}$$

has $c(0) = p$ and $c'(0) = v$. The definition of the directional derivative and a simple calculation now show that

$$L(v) = -\tilde{\nabla}_v \hat{n} = -(\hat{n} \circ c)'(0) = -\frac{1}{R}\begin{pmatrix} 0 \\ 1 \\ 0 \end{pmatrix} = -\frac{1}{R}v.$$

By symmetry it follows that L is simply $\frac{1}{R}$ times the identity map. $\Diamond$

The Weingarten map very much depends upon the way in which the surface sits in $\mathbb{R}^3$. A bug living on a surface, that could not see off the surface, could not determine the Weingarten map.

For ease of computation we make the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. The **second fundamental form** of M at p is the bilinear form $\mathrm{II}_p: T_pM \times T_pM \to \mathbb{R}$ induced by the linear map $L: T_pM \to T_pM$, that is

$$\mathrm{II}_p(v, w) = \langle L(v), w \rangle$$

for all $v, w \in T_pM$. We let II denote the second fundamental form at all points p in $x(U)$.The **second fundamental form** of M is the collection, denoted II, of all functions II_p at all points $p \in M$. $\Diamond$

As for the Weingarten map, the second fundamental form is an extrinsic quantity, and is only determined up to $\pm$.

Example 6.2.6. We continue Example 6.2.5. (1) Since the Weingarten map in this case is the zero map, the second fundamental form is the constantly zero bilinear form.

(2) Since the Weingarten map at each point is $\frac{1}{R}$ times the identity map, the second fundamental form at each point is

$$\mathrm{II}(v, w) = \langle \frac{1}{R}v, w \rangle = \frac{1}{R}\langle v, w \rangle. \quad \Diamond$$

Having used coordinate patches so far only to determine a choice of normal vectors, we now turn to calculations making specific use of a given coordinate patch $x: U \to M$. The vectors $\{x_1, x_2\}$ form a basis for T_pM, and we can use this basis to compute matrices for the Weingarten map $L: T_pM \to T_pM$ and the second fundamental form $\mathrm{II}_p: T_pM \times T_pM \to \mathbb{R}$ (as discussed in Section 5.5);

we will denote these matrices $\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix}$ and $\begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix}$ respectively, abbreviating them (L_{ij}) and (l_{ij}). We can think of the L_{ij} and the l_{ij} as functions $U \to \mathbb{R}$ for $i, j = 1, 2$; it will follow from Equations 6.2.1 and 6.2.3 below that the L_{ij} and l_{ij} are smooth functions. For ease of notation we will usually not write the variables in the L_{ij} and l_{ij}.

Since the bilinear form II is the bilinear form induced by the linear map L, it follows from Lemma 6.2.1 that

$$(l_{ij}) = (L_{ij})^t (g_{ij}),$$

noting that (g_{ij}) plays the role of the matrix G in the lemma. Since (g_{ij}) is symmetric, and using standard results about transposes of matrices, we can solve for (L_{ij}) to obtain

$$(L_{ij}) = (g_{ij})^{-1} (l_{ij})^t. \tag{6.2.1}$$

The following lemma shows how to calculate the l_{ij} (and hence the L_{ij}) in terms of the coordinate patch x. As usual n_1 and n_2 denote the partial derivatives of n with respect to the variables s and t.

Lemma 6.2.7. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point.*

(i) *(Weingarten Equations) For all $i = 1, 2$ we have*

$$n_i = -L_{1i} x_1 - L_{2i} x_2. \tag{6.2.2}$$

(ii) *For all $i, j = 1, 2$ the entries of the matrix (l_{ij}) can be computed by*

$$l_{ij} = -\langle n_i, x_j \rangle = \langle n, x_{ji} \rangle. \tag{6.2.3}$$

Proof. (i). By applying Exercise 5.7.2 to the vector field $\hat{n} = n \circ x^{-1}$, and using standard results from linear algebra, we see that

$$n_i = \frac{\partial (\hat{n} \circ x)}{\partial u_i} = \tilde{\nabla}_{x_i} \hat{n} = -L(x_i) = -L_{1i} x_1 - L_{2i} x_2,$$

where all functions are evaluated at $x^{-1}(p)$.

(ii). By the definition of the matrix of a bilinear form, and using the first part of the above computation, we have

$$l_{ij} = \mathrm{II}(x_i, x_j) = \langle L(x_i), x_j \rangle = -\langle \tilde{\nabla}_{x_i} \hat{n}, x_j \rangle = -\langle n_i, x_j \rangle,$$

which is the first equality we are trying to prove. To see the second equality, observe that $\langle n, x_j \rangle = 0$. Hence

$$0 = \frac{\partial}{\partial u_i} \langle n, x_j \rangle = \langle n_i, x_j \rangle + \langle n, x_{ji} \rangle.$$

The second equality we are proving now follows. □

The following result plays a crucial role in the study of curvature.

Lemma 6.2.8. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. The linear map L is self-adjoint, and the bilinear form* II *is symmetric.*

Proof. The symmetry of the matrix (l_{ij}) follows from Lemma 6.2.7 (ii) and the equality $x_{12} = x_{21}$. The rest of the lemma now follows using Lemma 6.2.2. □

Whereas the matrix (l_{ij}) is always symmetric, it is not guaranteed in the above lemma that the matrix (L_{ij}) will be symmetric; the basis $\{x_1, x_2\}$ is not orthonormal, so we cannot use Lemma 6.2.2. Further, using Lemma 6.2.8 we can re-write Equation 6.2.1 as

$$(L_{ij}) = (g_{ij})^{-1} (l_{ij}). \tag{6.2.4}$$

Example 6.2.9. (1) From Examples 6.2.5 (1) and 6.2.6 (1) it follows that for any coordinate patch for the plane $\mathbb{R}^2 \subset \mathbb{R}^3$, the matrices (l_{ij}) and (L_{ij}) are both zero matrices.

(2) We compute the (l_{ij}) and (L_{ij}) matrices for the saddle surface, discussed in Example 5.5.1 (2). Using the computations made in that example we see that

$$x_{11} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \quad x_{12} = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \quad x_{22} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix},$$

and

$$n = \frac{x_1 \times x_2}{\|x_1 \times x_2\|} = \frac{1}{\sqrt{1 + s^2 + t^2}} \begin{pmatrix} -t \\ -s \\ 1 \end{pmatrix}.$$

Using Lemma 6.2.7 (ii) we then have

$$l_{11} = \langle \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{1+s^2+t^2}} \begin{pmatrix} -t \\ -s \\ 1 \end{pmatrix} \rangle = 0,$$

$$l_{12} = l_{21} = \langle \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix}, \frac{1}{\sqrt{1+s^2+t^2}} \begin{pmatrix} -t \\ -s \\ 1 \end{pmatrix} \rangle = \frac{1}{\sqrt{1+s^2+t^2}},$$

$$l_{22} = \langle \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}, \frac{1}{\sqrt{1+s^2+t^2}} \begin{pmatrix} -t \\ -s \\ 1 \end{pmatrix} \rangle = 0.$$

From Equation 6.2.4 it now follows that

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} = \frac{1}{1+s^2+t^2} \begin{pmatrix} 1+s^2 & -st \\ -st & 1+t^2 \end{pmatrix} \frac{1}{\sqrt{1+s^2+t^2}} \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}$$
$$= \frac{1}{(1+s^2+t^2)^{3/2}} \begin{pmatrix} -st & 1+s^2 \\ 1+t^2 & -st \end{pmatrix}. \quad \Diamond$$

Exercises

6.2.1. Describe the Weingarten map for the cylinder $S^1 \times \mathbb{R}$.

6.2.2. For a general monge patch, as parametrized in Section 5.3, show that

$$(l_{ij}) = \frac{1}{\sqrt{1+(f_1)^2+(f_2)^2}} \begin{pmatrix} f_{11} & f_{12} \\ f_{12} & f_{22} \end{pmatrix},$$

where f_1 and f_2 are the partial derivatives of f.

6.2.3. Find (l_{ij}) and (L_{ij}) for the torus as parametrized in Section 5.3.

6.2.4. For a general surface of revolution, as parametrized in Section 5.3, show that

$$(l_{ij}) = \frac{1}{\sqrt{(r')^2+(z')^2}} \begin{pmatrix} r'z''-r''z' & 0 \\ 0 & rz' \end{pmatrix}.$$

6.2.5. For the right helicoid, as parametrized in Section 5.3, show that

$$(l_{ij}) = \frac{1}{\sqrt{b^2+t^2}}\begin{pmatrix} 0 & b \\ b & 0 \end{pmatrix} \quad \text{and} \quad (L_{ij}) = \frac{1}{\sqrt{b^2+t^2}}\begin{pmatrix} 0 & \frac{b}{b^2+t^2} \\ b & 0 \end{pmatrix}.$$

6.2.6. For a general rectifying developable surface, as parametrized in Section 5.3, show that

$$(l_{ij}) = \frac{1}{\sqrt{1+t^2\tau^2(s)}}\begin{pmatrix} 0 & -\tau(s) \\ -\tau(s) & \kappa(s)+t^2\kappa(s)\tau^2(s) \end{pmatrix},$$

where $\kappa(s)$ and $\tau(s)$ are the curvature and torsion of the curve c.

6.2.7*. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $Z\colon M \to \mathbb{R}^3$ be a smooth vector field such that $Z(p)$ is orthogonal to T_pM for all $p \in M$. If $p \in M$ is a point and $v \in T_pM$ is a vector, show that

$$\mathrm{II}(v, v) = -\frac{1}{\|Z(p)\|}\langle DZ(p)\, v, v\rangle.$$

6.2.8*. Let M and F be as in Exercise 5.9.10. If $p \in M$ is a point, let L_1 denote the Weingarten map of M at p and let L_2 denote the Weingarten map of $F(M)$ at $F(p)$. Show that $L_2(v) = R\,L_1(R^{-1}v)$ for all $v \in T_{F(p)}F(M)$.

6.3 Curvature — Second Attempt

The definition of curvature is now easy, using the tools we have developed. The Weingarten map summarizes how the normal vector field is changing at each point in the surface. Since curvature should assign a single number to every point on the surface, we need to squeeze a single number out of the Weingarten map; the determinant and trace are the two obvious candidates to get a number out of a linear map. (Recall that the concepts of "determinant" and "trace" apply to linear maps, not just matrices.)

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. We define two functions $K, H\colon M \to \mathbb{R}$ as follows. For each point $p \in M$ let L be the Weingarten map of M at p, and define

$$K(p) = \det L \quad \text{and} \quad H(p) = \frac{1}{2} \operatorname{tr} L.$$

The number $K(p)$ is called the **Gaussian curvature** at p, and the quantity $H(p)$ is called the **mean curvature** at p. ◊

To see whether the choice of Weingarten map (which is only defined up to $\pm$) has any effect on the above definition, note that if $F: \mathbb{R}^2 \to \mathbb{R}^2$ is a linear map then $\det(-F) = \det F$ and $\operatorname{tr}(-F) = -\operatorname{tr} F$ (this uses the even-dimensionality of the vector space $\mathbb{R}^2$). It follows that K is independent of the choice of Weingarten map, whereas H is only defined up to $\pm$.

According to Lemma 6.2.8 the linear map L is self-adjoint at each point p in the surface. Using the finite-dimensional spectral theorem for self-adjoint linear maps (see for example [LA1, p. 193]) we deduce that L has two real eigenvalues (not necessarily distinct), denoted k_1 and k_2; we label the eigenvalues so that $k_1 \geq k_2$. (The numbers k_1 and k_2 are really functions of M, but we drop the arguments.) If $k_1 = k_2$ then all vectors in T_pM are eigenvectors; if $k_1 \neq k_2$, then by the spectral theorem the eigenvectors for k_1 and k_2 are orthogonal. We then deduce from the definition of determinant and trace that

$$K(p) = k_1 k_2 \quad \text{and} \quad H(p) = \frac{1}{2}(k_1 + k_2).$$

Since these eigenvalues are of such significance geometrically, we give them a name.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. The eigenvalues k_1 and k_2 of the Weingarten map of M at p are called the **principal curvatures** of M at p, and their eigenvectors are called the **principal directions** of M at p. ◊

Both Gaussian and mean curvature were defined using the Weingarten map, which depends upon the way in which the surface sits in $\mathbb{R}^3$. Thus, neither type of curvature appears at first glance to be intrinsic, which was one of the properties we postulated for curvature in Section 6.1. Remarkably, however, Gaussian curvature turns out to be intrinsic even though it is not intrinsically defined. Gauss was so taken by this fact that he called it the Theorema Egregium, which means the outstanding (or remarkable) theorem in Latin. We will prove this theorem in Section 6.5. Mean curvature is definitely not intrinsic. Though mean curvature is still quite useful, Gaussian curvature is universally accepted as "the

curvature" for surfaces; if one encounters the unadorned word "curvature" in reference to surfaces, the reference is virtually always to Gaussian curvature. We will stick to Gaussian curvature from now on, except in some examples of calculations.

We can now compute Gaussian and mean curvature in a few simple cases.

Example 6.3.1. We continue Example 6.2.5. (1) The Weingarten map in this case is the zero map, and both its determinant and trace are zero. Therefore the Gaussian and mean curvatures are both zero for a point with a flat open neighborhood.

(2) The Weingarten map at each point is $\frac{1}{R}$ times the identity map, and the determinant of this map is $\frac{1}{R^2}$ and the trace is $\frac{2}{R}$. Hence the Gaussian curvature of any point on the sphere of radius R is $\frac{1}{R^2}$, and the mean curvature is $\frac{1}{R}$. As expected, a sphere of larger radius has smaller curvature, and as the radius goes to infinity the curvature goes to zero (which is reasonable, since as the radius goes to infinity the sphere looks locally more and more like a plane). $\Diamond$

To get a better feel for Gaussian curvature let us return to our original attempt at defining curvature in Section 6.1 via curves through a point p in a surface $M \subset \mathbb{R}^3$. We considered curves obtained by intersecting the surface M with planes $\Omega \subset \mathbb{R}^3$ that contained p and were parallel to $\hat{n}(p)$. To compute the planar curvature of these curves we will need to use oriented planes. Observe that oriented planes containing p and parallel to $\hat{n}(p)$ are in one-to-one correspondence with unit vectors in T_pM. For any such vector v, let Ω_v be the oriented plane in $\mathbb{R}^3$ containing p and parallel to the plane spanned by the vectors $\{v, \hat{n}(p)\}$, with the orientation of Ω_v given by the ordered basis $\{v, \hat{n}(p)\}$ (that is, we consider a rotation of the plane taking v to $\hat{n}(p)$ to be counterclockwise). Conversely, for any oriented plane Ω containing p and parallel to $\hat{n}(p)$, let v_Ω be the unique unit vector in T_pM such that Ω is parallel to the plane spanned by the vectors $\{v_\Omega, \hat{n}(p)\}$, and such that the oriented basis $\{v_\Omega, \hat{n}(p)\}$ corresponds to the orientation of Ω.

Let Ω be an oriented plane containing p and parallel to $\hat{n}(p)$. By Exercise 6.1.1 we know that there is some injective smooth curve $c_\Omega\colon (-\epsilon, \epsilon) \to M$ for some number $\epsilon > 0$ such that $c_\Omega((-\epsilon, \epsilon))$ is an open subset of $\Omega \cap M$ and $c_\Omega(0) = p$. Without loss of generality we may assume that the curve c_Ω is unit speed, and that $c_\Omega'(0) = v_\Omega$. Since c_Ω is contained in Ω it is a planar curve, and thus we can compute its planar curvature $\bar{\kappa}$ using the method of Section 4.7, making use of the orientation of Ω. The following proposition, the importance

of which becomes apparent with the subsequent theorem, relates the geometry of M and the curvature $\bar{\kappa}(0)$.

Proposition 6.3.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. If Ω is an oriented plane containing p and parallel to $\hat{n}(p)$, and if $c_\Omega: (-\epsilon, \epsilon) \to M$ is an injective smooth curve such that $c_\Omega((-\epsilon, \epsilon))$ is an open subset of $\Omega \cap M$ and $c_\Omega(0) = p$, then*

$$\mathrm{II}_p(v_\Omega, v_\Omega) = \bar{\kappa}(0),$$

where $\bar{\kappa}$ is the planar curvature of c_Ω.

Proof. By choosing $\epsilon > 0$ small enough we may assume that the image of c_Ω lies in $x(U)$. By the definition of $\hat{n}$ we see that

$$\langle \hat{n} \circ c_\Omega(t), c_\Omega'(t) \rangle = 0$$

for all $t \in (-\epsilon, \epsilon)$. Taking the derivative of both sides of this equation and rearranging we have

$$\langle (\hat{n} \circ c_\Omega)'(t), c_\Omega'(t) \rangle = - \langle \hat{n} \circ c_\Omega(t), c_\Omega''(t) \rangle. \tag{6.3.1}$$

Next, combining the fact that c_Ω is unit speed with Equation 4.7.1 we know

$$c_\Omega''(t) = \bar{T}'(t) = \bar{\kappa}(t)\bar{N}(t). \tag{6.3.2}$$

Because rotation from v_Ω to $\hat{n}(p)$ is considered counterclockwise, we have $\bar{N}(0) = \hat{n}(p)$. Thus

$$c_\Omega''(0) = \bar{\kappa}(0)\hat{n}(p). \tag{6.3.3}$$

Using the definition of the directional derivative and the above observations we now compute

$$\begin{aligned} \mathrm{II}_p(v_\Omega, v_\Omega) &= \langle L(v_\Omega), v_\Omega \rangle = \langle -\tilde{\nabla}_{v_\Omega} \hat{n}, v_\Omega \rangle = - \langle \hat{n} \circ c_\Omega'(0), c_\Omega'(0) \rangle \\ &= \langle \hat{n} \circ c_\Omega(0), c_\Omega''(0) \rangle = \langle \hat{n}(p), \bar{\kappa}(0)\hat{n}(p) \rangle = \bar{\kappa}(0). \quad \square \end{aligned}$$

The following theorem gives the relation of the curvature of curves in a surface to the curvature of the surface itself.

Theorem 6.3.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. The principal curvatures k_1 and k_2 of M at p are the maximum and minimum values of $II_p(v, v)$ over all unit vectors $v \in T_pM$. If v is any such vector, then*

$$II_p(v, v) = k_1 \cos^2\theta + k_2 \sin^2\theta,$$

where θ is the angle between v and the eigenvalue for k_1. If $k_1 \neq k_2$, the eigenvalues for k_1 and k_2 are the only critical points for $II_p(v, v)$ thought of as a function of θ.

The formula for $II_p(v, v)$ in this theorem is called Euler's formula. Combining Theorem 6.3.3 with Proposition 6.3.2 we see that k_1 is the maximal planar curvature of any of the curves c_Ω, and k_2 is the minimal such curvature. Thus the Gaussian and mean curvatures can be calculated from the planar curvatures of curves in the surface by multiplying and averaging respectively the maximal and minimal such curvatures. We have as a consequence of Theorem 6.3.3 and Proposition 6.3.2 the somewhat surprising fact that for any point in a smooth surface the maximal and minimal planar curvature of the curves c_Ω always occur at perpendicular curves (unless the planar curvatures of all c_Ω are equal).

Proof of Theorem 6.3.3. Let e_1 and e_2 be eigenvectors for k_1 and k_2 respectively. Since the linear map L is self-adjoint we may choose e_1 and e_2 to be orthogonal unit vectors (whether or not k_1 and k_2 are distinct). Give T_pM an orientation. By replacing e_2 with $-e_2$ if necessary, we may assume that $\{e_1, e_2\}$ determines the given orientation of T_pM. Any unit vector $v \in T_pM$ can thus be written as $v = \cos\theta\, e_1 + \sin\theta\, e_2$, where $\theta \in (-\pi, \pi]$ is the signed angle from e_1 to v. We then have

$$II_p(v, v) = \langle L(v), v\rangle = k_1 \cos^2\theta + k_2 \sin^2\theta,$$

where the last equality is obtained using the expression for v in terms of θ and the orthonormality of the basis $\{e_1, e_2\}$.

We can now view $II_p(v, v)$ as a function of θ, with domain all $\mathbb{R}$; this function is periodic with period 2π. To find the extrema of $II_p(v, v)$ we simply have to look at the critical points with respect to θ. We compute

$$\frac{d}{d\theta} II_p(v, v) = 2(k_1 - k_2) \sin\theta \cos\theta.$$

If $k_1 = k_2$ then $II_p(v, v)$ is constant, so every v is both a maximum point and a minimum point, and there is nothing to prove. Now assume that $k_1 \neq k_2$.

The extrema can only occur at $\theta = \frac{k\pi}{2}$ for $k \in \mathbb{Z}$. However, $\theta = k\pi$ for all $k \in \mathbb{Z}$ correspond to $v = \pm e_1$, and $\theta = \frac{(2k+1)\pi}{2}$ for all $k \in \mathbb{Z}$ correspond to $v = \pm e_2$. Substituting $\theta = 0$ into Euler's formula we have $\mathrm{II}_p(e_1, e_1) = k_1$, and substituting $\theta = \pi/2$ into Euler's formula we have $\mathrm{II}_p(e_2, e_2) = k_2$. Using the periodicity of $\mathrm{II}_p(v, v)$ as a function of θ it follows that k_1 and k_2 are the maximum and minimum values, respectively, of $\mathrm{II}_p(v, v)$. □

Example 6.3.4. (1) It is not hard to verify that at any point on the right circular cylinder $S^1 \times \mathbb{R}$ the maximal and minimal curvatures of curves of the form c_Ω occur at horizontal curves (with curvature 1) and at vertical curves (with curvature zero). Thus the maximal and minimal curvatures are indeed in perpendicular directions.

(2) The monkey saddle is the graph of the function $z = x^3 - 3xy^2$ (see Figure 6.3.1.). We want to compute the curvature of the monkey saddle at the origin. Suppose the function $\mathrm{II}_{O_3}(v, v)$ were not constantly zero. By the Extreme Value Theorem it would have a maximum value and a minimum value that are distinct, and by differentiability these extrema must occur at critical points. In Figure 6.3.1 there are three equally spaced straight lines through the origin contained in the surface. These three lines are curves of the form c_Ω for appropriate choices of Ω, and hence by Proposition 6.3.2 the function $\mathrm{II}_{O_3}(v, v)$, defined over all unit vectors $v \in T_{O_3}M$, equals zero for the six unit vectors in $T_{O_3}M$ along the three lines. At least one of the extrema of $\mathrm{II}_{O_3}(v, v)$ must occur strictly between two of these unit vectors. Since the six vectors divide the surface into six pieces, identical except for sign, then $\mathrm{II}_{O_3}(v, v)$ must therefore have at least six critical points, impossible by Theorem 6.3.3. Hence $\mathrm{II}_{O_3}(v, v)$ must be constantly zero for all unit vectors in $T_{O_3}M$. It follows from Theorem 6.3.3 that $k_1 = k_2 = 0$, and hence $K(O_3) = H(O_3) = 0$.

This result may seem somewhat counterintuitive, given that the monkey saddle seems to be "curving" a fair bit at the origin. However, the nature of smoothness forces a surface that has as many ups and downs as the monkey saddle does at the origin to flatten out so much that the Gaussian and mean curvatures must be zero. (In the simplicial case, by contrast, no such restriction occurs, and a simplicial monkey saddle would have non-zero curvature at the origin.) ◊

Theorem 6.3.3 helps us gain insight into the meaning of the sign of K. If $K(p) > 0$ then either $k_1, k_2 > 0$ or $k_1, k_2 < 0$; this means that the curves through p corresponding to the two principal directions both either curve in the

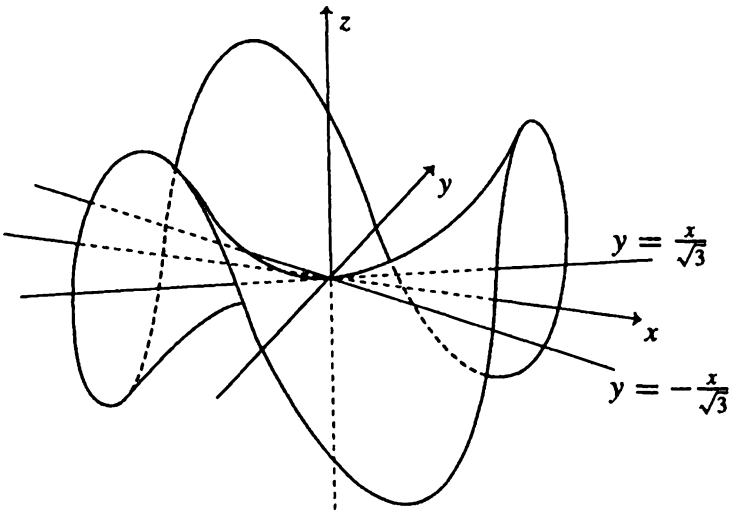


Figure 6.3.1

direction of $\hat{n}(p)$ or curve away from it. See Figure 6.3.2 (i). If $K(p) < 0$, then either $k_1 > 0$ and $k_2 < 0$ or vice-versa; this means that one of the curves through p corresponding to a principal direction curves in the direction of $\hat{n}(p)$ and the other curves away from $\hat{n}(p)$. See Figure 6.3.2 (ii). Finally, if $K(p) = 0$, then at least one of the principal curvatures is zero, though not necessarily both. Note, however, that the curve through p corresponding to a principal curvature with value zero need not be a straight line, it just needs to have zero planar curvature. See Figure 6.3.2 (iii).

Exercises

6.3.1. Find the Gaussian and mean curvatures for the cylinder $S^1 \times \mathbb{R}$.

6.3.2. Let $c\colon [a, b] \to \mathbb{R}^2$ be a smooth curve. The **generalized right cylinder** with cross section c is the surface $M \subset \mathbb{R}^3$ given by

$$M = \left\{ \begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3 \mid \begin{pmatrix} x \\ y \end{pmatrix} = c(s) \text{ for some } s \in [a, b] \right\}.$$

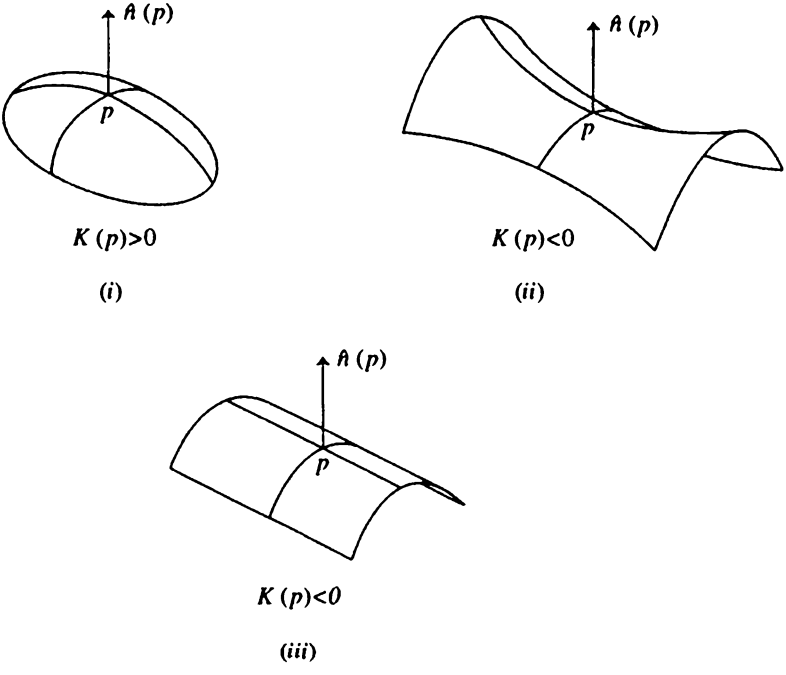


Figure 6.3.2

See Figure 6.3.3. (Observe that a generalized right cylinder is a special case of a ruled surface, where the rulings are all parallel to the z-axis.) Compute the Gaussian curvature at all points of a generalized cylinder.

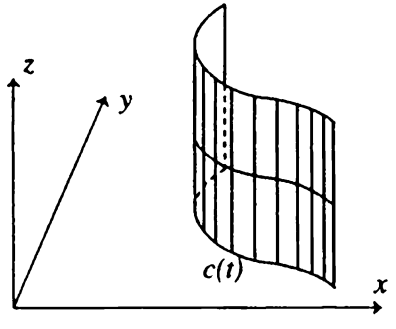


Figure 6.3.3

6.3.3. The "dog saddle" is the surface given by $z = 4x^3y - 4xy^3$. Sketch the graph of this surface. What is the curvature of the surface at the origin?

6.3.4. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. The point p is called **umbilic** if all vectors in T_pM are principal directions. For example, on a sphere all points are umbilic. (See [RT, index] for a tangential comment on the term "umbilic.") Show that if p is an umbilic point then there is some real number k such that the following properties hold.

(i) The Weingarten map at p is multiplication by the scalar k.

(ii) For any coordinate patch $x: U \to M$ we have $l_{ij}(\bar{p}) = kg_{ij}(\bar{p})$ and $n_i(\bar{p}) = -kx_i(\bar{p})$ for $i, j = 1, 2$ (where as usual $\bar{p} = x^{-1}(p)$).

6.3.5. The goal of this exercise is to compute the Gaussian and mean curvature functions on level surfaces (as defined in Section 5.3). More specifically, let $M \subset \mathbb{R}^3$ be a smooth surface given as $M = F^{-1}(a)$ for some smooth function $F: V \to \mathbb{R}$, where $V \subset \mathbb{R}^3$ is an open set and $a \in \mathbb{R}$ is a number. For this method to work we will assume additionally that all the mixed second partial derivatives of F are constantly zero, that is $F_{ij} = 0$ for $i \neq j$ (this assumption still allows us to handle the standard quadric surfaces). Our treatment follows [SK3 vol. III, Chapter 3]. Fix a point $p \in M$.

(1) Let $v \in T_pM$ be a vector; suppose $v = \begin{pmatrix} v_1 \\ v_2 \\ v_3 \end{pmatrix} \in \mathbb{R}^3$. Show that

$$\mathrm{II}_p(v, v) = -\frac{(v_1)^2 F_{11}(p) + (v_2)^2 F_{22}(p) + (v_3)^2 F_{33}(p)}{\|DF(p)\|}.$$

(2) Using Theorem 6.3.3 we can find the principal curvatures of M at p by finding the maximum and minimum values of $\mathrm{II}_p(v, v)$ with respect to the variables v_1, v_2 and v_3, subject to the constraint that v is a unit tangent vector; this latter condition is expressed by the equations $\|v\|^2 = 1$ and $\langle v, (DF(p))^t \rangle = 0$. To find the extrema of $\mathrm{II}_p(v, v)$ we use the method of Lagrange multipliers, which states that the extrema will occur at the solutions to the three equations

$$D(\mathrm{II}_p(v, v)) = \lambda\, D(\|v\|^2) + \mu\, D(\langle v, (DF(p))^t \rangle)$$
$$\|v\|^2 = 1$$
$$\langle v, (DF(p))^t \rangle = 0,$$

where the derivatives in the first equation are with respect to the v_i. Show that the extrema of $\mathrm{II}_p(v, v)$ occur at v_1, v_2, v_3, λ and μ satisfying

$$\begin{pmatrix} (F_{11}-\lambda) & 0 & 0 & F_1 \\ 0 & (F_{22}-\lambda) & 0 & F_2 \\ 0 & 0 & (F_{33}-\lambda) & F_1 \\ F_1 & F_2 & F_3 & 0 \end{pmatrix} \begin{pmatrix} v_1 \\ v_2 \\ v_3 \\ -\frac{\mu}{2} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 0 \end{pmatrix}, \tag{6.3.4}$$

where for convenience we drop the argument p in the partial derivatives of F.

(3) Show that

$$\lambda = (v_1)^2 F_{11} + (v_2)^2 F_{22} + (v_3)^2 F_{33} \tag{6.3.5}$$

for any solution of Equation 6.3.4.

(4) If λ_1 and λ_2 denote the maximal and minimal values of λ, show that the maximal and minimal values of $\mathrm{II}_p(v, v)$ are

$$k_1 = -\frac{\lambda_1}{\|DF\|} \quad \text{and} \quad k_2 = -\frac{\lambda_2}{\|DF\|}.$$

(5) Show that the matrix in Equation 6.3.4 must have zero determinant, and derive a quadratic equation for λ. Find $\lambda_1\lambda_2$ and $\lambda_1 + \lambda_2$ using facts about the relation between the solutions of a quadratic equation and the coefficients of the equation. Deduce that

$$K(p) = \frac{(F_1)^2 F_{22} F_{33} + (F_2)^2 F_{11} F_{33} + (F_3)^2 F_{11} F_{22}}{\left((F_1)^2 + (F_2)^2 + (F_3)^2\right)^2}$$

$$H(p) = \frac{(F_1)^2(F_{22} + F_{33}) + (F_2)^2(F_{11} + F_{33}) + (F_3)^2(F_{11} + F_{22})}{2\left((F_1)^2 + (F_2)^2 + (F_3)^2\right)^{3/2}}.$$

6.2.6. Find formulas for the Gaussian and mean curvature functions of

(1) the ellipsoid $\dfrac{x^2}{a^2} + \dfrac{y^2}{b^2} + \dfrac{z^2}{c^2} = 1$;

(2) the hyperbolic paraboloid $\dfrac{x^2}{a^2} - \dfrac{y^2}{b^2} - z = 0$.

6.2.7*. Let M and F be as in Exercises 5.9.10 and 6.2.8. Show that $K(F(p)) = K(p)$ and $H(F(p)) = H(p)$ for each point $p \in M$, where the curvatures at $F(p)$ are computed for the surface $F(M)$, and the curvatures at p are computed for the surface M.

6.4 Computations of Curvature Using Coordinates

Our discussion of Gaussian and mean curvature so far has not involved the use of coordinate patches (other than to choose a well-defined Gauss map), to emphasize that our curvature functions do not depend upon coordinate patches. To compute the Gaussian or mean curvature for any but the simplest surfaces, however, coordinate patches are definitely needed. In principle we could compute K and H by simply taking the determinant and half the trace of (L_{ij}), which is just the matrix for L with respect to the basis $\{x_1, x_2\}$, and which we saw how to compute in Section 6.2. To simplify matters we use the following notation and lemma (though we will at times revert to our previous notation when we need to use summation notation). As usual, let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let $x: U \to M$ be a coordinate patch such that $p \in x(U)$. We let

$$E = g_{11}, \quad F = g_{12} = g_{21}, \quad G = g_{22}$$

and

$$A = l_{11}, \quad B = l_{12} = l_{21}, \quad C = l_{22}.$$

Recall from Lemma 5.5.2 that $EG - F^2 = \|x_1 \times x_2\|^2 \neq 0$. The following proposition is our main computational tool for curvature of surfaces.

Proposition 6.4.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. Then*

$$K(p) = \frac{AC - B^2}{EG - F^2}, \tag{6.4.1}$$

and

$$H(p) = \frac{EC - 2FB + GA}{2(EG - F^2)}, \tag{6.4.2}$$

where $A, \ldots, G$ are evaluated at $x^{-1}(p)$.

Proof. We will drop all arguments p and $\bar{p}$ from our equations. By Equation 6.2.4 we have

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} = \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} \begin{pmatrix} A & B \\ B & C \end{pmatrix}. \tag{6.4.3}$$

Hence

$$K = \det L = \det \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1} \det \begin{pmatrix} A & B \\ B & C \end{pmatrix} = \frac{AC - B^2}{EG - F^2}.$$

The formula for H is obtained by multiplying the two matrices in Equation 6.4.3 together after taking the appropriate inverse, and then taking half the trace. □

Because the quantities E, F, G, A, B and C are all smooth functions $U \to \mathbb{R}$, so are the functions K and H. For convenience we now summarize all the formulas needed for calculating the curvature of smooth surfaces:

$$\begin{gathered} E = \langle x_1, x_1 \rangle, \quad F = \langle x_1, x_2 \rangle, \quad G = \langle x_2, x_2 \rangle \\ n = \frac{x_1 \times x_2}{\|x_1 \times x_2\|} = \frac{x_1 \times x_2}{\sqrt{EG - F^2}} \\ A = \langle n, x_{11} \rangle, \quad B = \langle n, x_{12} \rangle, \quad C = \langle n, x_{22} \rangle \\ K = \frac{AC - B^2}{EG - F^2}, \quad H = \frac{EC - 2FB + GA}{2(EG - F^2)}. \end{gathered} \tag{6.4.4}$$

Now comes the payoff — we can compute the curvature of a variety of surfaces almost mechanically.

Example 6.4.2. We compute the curvature of the right helicoid, using the parametrization given in Section 5.3. We have

$$x_1 = \begin{pmatrix} -t \sin s \\ t \cos s \\ b \end{pmatrix}, \quad x_2 = \begin{pmatrix} \cos s \\ \sin s \\ 0 \end{pmatrix}$$

$$x_{11} = \begin{pmatrix} -t \cos s \\ -t \sin s \\ 0 \end{pmatrix}, \quad x_{12} = \begin{pmatrix} -\sin s \\ \cos s \\ 0 \end{pmatrix}, \quad x_{22} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}$$

$$E = b^2 + t^2, \quad F = 0, \quad G = 1$$

$$n = \frac{1}{\sqrt{b^2 + t^2}} \begin{pmatrix} -b \sin s \\ b \cos s \\ -t \end{pmatrix}$$

$$A = 0, \quad B = \frac{b}{\sqrt{b^2 + t^2}}, \quad C = 0$$

$$K = -\frac{b^2}{(b^2 + t^2)^2}, \quad H = 0.$$

Observe that the curvature is always negative (since b is assumed positive), that the maximal curvature occurs at $t = 0$ (which is along the z-axis), and that as t goes to infinity the curvature goes to zero. Since H is constantly zero the right helicoid is a minimal surface. $\Diamond$

We now have the tools to ascertain that our definition of Gaussian curvature is equivalent to Gauss' original definition, as discussed in Section 6.1. Given that our discussion of Gauss' original definition was not entirely rigorous (especially when it came to the limiting process involved), we cannot be entirely rigorous here either. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in M$ be a point and let $\tilde{p} = x^{-1}(p)$. Recall that Gauss' original definition of curvature was

$$\lim_{T \to \{p\}} \frac{\text{Area}_0(\hat{n}(T))}{\text{Area}(T)} = \lim_{T \to \{p\}} \frac{\iint_{x^{-1}(T)} \langle n_1 \times n_2, n \rangle \, ds \, dt}{\iint_{x^{-1}(T)} \sqrt{\det(g_{ij})} \, ds \, dt},$$

where the terms were defined in Section 6.1. To evaluate this limit (which is reminiscent of L'Hôpital's rule) we use the Mean Value Theorem for multiple integrals (see [BT, §24]). From this theorem it follows that for each set T there must be points $a_T, b_T \in x^{-1}(T)$ such that

$$\iint_{x^{-1}(T)} \langle n_1 \times n_2, n \rangle \, ds \, dt = \langle n_1(a_T) \times n_2(a_T), n(a_T) \rangle \, \text{Area}(x^{-1}(T)),$$

$$\iint_{x^{-1}(T)} \sqrt{\det(g_{ij})} \, ds \, dt = \sqrt{\det(g_{ij}(b_T))} \, \text{Area}(x^{-1}(T)).$$

Observing that as $T \to \{p\}$ then $a_T \to \tilde{p}$ and $b_T \to \tilde{p}$, and making use of Exercise 6.4.5, we now see that

$$\lim_{T \to \{p\}} \frac{\text{Area}_0(\hat{n}(T))}{\text{Area}(T)} = \lim_{T \to \{p\}} \frac{\langle n_1(a_T) \times n_2(a_T), n(a_T) \rangle \, \text{Area}(x^{-1}(T))}{\sqrt{\det(g_{ij}(b_T))} \, \text{Area}(x^{-1}(T))}$$

$$= \frac{\langle n_1(\tilde{p}) \times n_2(\tilde{p}), n(\tilde{p}) \rangle}{\sqrt{\det(g_{ij}(\tilde{p}))}} = \frac{K(p)\sqrt{\det(g_{ij}(\tilde{p}))}}{\sqrt{\det(g_{ij}(\tilde{p}))}} = K(p).$$

Exercises

6.4.1*. For the catenoid (see Exercise 5.5.3) show that

$$K = \frac{-1}{(1+t^2)^2}, \quad H = 0.$$

6.4.2. Show that for a general monge patch (as parametrized in Section 5.3)

$$K = \frac{f_{11}f_{22} - (f_{12})^2}{\big(1 + (f_1)^2 + (f_2)^2\big)^2}$$
$$H = \frac{\big(1 + (f_1)^2\big)f_{22} - f_1 f_2 f_{12} + \big(1 + (f_2)^2\big)f_{11}}{2\big(1 + (f_1)^2 + (f_2)^2\big)^{3/2}}.$$

6.4.3*. Show that for a general surface of revolution (as parametrized in Section 5.3)

$$K = \frac{z'(r'z'' - r''z')}{r\big((r')^2 + (z')^2\big)^2}$$
$$H = \frac{r(r'z'' - r''z') + z'((r')^2 + (z')^2)}{2r\big((r')^2 + (z')^2\big)^{3/2}}.$$

6.4.4*. For the surface in Exercise 5.5.4 show that

$$K = -\frac{1}{(1+t^2)^2}$$
$$H = \frac{t}{2(1+t^2)^{3/2}}.$$

6.4.5*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in M$ be a point and let $\bar{p} = x^{-1}(p)$. Show that

$$\langle n_1(\bar{p}) \times n_2(\bar{p}), n(\bar{p})\rangle = K(p)\sqrt{\det\big(g_{ij}(\bar{p})\big)}.$$

6.4.6*. Suppose that in a coordinate patch with connected domain we know that $A = B = C = 0$ at all points in the domain of the coordinate patch. Show that the image of the coordinate patch is contained in a plane.

6.4.7*. Suppose that $K = H = 0$ everywhere on a connected smooth surface. Show that the Weingarten map at all points is the zero map, and that the surface is contained in a plane. Give an example to show that $K = 0$ alone at all points does not suffice to imply that the surface is contained in a plane.

6.4.8. Suppose that all points on a given connected surface M are umbilic (see Exercise 6.3.4). Show that the following hold.

(i) The value k as in Exercise 6.3.4 is a constant over the whole surface.

(ii) If $k = 0$ the surface is contained in a plane.

(iii) If $k \neq 0$ the surface is contained in a sphere of radius $\frac{1}{|k|}$.

6.4.9*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in x(U)$ be a point and let $v \in T_pM$ be a vector. Let $\bar{p} = x^{-1}(p)$, and write v in coordinates as $v = v^1 x_1(\bar{p}) + v^2 x_2(\bar{p})$. Show that v is a principal direction of M at p iff

$$(EB - FA)(v^1)^2 + (EC - GA)v^1v^2 + (FC - GB)(v^2)^2 = 0,$$

where all the functions are evaluated at $\bar{p}$.

6.4.10*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch, let $p \in x(U)$ be a non-umbilic point (see Exercise 6.3.4) and let $\bar{p} = x^{-1}(p)$. Show that $x_1(\bar{p})$ and $x_2(\bar{p})$ are principal directions of M at p iff $F(\bar{p}) = 0 = B(\bar{p})$.

6.4.11*. Let $M \subset \mathbb{R}^3$ be a compact smooth surface. The goal of this exercise is to show that M has positive Gaussian curvature on some non-empty open subset. (This result is not true for surfaces in higher-dimensional Euclidean space.)

(1) Suppose we can show that there is a single point on M with positive Gaussian curvature. Use the continuity of K to show that K must be positive on a non-empty open subset of M.

(2) The rest of the proof is to find a single point of positive Gaussian curvature. Define $f: M \to \mathbb{R}$ by $f(x) = \langle x, x \rangle$. Show that f takes on a maximal value at some point $q \in M$.

(3) Let $x: U \to M$ be a coordinate patch containing q in its image, and let $\bar{q} = x^{-1}(q)$. It follows from Exercise 5.9.9 that we can choose x so that $x_1(\bar{q})$ and $x_2(\bar{q})$ are principal directions of M at q. Show that

$$(f \circ x)_1(\bar{q}) = 0 = (f \circ x)_2(\bar{q}),$$
$$(f \circ x)_{11}(\bar{q}) \leq 0, \quad (f \circ x)_{22}(\bar{q}) \leq 0,$$

where as usual the subscripts indicate partial derivatives.

(4) Show that $\|x(\bar{q})\| > 0$ and that

$$n(\bar{q}) = \pm \frac{x(\bar{q})}{\|x(\bar{q})\|}.$$

Assume without loss of generality that there is a plus in the above equation.

(5) Let $R = \|x(\bar{q})\|$. Show that

$$\frac{A(\bar{q})}{E(\bar{q})} \cdot \frac{C(\bar{q})}{G(\bar{q})} \leq -\frac{1}{R}.$$

(6) Deduce that $K(q) \geq \frac{1}{R^2} > 0$.

6.5 Theorema Egregium and the Fundamental Theorem of Surfaces

The two main results in this chapter, which we now present, are the Theorema Egregium and the Fundamental Theorem of Surfaces. Our discussion of both theorems relies upon formulas known as the Gauss Equation and the Codazzi–Mainardi Equations, which are given in Theorem 6.6.2. To prove this theorem, we start by finding the analog for surfaces of the Frenet–Serret Theorem (Theorem 4.5.5).

The Frenet–Serret Theorem tells us the derivatives of the vectors T, N and B. The best analogs we have for surfaces of T, N and B are the vectors x_1, x_2 and n, though the latter are not orthonormal. Note also that x_1, x_2 and n are only defined for a coordinate patch, not for the whole surface, and that they very much depend on the choice of coordinate patch; thus we will restrict our attention to a single coordinate patch $x: U \to M$. Since x_1, x_2 and n are functions of two variables, we are looking for formulas that express the partial

derivatives of x_1, x_2 and n as linear combinations of these three vectors. We already have one such formula, namely the Weingarten equations from Lemma 6.2.7, which gives the partial derivatives of n. The following proposition gives us the desired formulas for x_1 and x_2. As usual we will let x_{ij} denote the jth partial derivative of x_i; we will most often drop the arguments in all functions defined on U.

Theorem 6.5.1 (Gauss Formulas). *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x: U \to M$ be a coordinate patch. Then for all $i, j = 1, 2$*

$$x_{ij} = \Gamma_{ij}^1 x_1 + \Gamma_{ij}^2 x_2 + l_{ij} n. \tag{6.5.1}$$

Proof. Since $\{x_1, x_2, n\}$ form a basis for $\mathbb{R}^3$, there must be unique numbers P_{ij}^k and Q_{ij} such that

$$x_{ij} = P_{ij}^1 x_1 + P_{ij}^2 x_2 + Q_{ij} n, \tag{6.5.2}$$

for each $i, j = 1, 2$. Taking the inner product of both sides of this equation with each of the three basis vectors yields

$$\langle x_{ij}, x_1 \rangle = P_{ij}^1 g_{11} + P_{ij}^2 g_{12} \tag{6.5.3}$$

$$\langle x_{ij}, x_2 \rangle = P_{ij}^1 g_{21} + P_{ij}^2 g_{22} \tag{6.5.4}$$

$$\langle x_{ij}, n \rangle = Q_{ij}. \tag{6.5.5}$$

It follows from Equations 6.5.5 and 6.2.3 that $Q_{ij} = l_{ij}$. To solve for the P_{ij}^k let us rewrite Equations 6.5.3 and 6.5.4 in matrix form:

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} P_{ij}^1 \\ P_{ij}^2 \end{pmatrix} = \begin{pmatrix} \langle x_{ij}, x_1 \rangle \\ \langle x_{ij}, x_2 \rangle \end{pmatrix}.$$

Using Exercise 5.5.9 we have

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} P_{ij}^1 \\ P_{ij}^2 \end{pmatrix} = \frac{1}{2} \begin{pmatrix} \frac{\partial g_{j1}}{\partial u_i} + \frac{\partial g_{i1}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_1} \\ \frac{\partial g_{j2}}{\partial u_i} + \frac{\partial g_{i2}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_2} \end{pmatrix}.$$

Multiplying both sides of this last equation by $(g_{ij})^{-1}$, and then using Lemma 5.7.2, we deduce that $P_{ij}^k = \Gamma_{ij}^k$, which is precisely what we needed to show. □

We now turn to the main technical tool of this section. The equations in this result, which may appear somewhat unmotivated, are derived from the equality of mixed third partial derivatives.

Theorem 6.5.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $x: U \to M$ be a coordinate patch. Then the following equations hold.*

(i) (Gauss Equation)

$$l_{11}l_{22} - (l_{12})^2 = \sum_{r=1}^{2} g_{1r}\left\{\frac{\partial \Gamma_{22}^r}{\partial u_1} - \frac{\partial \Gamma_{21}^r}{\partial u 1_2} + \sum_{m=1}^{2}\left(\Gamma_{22}^m \Gamma_{m1}^r - \Gamma_{21}^m \Gamma_{m2}^r\right)\right\}. \quad (6.5.6)$$

(ii) (Codazzi–Mainardi Equations)

$$\frac{\partial l_{12}}{\partial u_1} - \frac{\partial l_{11}}{\partial u_2} + \sum_{r=1}^{2}\left(\Gamma_{12}^r l_{r1} - \Gamma_{11}^r l_{r2}\right) = 0. \quad (6.5.7)$$

$$\frac{\partial l_{22}}{\partial u_1} - \frac{\partial l_{21}}{\partial u_2} + \sum_{r=1}^{2}\left(\Gamma_{22}^r l_{r1} - \Gamma_{21}^r l_{r2}\right) = 0. \quad (6.5.8)$$

Proof. Starting with the Gauss formulas (Equation 6.5.1) and taking the partial derivative with respect to u_k of both sides of the equation yields

$$x_{ijk} = \frac{\partial \Gamma_{ij}^1}{\partial u_k} x_1 + \Gamma_{ij}^1 x_{1k} + \frac{\partial \Gamma_{ij}^2}{\partial u_k} x_2 + \Gamma_{ij}^2 x_{2k} + \frac{\partial l_{ij}}{\partial u_k} n + l_{ij} n_k.$$

Substituting in Equations 6.5.1 and 6.2.2 we obtain

$$\begin{aligned} x_{ijk} &= \frac{\partial \Gamma_{ij}^1}{\partial u_k} x_1 + \Gamma_{ij}^1\left(\Gamma_{1k}^1 x_1 + \Gamma_{1k}^2 x_2 + l_{1k} n\right) \\ &\quad + \frac{\partial \Gamma_{ij}^2}{\partial u_k} x_2 + \Gamma_{ij}^2\left(\Gamma_{2k}^1 x_1 + \Gamma_{2k}^2 x_2 + l_{2k} n\right) \\ &\quad + \frac{\partial l_{ij}}{\partial u_k} n + l_{ij}\left(- L_{1k} x_1 - L_{2k} x_2\right) \\ &= \left\{\frac{\partial \Gamma_{ij}^1}{\partial u_k} + \Gamma_{ij}^1 \Gamma_{1k}^1 + \Gamma_{ij}^2 \Gamma_{2k}^1 - l_{ij} L_{1k}\right\} x_1 \\ &\quad + \left\{\frac{\partial \Gamma_{ij}^2}{\partial u_k} + \Gamma_{ij}^1 \Gamma_{1k}^2 + \Gamma_{ij}^2 \Gamma_{2k}^2 - l_{ij} L_{2k}\right\} x_2 \\ &\quad + \left\{\Gamma_{ij}^1 l_{1k} + \Gamma_{ij}^2 l_{2k} + \frac{\partial l_{ij}}{\partial u_k}\right\} n. \end{aligned} \quad (6.5.9)$$

By interchanging the roles of j and k in the above computation we also obtain

$$
\begin{aligned}
x_{ikj} = &\left\{\frac{\partial \Gamma^1_{ik}}{\partial u_j} + \Gamma^1_{ik}\Gamma^1_{1j} + \Gamma^2_{ik}\Gamma^1_{2j} - l_{ik}L_{1j}\right\}x_1 \\
&+ \left\{\frac{\partial \Gamma^2_{ik}}{\partial u_j} + \Gamma^1_{ik}\Gamma^2_{1j} + \Gamma^2_{ik}\Gamma^2_{2j} - l_{ik}L_{2j}\right\}x_2 \\
&+ \left\{\Gamma^1_{ik}l_{1j} + \Gamma^2_{ik}l_{2j} + \frac{\partial l_{ik}}{\partial u_j}\right\}n.
\end{aligned} \tag{6.5.10}
$$

The equality of mixed partial derivatives implies that $x_{ijk} = x_{ikj}$, so we can equate the final expressions in Equations 6.5.9 and 6.5.10; since x_1, x_2 and n form a basis for $\mathbb{R}^3$ we can equate the coefficients of these three vectors in the two expressions that have been set equal, yielding

$$
\frac{\partial \Gamma^1_{ij}}{\partial u_k} + \Gamma^1_{ij}\Gamma^1_{1k} + \Gamma^2_{ij}\Gamma^1_{2k} - l_{ij}L_{1k} = \frac{\partial \Gamma^1_{ik}}{\partial u_j} + \Gamma^1_{ik}\Gamma^1_{1j} + \Gamma^2_{ik}\Gamma^1_{2j} - l_{ik}L_{1j}, \tag{6.5.11}
$$

$$
\frac{\partial \Gamma^2_{ij}}{\partial u_k} + \Gamma^1_{ij}\Gamma^2_{1k} + \Gamma^2_{ij}\Gamma^2_{2k} - l_{ij}L_{2k} = \frac{\partial \Gamma^2_{ik}}{\partial u_j} + \Gamma^1_{ik}\Gamma^2_{1j} + \Gamma^2_{ik}\Gamma^2_{2j} - l_{ik}L_{2j}, \tag{6.5.12}
$$

$$
\Gamma^1_{ij}l_{1k} + \Gamma^2_{ij}l_{2k} + \frac{\partial l_{ij}}{\partial u_k} = \Gamma^1_{ik}l_{1j} + \Gamma^2_{ik}l_{2j} + \frac{\partial l_{ik}}{\partial u_j}. \tag{6.5.13}
$$

Rearranging the terms of Equation 6.5.13 yields

$$
\frac{\partial l_{ij}}{\partial u_k} - \frac{\partial l_{ik}}{\partial u_j} + \sum_{r=1}^{2}\left(\Gamma^r_{ij}l_{rk} - \Gamma^r_{ik}l_{rj}\right) = 0. \tag{6.5.14}
$$

Substituting the values $i = k = 1$ and $j = 2$ into Equation 6.5.14 yields Equation 6.5.7, and substituting in $i = j = 2$ and $k = 1$ yields Equation 6.5.8.

By substituting $i = k = 2$ and $j = 1$ into Equations 6.5.11 and 6.5.12 and doing some rearranging we obtain

$$
\begin{aligned}
l_{22}L_{11} - l_{12}L_{21} &= \frac{\partial \Gamma^1_{22}}{\partial u_1} - \frac{\partial \Gamma^1_{21}}{\partial u_2} + \sum_{m=1}^{2}\left(\Gamma^m_{22}\Gamma^1_{m1} - \Gamma^m_{21}\Gamma^1_{m2}\right). \\
l_{22}L_{21} - l_{21}L_{22} &= \frac{\partial \Gamma^2_{22}}{\partial u_1} - \frac{\partial \Gamma^2_{21}}{\partial u_2} + \sum_{m=1}^{2}\left(\Gamma^m_{22}\Gamma^2_{m1} - \Gamma^m_{21}\Gamma^2_{m2}\right).
\end{aligned} \tag{6.5.15}
$$

If we let

$$T_r = \frac{\partial \Gamma_{22}^r}{\partial u_1} - \frac{\partial \Gamma_{21}^r}{\partial u_2} + \sum_{m=1}^{2} (\Gamma_{22}^m \Gamma_{m1}^r - \Gamma_{21}^m \Gamma_{m2}^r)$$

for $r = 1, 2$, then we can rewrite Equations 6.5.15 in matrix form as

$$\begin{pmatrix} L_{11} & L_{12} \\ L_{21} & L_{22} \end{pmatrix} \begin{pmatrix} l_{22} \\ -l_{21} \end{pmatrix} = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}.$$

Using Equation 6.2.4 we obtain

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} l_{22} \\ -l_{21} \end{pmatrix} = \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}.$$

Hence

$$\begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix} \begin{pmatrix} l_{22} \\ -l_{21} \end{pmatrix} = \begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} \begin{pmatrix} T_1 \\ T_2 \end{pmatrix}.$$

Multiplying both sides of this equation out and setting the top entries equal yields Equation 6.5.6. □

Our next result, Gauss' famous Theorem Egregium, is one of the more amazing and important theorems of differential geometry. It is this theorem that ultimately makes Gaussian curvature so special. The notion of a quantity associated to surfaces being intrinsic was discussed in Section 5.9; the Theorema Egregium states that Gaussian curvature is intrinsic. Though not difficult to state formally, this theorem is nonetheless somewhat mysterious; it is not clear how Gauss thought of it, and the proof is not particularly enlightening.

Theorem 6.5.3 (Theorema Egregium). *Let $M \subset \mathbb{R}^3$ and $N \subset \mathbb{R}^3$ be smooth surfaces and let $f: M \to N$ be a local isometry. Then $K(f(p)) = K(p)$ for each point $p \in M$, where the curvature at $f(p)$ is computed for the surface N, and the curvature at p is computed for the surface M.*

Proof. It follows from Proposition 5.9.4 (3) that to prove this theorem it would suffice to show that, given a surface and a coordinate patch for the surface, Gaussian curvature can be expressed entirely in terms of the functions E, F and G and their derivatives. More specifically, let $M \subset \mathbb{R}^3$ be a smooth surface, let $x: U \to M$ be a coordinate patch and let $p \in x(U)$ be a point. Throughout this proof all functions defined on U are to be evaluated at $x^{-1}(p)$.

From Proposition 6.4.1 we know that

$$K(p) = \frac{AC - B^2}{EG - F^2}.$$

It will thus suffice to show that the quantity $AC - B^2$ can be expressed entirely in terms of the functions E, F and G and their derivatives (although the individual functions A, B and C cannot be expressed entirely in this way). Note that $AC - B^2$ is just another notation for $l_{11}l_{22} - (l_{12})^2$, and Theorem 6.5.2 says that this quantity only depends upon the g_{ij} and the Γ_{ij}^k. By Lemma 5.7.2 we know that the Γ_{ij}^k can be expressed in terms of the g_{ij} and their derivatives, and thus $AC - B^2$ can be expressed in terms of the g_{ij} and their derivatives alone. □

To get an intuitive feel for the Theorema Egregium, think of local isometries as transformations of surfaces that do not stretch or shrink the domain. As discussed in Example 5.9.5, a classic example of such a map is obtained by taking a piece of paper and rolling it up into a right circular cylinder. The Theorema Egregium then says that the Gaussian curvature at all points of a cylinder must be zero, as indeed it is; in Section 6.3 it was shown that the Gaussian curvature at all points of a generalized cylinder is zero, and the standard right circular cylinder is a special case. (This example also shows that the analog of the Theorema Egregium does not hold for mean curvature, since the mean curvature of the plane is constantly zero, and the mean curvature of the unit right circular cylinder is constantly 1/2.) On the other hand, we now see why a ball cannot be wrapped with wrapping paper without crumpling or tearing the paper. The curvature of a plane (the paper) is zero at every point, whereas the curvature at every point of a sphere of radius R is $\frac{1}{R^2}$. Thus by the Theorema Egregium there can be no local isometry from the plane to the sphere. Any map from the plane to the sphere must stretch or shrink the plane, and wrapping paper (unless it is made of rubber) cannot be stretched.

We end this section by pointing out that the converse to the Theorema Egregium does not hold. More precisely, there are smooth surfaces $M \subset \mathbb{R}^3$ and $N \subset \mathbb{R}^3$ for which there is a smooth function $f: M \to N$ such that $K(f(p)) = K(p)$ for all $p \in M$, and yet the function is not a local isometry. A standard example (following [SK3 vol. III, p. 242]) is to let M be the surface in Exercise 5.5.4 and N be the right helicoid (as parametrized in Section 5.3) with $b = 1$. Both these surfaces are covered by one coordinate patch each; let $y: (0, \infty) \times (-\pi, \pi) \to \mathbb{R}^3$ denote the coordinate patch for M and let $x: \mathbb{R}^2 \to$

$\mathbb{R}^3$ denote the coordinate patch for N. Define a map $f: M \to N$ by setting $f(y(\binom{t}{\theta})) = x(\binom{\theta}{t})$. Using a remark in Section 5.2 it is not hard to see that this map is smooth.

Using Example 6.4.2 and Exercise 6.4.4 we see that for all $\binom{t}{\theta}$ in the domain of y we have

$$K(f(y(\binom{t}{\theta}))) = K(x(\binom{\theta}{t})) = \frac{-1}{(1+t^2)^2} = K(y(\binom{t}{\theta})).$$

Thus the map f satisfies $K(f(p)) = K(p)$ for all $p \in M$. The f is not a local isometry, however. If it were a local isometry then by Proposition 5.9.4 the coordinate patches x and y would have the same metric coefficients (with the roles of t and θ reversed), and this is seen to be not true using Exercises 5.5.8 and 5.5.4.

The Fundamental Theorem of Surfaces is the analog for smooth surfaces of the Fundamental Theorem of Curves (Theorem 4.6.1), which said that curvature and torsion determine a smooth curve up to rotation and translation of $\mathbb{R}^3$. We have no notion of torsion for surfaces, but perhaps the curvature of a surface determines the surface up to rotation and translation. Both a plane and a right circular cylinder have constant curvature $K = 0$, and yet one surface cannot be obtained from the other by a rotation and translation of $\mathbb{R}^3$, so Gaussian curvature does not determine the surface. (We can map the plane onto the cylinder by a local isometry, but such a map is not an isometry of all of $\mathbb{R}^3$ taking one surface to the other.) Perhaps K and H together determine the surface? Unfortunately, the right helicoid (Section 5.3) and the catenoid (Exercise 5.5.3) have the same formulas for K and H (see Example 6.4.2 and Exercise 6.4.1), but one surface cannot be obtained from the other by a rotation and translation of $\mathbb{R}^3$.

The problem is that we are viewing the question incorrectly; we need to look not at the statement of the Fundamental Theorem of Curves but at its proof. The reason that curvature and torsion determined the curve up to rotation and translation is that they were the coefficients in the Frenet–Serret Theorem, which we integrated to find the curve. The combination of the Gauss Formulas (Equation 6.5.1) and Weingarten Equations (Equation 6.2.2) are the analogs of the Frenet–Serret Theorem, so we need to look at the coefficients in these equations. At first glance there seem to be too many different quantities involved in these coefficients. However, the Γ_{ij}^k can be expressed in terms of the g_{ij} and their derivatives using Lemma 5.7.2, and the L_{ij} can be expressed in terms of

the g_{ij} and the l_{ij} using Equation 6.2.4. Thus the eight functions g_{ij} and l_{ij} determine the coefficients of x_1, x_2 and n in Equations 6.2.2 and 6.5.1, so they are good candidates to determine surfaces up to rotation and translation.

Just as there was the restriction that $\kappa > 0$ in the Fundamental Theorem of Curves, we can make some obvious restrictions on the g_{ij} and l_{ij}. First, we know that $g_{ij} = g_{ji}$ and $l_{ij} = l_{ji}$. Also, we know that $g_{11} = \langle x_1, x_1 \rangle = \|x_1\|^2 > 0$, and similarly $g_{22} > 0$. Moreover, we know that

$$g_{11}g_{22} - (g_{12})^2 = \|x_1 \times x_2\|^2 > 0$$

using Lemma 5.5.2 (i) and the definition of coordinate patches. This last inequality combined with $g_{11} > 0$ implies automatically that $g_{22} > 0$. There are no such simple positivity restrictions on the l_{ij}. We now ask: Given eight smooth functions g_{ij} and l_{ij} defined on an open subset $U \subset \mathbb{R}^2$, satisfying the symmetry and positivity restrictions just mentioned, is there a coordinate patch $x: U \to \mathbb{R}^3$ with these functions forming the entries of the matrices for the first and second fundamental forms of the coordinate patch?

The answer, unfortunately, is still no, as seen in the following example. Let functions $g_{ij}, l_{ij}: \mathbb{R}^2 \to \mathbb{R}^3$ for $i, j = 1, 2$ be defined to be the constant functions

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}.$$

Certainly these functions satisfy all the symmetry and positivity conditions mentioned. We attempt to solve the Gauss formulas and Weingarten equations for x_1, x_2 and n given these g_{ij} and l_{ij}; if there were a coordinate patch $x: U \to M$ with the desired first and second fundamental form, then these equations would certainly have a solution. We start by computing the other coefficients for x_1, x_2 and n. First, observe that $(g_{ij})^{-1}$ is also constantly the identity matrix. Next, using Lemma 5.7.2 as the definition of the Γ_{ij}^k in terms of the g_{ij}, we see that all the $\Gamma_{ij}^k = 0$, since the g_{ij} are all constant functions. It follows from Equation 6.2.4 that (L_{ij}) is constantly the identity matrix. Substituting these values into Equations 6.2.2 and 6.5.1 yields

$$x_{11} = n, \quad x_{12} = x_{21} = 0, \quad x_{22} = n, \tag{6.5.16}$$

$$n_1 = -x_1, \quad n_2 = -x_2. \tag{6.5.17}$$

Suppose there were a solution to this system of linear partial differential equations. By taking the various partial derivatives of n_1 and n_2 as given in Equation 6.5.17, and substituting the values from Equation 6.5.16, we obtain

$$n_{11} = -x_{11} = -n, \tag{6.5.18}$$

$$n_{12} = n_{21} = -x_{21} = 0, \tag{6.5.19}$$

$$n_{22} = -x_{22} = -n. \tag{6.5.20}$$

Since the only partial derivatives in Equation 6.5.18 are with respect to u_1 we can solve the equation as if it were an ordinary differential equation, but having our constants be functions of u_2. Thus viewed, Equation 6.5.18 is a standard differential equation, and its solution is

$$n(\begin{pmatrix} u_1 \\ u_2 \end{pmatrix}) = f_1(u_2)\sin u_1 + f_2(u_2)\cos u_1, \tag{6.5.21}$$

for some vector-valued functions f_1 and f_2 of u_2 alone. Substituting this formula for n into Equation 6.5.20 yields

$$f_1''(u_2)\sin u_1 + f_2''(u_2)\cos u_1 = -f_1(u_2)\sin u_1 - f_2(u_2)\cos u_1. \tag{6.5.22}$$

Equating coefficients we see that $f_i''(u_2) = -f_i(u_2)$ for $i = 1, 2$. As before, we deduce that

$$f_i(u_2) = A_i \sin u_2 + B_i \cos u_2 \tag{6.5.23}$$

for $i, = 1, 2$, where the A_i and B_i are constant vectors. By substituting Equation 6.5.23 into Equation 6.5.21, and then plugging the resulting expression for n into Equation 6.5.19, the reader can verify that $A_1 = B_1 = A_2 = B_2 = 0$. Thus n is the zero function. It follows from Equation 6.5.17 that x_1 and x_2 are constantly zero, which is certainly never the case in a coordinate patch. Hence there can be no coordinate patch with the desired first and second fundamental form.

What went wrong in the above attempt was nothing more than our overly high expectations. Unlike ordinary differential equations, for which there are nice existence theorems, there are no such simple existence theorems for partial differential equations, so we should not have expected a solution for any choice of g_{ij} and l_{ij} subject only to our simple symmetry and positivity conditions.

It turns out that in order to guarantee solutions to partial differential equations certain integrability conditions must be satisfied; we will not state these conditions in general (see [SK3 vol. I, Chapter 6] or [FL, pp. 92–101]), though in essence the integrability conditions express the equality of mixed partial derivatives. In our particular situation these integrability conditions should be some equations that the g_{ij} and l_{ij} must satisfy for any coordinate patch. Remarkably enough, the necessary equations are precisely the Gauss Equation and the Codazzi–Mainardi Equations (Theorem 6.5.2).

Theorem 6.5.4 (Fundamental Theorem of Surfaces). *Let $U \subset \mathbb{R}^2$ be a connected open set. Suppose that there are eight smooth functions $g_{ij}, l_{ij}: U \to \mathbb{R}$ for $i, j = 1, 2$ satisfying the following two conditions.*

(1) $$g_{ij} = g_{ji}, \quad \text{and} \quad l_{ij} = l_{ji}$$

for all $i, j = 1, 2$, and

$$g_{11} > 0, \quad \text{and} \quad g_{11}g_{22} - (g_{12})^2 > 0.$$

(2) If eight functions $\Gamma_{ij}^k: U \to \mathbb{R}$, for all $i, j, k = 1, 2$, are defined by

$$\begin{pmatrix} \Gamma_{ij}^1 \\ \Gamma_{ij}^2 \end{pmatrix} = \frac{1}{2}\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix}^{-1} \begin{pmatrix} \frac{\partial g_{j1}}{\partial u_i} + \frac{\partial g_{i1}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_1} \\ \frac{\partial g_{j2}}{\partial u_i} + \frac{\partial g_{i2}}{\partial u_j} - \frac{\partial g_{ij}}{\partial u_2} \end{pmatrix},$$

then

$$l_{11}l_{22} - (l_{12})^2 = \sum_{r=1}^{2} g_{1r}\left\{\frac{\partial \Gamma_{22}^r}{\partial u_1} - \frac{\partial \Gamma_{21}^r}{\partial u_2} + \sum_{m=1}^{2}\left(\Gamma_{22}^m\Gamma_{m1}^r - \Gamma_{21}^m\Gamma_{m2}^r\right)\right\},$$

$$\frac{\partial l_{12}}{\partial u_1} - \frac{\partial l_{11}}{\partial u_2} + \sum_{r=1}^{2}\left(\Gamma_{12}^r l_{r1} - \Gamma_{11}^r l_{r2}\right) = 0,$$

$$\frac{\partial l_{22}}{\partial u_1} - \frac{\partial l_{21}}{\partial u_2} + \sum_{r=1}^{2}\left(\Gamma_{22}^r l_{r1} - \Gamma_{21}^r l_{r2}\right) = 0.$$

Then for each point $p \in U$ there is an open subset $V \subset U$ containing p and a coordinate patch $x: V \to \mathbb{R}^3$ for which (g_{ij}) and (l_{ij}) are the matrices for the first and second fundamental forms respectively. Any one such coordinate patch can be obtained from any other by a rotation and translation of $\mathbb{R}^3$.

Proofs of this theorem may be found in [DO1, p. 311] or [SK3 vol. III, p. 79]. The connectivity of the set U in the theorem is needed to insure the uniqueness of the image of x up to rotation and translation, since if U were not connected we could rotate and translate the image of only one component.

Exercises

6.5.1*. Let $M \subset \mathbb{R}^3$ be a surface and let $x: U \to M$ be a coordinate patch. Suppose that x has metric coefficients $E = 1$ and $F = 0$ on all of U. This exercise has two steps.

(1) Express the quantity $AC - B^2$ in terms of G.

(2) Show that

$$K \circ x = -\frac{1}{\sqrt{G}} \frac{\partial^2 \sqrt{G}}{\partial s^2}.$$

6.5.2. Find a formula for the Gaussian curvature at all points of a rectifying developable surface (as parametrized in Section 5.3). Deduce that $K \leq 0$ at all points on the surface, and that $K(x(\binom{s}{t})) = 0$ iff $\tau(s) = 0$.

6.5.3. Let functions $g_{ij}, l_{ij}: \mathbb{R}^2 \to \mathbb{R}^3$ for $i, j = 1, 2$ be defined by

$$\begin{pmatrix} g_{11} & g_{12} \\ g_{21} & g_{22} \end{pmatrix} = \begin{pmatrix} (u_1)^2 + 1 & 0 \\ 0 & 1 \end{pmatrix} \quad \text{and} \quad \begin{pmatrix} l_{11} & l_{12} \\ l_{21} & l_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 0 & u_1 \end{pmatrix}.$$

Is there a coordinate patch $x: \mathbb{R}^2 \to \mathbb{R}^3$ with these g_{ij} and l_{ij}?

Endnotes

Notes for Section 6.1

Gauss laid out the foundations of modern differential geometry in his landmark work [GA]. See [SK3 vol. II, Chapter 3] for an exposition of Gauss' original approach restated in more modern terminology. The importance of Gauss' work cannot be overestimated.

Notes for Section 6.2

(A) Defining curvature via the Weingarten map is not the only possible approach for surfaces, though all approaches yield Gaussian curvature. See [SK3 vol. II] for a variety of approaches in the more general setting of manifolds, and [OS2] for a survey of some recent work on curvature.

(B) In some books the Weingarten map is referred to as the *shape operator*.

(C) Unlike the notation (g_{ij}), which is fairly universal, there is no universal notation for the matrices of L and II.

(D) Although we will be sticking to the matrix equation given in Equation 6.2.1, the reader might find in other differential geometry books that the same equality is expressed by equations of the form

$$L_{ij} = \sum_{r=1}^{2} g^{ir} l_{rj}, \tag{6.E.1}$$

where the g^{kl} denote the entries of the matrix $(g_{ij})^{-1}$; to obtain Equation 6.E.1 simply multiply out the right hand side of Equation 6.2.1 using the g^{kl} notation and take a typical element. Equations of the sort seen in Equation 6.E.1, involving a careful use of subscripts and superscripts, predominated in older treatments of differential geometry, and are still often used by physicists (for example, in general relativity). Though the eyesore of indices gets worse in higher dimensions, if we want to do computation using coordinate patches in higher dimensions the use of such formulae is probably unavoidable; in the case of surfaces we can fortunately reformulate everything in terms of matrices (in higher dimensions we would need something like $n \times n \times \cdots \times n$ "matrices"). See [SK3 vol. II] for a very thorough comparative treatment with and without indices.

Notes for Section 6.3

(A) Mean curvature is particularly useful for the study of minimal surfaces; see [OS1], [DO1, §3-5] or [SK3 vol. III, Chapter 3].

(B) The reader might wonder why something called Euler's formula appears after we have defined something called Gaussian curvature, since Euler lived before Gauss. If we combine Proposition 6.3.2 with Theorem 6.3.3 we see that

Euler's formula can be stated without any reference to the second fundamental form, and that it is really a theorem about the curvature of certain curves in surfaces; the curvature of curves was known by Euler's time.

Notes for Section 6.4

The notation E, F and G was used by Gauss, and is quite standard. The use of A, B and C is not standard, but there appears to be no standard usage in this case.

Notes for Section 6.5

Another proof of the Theorema Egregium can be found in [SK3 vol. II, chapter 3], who claims that this proof is not all that different from Gauss' proof. This other proof is simply a long calculation, but it has the advantage of avoiding the Christoffel symbols entirely.

CHAPTER VII

Geodesics

7.1 Introduction

Geodesics are curves in surfaces that play a role analogous to that of straight lines in the plane. Straight lines in the plane have three important properties:

(1) There is a unique straight line containing any two distinct points;

(2) the straight line between any two points is the shortest path between the points;

(3) a straight line "does not bend to the left or right as we travel along it." (This could be restated more precisely by saying that some appropriate derivative is constant.)

Although all three of these properties hold simultaneously for straight lines in the plane, things are not so simple on more general surfaces. Consider driving on a road that is part of a great circle on Earth, that is, a circle of largest possible diameter, such as the equator or a longitude. It would feel as if you were going in a straight line, since relative to Earth you would not be veering to the right or the left. Thus, whatever a rigorous definition of our notion of "straight lines" on surfaces might be, great circles intuitively satisfy property (3). On the other hand, they do not satisfy either property (1) or (2). For property (1), observe that there are infinitely many longitudes between the North and South poles on the Earth. For property (2), take two nearby points on a sphere, and join them by the piece of the great circle that goes around "the long way." Such a great circle still doesn't bend, but it is certainly not the shortest path between the two points.

We see that not all three of the properties (1)–(3) will always hold simultaneously, and we thus have to choose the most useful of the three properties to use as the basis for defining geodesics. Though it may appear at first to be the least appealing, we choose property (3). This property is a local property (that is, it depends only on what is happening near a given point), making it more suitable to differential methods. Properties (1) and (2) are, by contrast, global properties, since they need to take into account what happens at possibly far

away points. We now turn to a more precise look at property (3), though we will return to properties (1) and (2) briefly later on.

7.2 Geodesics

What criteria can be used to generalize property (3) to surfaces? We note that nothing as simple as requiring curves to have zero curvature as curves in $\mathbb{R}^3$ can be used, since a great circle on a sphere does have non-zero curvature when viewed as a curve in $\mathbb{R}^3$, and yet it is what we want to call a geodesic. Rather, we use the idea of vectors remaining parallel as they are moved around on a surface. The concept of "parallel" is really in the eye of the beholder. Suppose that you are driving along a straight road (which is a great circle on the surface of Earth), and suppose you are holding a stick out straight in front of your car. Since you are driving on what appears to you to be a straight road, and since the stick is always pointing straight in front of your car, you would certainly say that the stick is being held parallel to itself at all moments. However, if someone were looking down at you from outer space, the stick would not appear to be kept parallel to itself, since as you drive around Earth the direction of the stick in space changes. We need to be able to see things from the point of view of a person on the surface.

Let us rephrase this idea in terms of vector fields. When does it happen that from the point of view of a creature on the surface, a vector field along a curve on the surface appears as if it remains parallel to itself along the curve? Given that we consider all vectors as being translated so as to start at the origin, what we want is a vector field that appears constant to a creature on the surface. Consider the fact that when we stand at a particular spot on Earth, the plane we think of as Earth is really just the tangent plane to Earth at the point in question. To calculate the rate of change of a vector field along a curve in a surface from the point of view of a creature on the surface, we thus only want to consider the component in the tangent plane of the rate of change of the vector field. This is precisely what $\frac{DZ}{dt}$ was defined for in Section 5.6. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $c\colon (a, b) \to M$ be a smooth curve and let $Z\colon (a, b) \to \mathbb{R}^3$ be a smooth vector field along c that is tangent to M along c. The vector field Z appears to be constant from the point of view of a creature on the surface if $\frac{DZ}{dt} = 0$ at all points $t \in (a, b)$.

Rather than considering all vector fields along a curve c in a surface, we are really interested in the one vector field along c that best tells us about the way in which c is bending, namely c' (which is necessarily a tangent vector field). A curve c will appear not to be bending to the right or left to someone travelling along the curve precisely if the vector field c' appears to be parallel to itself along c; hence the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $c\colon (a, b) \to M$ be a smooth curve. The curve c is a **geodesic** if

$$\frac{Dc'(t)}{dt} = 0$$

for all $t \in (a, b)$. $\Diamond$

Example 7.2.1. (1) We want find all geodesics in the plane $\mathbb{R}^2$. As mentioned in Example 5.4.1 (1), the tangent plane at each point of $\mathbb{R}^2$ is simply $\mathbb{R}^2$ itself. Thus a tangent vector field along a curve in $\mathbb{R}^2$ is simply a vector field in $\mathbb{R}^2$ along the curve; the derivative $\frac{d}{dt}$ of such a tangent vector field must also be a vector field in $\mathbb{R}^2$, thus for tangent vector fields along a curve in $\mathbb{R}^2$ the covariant derivative $\frac{D}{dt}$ equals the regular derivative $\frac{d}{dt}$. Therefore a geodesic in the plane will be a curve $c\colon (a, b) \to \mathbb{R}^2$ such that

$$\frac{dc'(t)}{dt} = c''(t) = 0$$

for all t. It follows that c has the form

$$c(t) = \begin{pmatrix} mt + p \\ nt + q \end{pmatrix}, \tag{7.2.1}$$

and thus c is a straight line. Conversely, any straight line in the plane can be parametrized as in Equation 7.2.1, and is therefore a geodesic.

(2) We show that every great circle on the unit sphere S^2, when parametrized correctly, is a geodesic. By the symmetry of the sphere it will suffice to show that the equator of S^2, which we parametrize as $c\colon (-\pi, \pi) \to S^2$ given by

$$c(t) = \begin{pmatrix} \cos t \\ \sin t \\ 0 \end{pmatrix},$$

is a geodesic. It is straightforward to compute that

$$\frac{dc'(t)}{dt} = \begin{pmatrix} -\cos t \\ -\sin t \\ 0 \end{pmatrix} = -c(t).$$

Recall that S^2 has the property that the normal vector at each point $p \in S^2$ is simply p (or $-p$). It follows that the projection of $\frac{dc'(t)}{dt}$ onto the tangent plane $T_{c(t)}S^2$ is always zero, and hence $\frac{Dc'(t)}{dt} = 0$. $\Diamond$

Contrary to our intuition, in which the image of a curve should either be a geodesic or not, according to our definition a geodesic is a particular parametrization of a curve. For example, a careful examination of Example 7.2.1 (1) reveals that it was only certain parametrizations of straight lines that were shown to be geodesics, namely the linear parametrizations. By contrast, consider the curve $c: (-\infty, \infty) \to \mathbb{R}^2$ given by $c(t) = \begin{pmatrix} t^3 \\ t^3 \\ 0 \end{pmatrix}$. The image of this curve is clearly a straight line. However, computing as we did in Example 7.2.1 (1), we see that

$$\frac{Dc'(t)}{dt} = \frac{dc'(t)}{dt} = \begin{pmatrix} 3t^2 \\ 3t^2 \\ 0 \end{pmatrix},$$

which is not zero when $t \neq 0$. Thus the choice of parametrization really does matter in determining whether a given curve is a geodesic or not. The following lemma, which will be used in Chapter VIII, shows the extent to which reparametrizations of geodesics are still geodesics.

Lemma 7.2.2. *Let $M \subset \mathbb{R}^3$ be a surface and let $c: (a, b) \to M$ be a non-constant geodesic. Then the following hold.*

(i) *The curve c has constant non-zero speed; that is, $\|c'(t)\|$ is a non-zero constant for all $t \in (a, b)$.*

(ii) *Let $g: (d, e) \to (a, b)$ be a smooth map for some open interval (d, e). Then $c \circ g$ is a geodesic iff g has the form $g(s) = ms + n$ for some real numbers m and n.*

(iii) *Suppose that c is a homeomorphism from (a, b) onto $c((a, b))$. If $\widehat{c}: (d, e) \to M$ is a curve with $\widehat{c}((d, e)) \subset c((a, b))$, then $\widehat{c}$ is a geodesic iff it has constant speed.*

Proof. (i). Using Lemma 5.6.8 (iii) and the definition of a geodesic we compute

$$\frac{d}{dt}\|c'(t)\|^2 = \frac{d}{dt}\langle c'(t), c'(t)\rangle = 2\langle \frac{Dc'(t)}{dt}, c'(t)\rangle = 0.$$

Thus $\|c'(t)\|$ is a constant. Since we are assuming that c is not constant, we see that $\|c'(t)\| \neq 0$.

(ii). Let $\widetilde{c} = c \circ g$. Suppose first that $g(s) = ms + n$ for some real numbers m and n. If we let $t = g(s)$, we now have $\widetilde{c}'(s) = m\, c'(t)$. Hence

$$\frac{D\widetilde{c}'(s)}{ds} = \frac{D\widetilde{c}'(s)}{dt}\frac{dt}{ds} = \frac{Dmc'(t)}{dt}m = m^2\frac{Dc'(t)}{dt} = 0,$$

where the last equality holds because c is a geodesic. Hence $\widetilde{c}$ is a geodesic as well.

Now suppose that $\widetilde{c}$ is a geodesic. By part (i) both c and $\widetilde{c}$ have constant speed; say $\|c'(t)\| = p$ for all $t \in (a, b)$ and $\|\widetilde{c}'(s)\| = q$ for all $s \in (d, e)$, where p and q are constants. By hypothesis on c and part (i) we know $p \neq 0$. By the definition of $\widetilde{c}$ we have

$$q = \|\widetilde{c}'(s)\| = \|c'(t)\|\,|g'(s)| = p|g'(s)|$$

for all $s \in (d, e)$. It now follows from the smoothness of g that $g'(s)$ is either constantly $\frac{q}{p}$ or constantly $-\frac{q}{p}$, and therefore g has the desired form.

(iii). If $\widehat{c}$ does not have constant speed then it clearly cannot be a geodesic by part (i). Now assume that $\widehat{c}$ has constant speed. If $\widehat{c}$ has constant speed zero the result is trivial, so assume otherwise. Let $s_0 \in (d, e)$ be fixed. We need to show that $\frac{D\widehat{c}'(s)}{ds}(s_0) = 0$. We now proceed similarly to the start of the proof of Proposition 5.2.5 (i). By the injectivity of c it follows that there is a unique point $q \in (a, b)$ such that $c(q) = \widehat{c}(s_0)$. By part (i) of this lemma $c'(q) \neq 0$. Using Exercise 4.2.1 there is a number $\delta > 0$, an open subset $V \subset \mathbb{R}^3$ containing $c(q)$ and a smooth map $G\colon V \to \mathbb{R}^3$ such that $G(V)$ is open in $\mathbb{R}^3$, that G is a diffeomorphism from V onto $G(V)$, that $c((q - \delta, q + \delta)) \subset V$ and that

$$G \circ c(t) = \begin{pmatrix} t \\ 0 \\ 0 \end{pmatrix}$$

for all $t \in (q - \delta, q + \delta)$. For convenience, let $J = (q - \delta, q + \delta)$. Let $\pi_1\colon \mathbb{R}^3 \to \mathbb{R}$ be projection onto the first coordinate, that is $\pi_1(\begin{pmatrix} x \\ y \\ z \end{pmatrix}) = x$ for

all $\begin{pmatrix} x \\ y \\ z \end{pmatrix} \in \mathbb{R}^3$. The map π_1 is continuous, as seen in Example 1.3.1. Observe that $\pi_1 \circ G|c(J) \circ c|J = 1_J$, and since $c|J$ is bijective it is straightforward to verify that $(c|J)^{-1} = \pi_1 \circ G|c(J)$.

Since c is a homeomorphism onto its image, it follows that $c(J)$ is open in $c((a, b))$. We can view $\widehat{c}$ as a map $(d, e) \to c((a, b))$, and hence by the continuity of $\widehat{c}$ there is some number $\epsilon > 0$ such that

$$(s_0 - \epsilon, s_0 + \epsilon) \subset (\widehat{c})^{-1}(c(J)).$$

For convenience let c_* denote the restriction of $\widehat{c}$ to $(s_0 - \epsilon, s_0 + \epsilon)$. We now define a map h: $(s_0 - \epsilon, s_0 + \epsilon) \to J$ by letting $h = \pi_1 \circ G \circ c_*$. The map h is smooth since it is the composition of smooth maps. We now compute

$$\begin{aligned} c \circ h &= c|J \circ h = c|J \circ (\pi_1 \circ G \circ c_*) \\ &= c|J \circ 1_J \circ (\pi_1 \circ G|c(J)) \circ c_* \\ &= c|J \circ (\pi_1 \circ G|c(J) \circ c|J)^{-1} \circ (\pi_1 \circ G|c(J)) \circ c_* \\ &= c|j \circ (c|J)^{-1} \circ (\pi_1 \circ G|c(J))^{-1} \circ (\pi_1 \circ G|c(J)) \circ c_* = c_*. \end{aligned}$$

It now follows that $c_*'(s) = c'(h(s))\, h'(s)$ for all $s \in (s_0 - \epsilon, s_0 + \epsilon)$, and thus $\|c_*'(s)\| = \|c'(h(s))\|\, |h'(s)|$. Since both c and c_* have constant speed, and since c has non-zero speed, it follows that $|h'(s)|$ is a constant. Since h is a smooth function, it follows that h' must be continuous; hence $h'(s)$ is a constant, say $h'(s) = m$. Thus $h(s) = ms + n$ for some real number n. It follows from part (ii) that c_* is a geodesic. Hence

$$\frac{D\widehat{c}'(s)}{ds}(s_0) = \frac{Dc_*'(s)}{ds}(s_0) = 0. \quad \square$$

For later use we need the following definition and lemma.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $A \subset M$ be an arc (as defined in Section 2.2). The arc A is a **regular arc** (respectively a **geodesic arc**) if there is a regular curve (respectively a geodesic) c: $(a, b) \to M$ such that $A = c([x, y])$ for some closed interval $[x, y] \subset (a, b)$. $\Diamond$

Lemma 7.2.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $A \subset M$ be a regular arc. If c: $(a, b) \to M$ is a regular curve such that $A = c([x, y])$ for some*

closed interval $[x, y] \subset (a, b)$, *then* $c|[x, y]$ *is injective, and* c *maps* x *and* y *to the endpoints of* A.

Proof. Exercise 7.2.6.

We have gone about as far as we can without using coordinate patches. The following lemma shows that the criterion for a curve being a geodesic becomes a system of differential equations when expressed in terms of a coordinate patch. We will then be able to use standard results about the existence and uniqueness of solutions of differential equations (summarized in Section 4.2) to determine the existence and uniqueness of geodesics in certain situations. Let $\bar{c}$, c_1 and c_2 be as in Section 5.2.

Proposition 7.2.4. *Let* $M \subset \mathbb{R}^3$ *be a smooth surface, let* $x: U \to M$ *be a coordinate patch and let* $c: (a, b) \to x(U)$ *be a smooth curve. Then* c *is a geodesic iff*

$$\begin{aligned} c_1'' + \Gamma_{11}^1 (c_1')^2 + 2\Gamma_{12}^1 c_1' c_2' + \Gamma_{22}^1 (c_2')^2 = 0, \\ c_2'' + \Gamma_{11}^2 (c_1')^2 + 2\Gamma_{12}^2 c_1' c_2' + \Gamma_{22}^2 (c_2')^2 = 0, \end{aligned} \tag{7.2.2}$$

where the c_i *and their derivatives are evaluated at* t, *and the* Γ_{ij}^k *are evaluated at* $\bar{c}(t)$.

Proof. We did all the work for this proof when we discussed covariant derivatives using coordinate patches. Consider $c' = c_1' x_1 + c_2' x_2$ as a vector field along c which is tangent to M along c; applying Proposition 5.7.5 to this vector field we see that c is a geodesic iff

$$0 = \frac{Dc'}{dt} = \sum_{k=1}^{2} \left(\frac{dc_k'}{dt} + \sum_{i=1}^{2} \sum_{j=1}^{2} \Gamma_{ij}^k c_j' c_i' \right) x_k.$$

Since $\{x_1, x_2\}$ is a basis for the tangent plane, this last equation is equivalent to

$$\frac{dc_k'}{dt} + \sum_{i=1}^{2} \sum_{j=1}^{2} \Gamma_{ij}^k c_j' c_i' = 0$$

for $k = 1, 2$. The desired result now follows. □

The following result is a nice application of the above proposition.

Proposition 7.2.5. *Let $M \subset \mathbb{R}^3$ be a surface of revolution with unit speed profile curve $d(t) = \binom{r(t)}{z(t)}$. Then the following hold.*

(i) *Every meridian of M can be parametrized as a geodesic.*

(ii) *A circle of latitude of M can be parametrized as a geodesic iff the tangent vector to the profile curve at its point of intersection with the circle of latitude is parallel to the axis of revolution.*

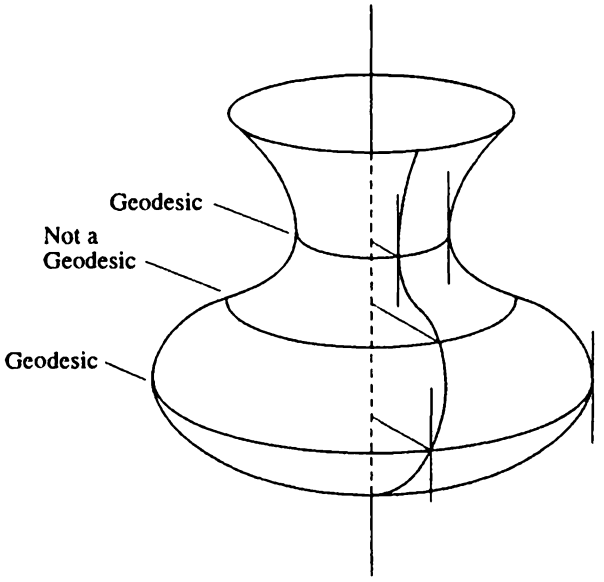


Figure 7.2.1

Proof. Since the profile curve is unit speed we know from Exercise 5.7.3 that $\Gamma^1_{22} = -rr'$, $\Gamma^2_{12} = \Gamma^2_{21} = r'/r$ and all other Γ^k_{ij} are 0. We will write $t(s)$ and $\theta(s)$ instead of c_1 and c_2 respectively; Equation 7.2.2 thus becomes

$$t'' - rr'(\theta')^2 = 0, \tag{7.2.3}$$

$$\theta'' + 2\frac{r'}{r}t'\theta' = 0, \tag{7.2.4}$$

where t, θ and their derivatives are evaluated at s, and r and its derivative are evaluated at $t(s)$.

(i) A meridian on a surface of revolution is a curve obtained by holding θ constant and varying t; such a curve can be parametrized by $c(s) = x(\begin{pmatrix} s \\ \theta_0 \end{pmatrix})$ for some constant θ_0. We thus take $t(s) = s$ and $\theta(s) = \theta_0$. Hence $t'' = 0$ and $\theta'' = \theta' = 0$. This parametrization c of a typical meridian therefore satisfies Equations 7.2.3 and 7.2.4, which proves the desired result.

(ii) A circle of latitude on a surface of revolution is a curve obtained by holding t constant and varying θ; such a curve has the form $c(s) = x(\begin{pmatrix} t_0 \\ \theta(s) \end{pmatrix})$ for some constant t_0. Since $t(s) = t_0$ it follows that $t'' = t' = 0$. Suppose that this circle of latitude c is a geodesic. The velocity vector of the circle of latitude is $c'(s) = \theta'(s)x_2$, where for convenience we drop the arguments in x_2. Hence $\|c'(s)\| = |\theta'(s)|\|x_2\|$. Since $c(s)$ is a geodesic it must have constant speed, so $|\theta'(s)|\|x_2\|$ is a constant. It can be verified that $\|x_2\| = r(t(s)) = r(t_0)$. Thus $|\theta'(s)|$ is a non-zero constant. By smoothness it follows that $\theta'(s)$ is a non-zero constant. The curve c must satisfy Equation 7.2.3, so $rr' = 0$. Since we always assume that $r(t_0) > 0$ for any surface of revolution, it follows that $r'(t_0) = 0$, and this latter condition implies that the tangent vector to the profile curve at its point of intersection with the circle of latitude is parallel to the axis of revolution.

Now suppose that the tangent vector to the profile curve at its point of intersection with the circle of latitude $c(s)$ is parallel to the axis of revolution. It follows that $r'(t_0) = 0$. Since $t'' = 0$, we see that c satisfies Equation 7.2.3. Since we are only claiming that there is some parametrization of the circle of latitude that is a geodesic, let us choose $\theta(s) = s$. It then follows that $\theta'' = 0$, and since $t' = 0$ we deduce that c satisfies Equation 7.2.4. Hence c is a geodesic. □

The reason that we can only prove that some parametrizations of meridians and appropriate circles of latitude are geodesics is that in general not every reparametrization of a geodesic is still a geodesic. Also, note that the above proposition does not claim to characterize all geodesics on surfaces of revolution, only those geodesics that are meridians or circles of latitude. This proposition can be used to show that any smooth, injective, unit speed planar curve, with the property that it does not intersect some straight line in the plane, is actually a geodesic on some surface. Simply rotate the curve about the line that it does not intersect, and the curve is then a meridian on the resulting surface of revolution, which is a geodesic by the above proposition. It turns out that in general any

smooth, injective, unit-speed curve is a geodesic on some surface, though the proof is not as simple; see Exercise 7.2.4 for details.

Can any two points on a smooth surface be joined by a geodesic arc? If so, then geodesics would satisfy the analog of property (1) of Section 7.3. The example of great circles on the sphere, mentioned in the previous section, shows that in general we cannot hope to find a unique geodesic arc joining any two distinct points. Can we always find at least one? Consider the surface $M = \mathbb{R}^2 - \{O_2\}$. The geodesics in any subset of the plane are still straight lines (the same argument used for the whole plane still works here); hence there can be no geodesic arc in M from $\binom{1}{0}$ to $\binom{-1}{0}$, since the line segment from one point to the other would have to pass though the origin, which is not in the surface.

The Hopf–Rinow Theorem (stated in Section 8.5) says that if a surface has no "holes" (as in the previous example) then any two distinct points lie on some geodesic, which furthermore has the shortest length of any curve joining the two points. The question of what constitutes a hole is subtle, however. An infinite circular cylinder certainly has a "hole" in a straightforward geometric sense, but any two points on the cylinder can in fact be joined by a length-minimizing geodesic. The infinite circular cylinder and a plane with a point removed are homeomorphic, so whatever is meant by a "hole" from the point of view of geodesics is not a topological invariant. See see [KL] or [DO1] for more details.

Instead of trying to join two points by a geodesic, which is tricky, we think of starting from a point and heading in a given direction. It seems reasonable, based on our experience on Earth, that, from any starting point, there is one and only one way to move forward in a given direction if we do not want to feel as if we are turning to the right or the left as we go forward. We cannot always expect to continue very far in this manner, since there might be an obstacle in our path. The following theorem gives the analog of this idea for all surfaces. For the concept of "direction" we use tangent vectors. The fact that we may run into an obstacle (such as a hole) is taken into account by the fact that the length of the geodesic depends upon the starting point and direction.

Theorem 7.2.6. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let p be a point in M and let $v \in T_pM$ be a vector. Then there exists a number $\epsilon > 0$ and a geodesic $c\colon (-\epsilon, \epsilon) \to M$ such that $c(0) = p$ and $c'(0) = v$. The geodesic c is unique in the following sense: If for some number $\delta > 0$ the curve $\widetilde{c}\colon (-\delta, \delta) \to M$ is another geodesic such that $\widetilde{c}(0) = p$ and $\widetilde{c}'(0) = v$, then $\widetilde{c}(t) = c(t)$ for all t*

in the intersection of the domains of the two curves.

Proof. Let $x: U \to M$ be a coordinate patch such that $p \in x(U)$. The proof consists of applying the existence and uniqueness theorem for the solutions of ordinary differential equations with initial conditions to Equation 7.2.2. These differential equations are second order equations, but if we introduce the two new variables $d_1 = c_1'$ and $d_2 = c_2'$ we can rewrite the equation as the following system of first order differential equations:

$$\begin{aligned} c_1' &= d_1 \\ c_2' &= d_2 \\ d_1' &= -\Gamma_{11}^1 (d_1)^2 - 2\Gamma_{12}^1 d_1 d_2 - \Gamma_{22}^1 (d_2)^2 \\ d_2' &= -\Gamma_{11}^2 (d_1)^2 - 2\Gamma_{12}^2 d_1 d_2 - \Gamma_{22}^2 (d_2)^2, \end{aligned} \tag{7.2.5}$$

where the Γ_{ij}^k are functions of c_1 and c_2. Rather than thinking of this system as four equations in four unknowns we can think of it as a single vector-valued differential equation, the domain of which is

$$\left\{ \begin{pmatrix} c_1 \\ c_2 \\ d_1 \\ d_2 \end{pmatrix} \in \mathbb{R}^4 \mid \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} \in U \text{ and } \begin{pmatrix} d_1 \\ d_2 \end{pmatrix} \in \mathbb{R}^2 \right\} = U \times \mathbb{R}^2.$$

The initial conditions for our vector-valued differential equation are

$$\begin{pmatrix} c_1(0) \\ c_2(0) \\ d_1(0) \\ d_2(0) \end{pmatrix} = \begin{pmatrix} p_1 \\ p_2 \\ v_1 \\ v_2 \end{pmatrix},$$

where $x^{-1}(p) = \bar{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$ and $v = v_1 x_1(\bar{p}) + v_2 x_2(\bar{p})$. Applying Theorem 4.2.4 to our differential equation and initial condition, we deduce that there is a number $\epsilon > 0$ and a smooth function $C: (-\epsilon, \epsilon) \to U \times \mathbb{R}^2$, written

$$C(t) = \begin{pmatrix} c_1(t) \\ c_2(t) \\ d_1(t) \\ d_2(t) \end{pmatrix},$$

such that the functions $c_1(t)$, $c_2(t)$, $d_1(t)$ and $d_2(t)$ satisfy our system of differential equations and initial condition. Tracing through our construction we see

that the curve $c: (-\epsilon, \epsilon) \to U$ defined by $c(t) = x\left(\begin{pmatrix} c_1(t) \\ c_2(t) \end{pmatrix}\right)$ will be a geodesic with $c(0) = p$ and $c'(0) = v$. The uniqueness condition in Theorem 4.2.4 guarantees the uniqueness of c. □

Example 7.2.7. We will show that the only geodesics on the surface S^2 are great circles (or pieces of them). We saw in Example 7.2.1 (2) that all great circles are geodesics. However, given any point $p \in S^2$ and any tangent vector $v \in T_pS^2$ it is not hard to show that there is a parametrization $c: (-\infty, \infty) \to S^2$ of a great circle such that $c(0) = p$ and $c'(0) = v$. Theorem 7.2.6 now implies that any other geodesic $d: (-\epsilon, \epsilon) \to S^2$ with $d(0) = p$ and $d'(0) = v$ must agree with c on $(-\epsilon, \epsilon)$. ◊

For later use we need the following definition and lemma.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point and let $v \in T_pM$ be a vector. The number ρ_v is defined to be

$$\rho_v = \text{lub}\{r \in \mathbb{R} \mid \text{there is a geodesic } c: (-r, r) \to M \text{ such that } c(0) = p \text{ and } c'(0) = v\}$$

if the set is bounded (so that the least upper bound exists), and $\rho_v = \infty$ if the set is not bounded. ◊

Lemma 7.2.8. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point and let $v \in T_pM$ be a vector. Then the following hold.*

(i) $\rho_v > 0$.

(ii) *There exists a geodesic $c: (-\rho_v, \rho_v) \to M$ such that $c(0) = p$ and $c'(0) = v$.*

(iii) *If $s \in \mathbb{R} - \{0\}$ then $\rho_{sv} = \frac{\rho_v}{|s|}$; if $s = 0$ then $\rho_{sv} = \infty$.*

Proof. Exercise 7.2.7.

Exercises

7.2.1. Which circles of latitude of the torus, parametrized as a surface of revolution as in Section 5.3, are geodesics?

7.2.2. Let $c\colon (a, b) \to \mathbb{R}^2$ be an injective unit-speed curve, and let $M \subset \mathbb{R}^3$ be the right cylinder with cross section the image of the curve c; as mentioned in Section 5.3 the surface M can be parametrized as the rectifying developable surface generated by c. Describe the geodesics on M.

7.2.3*. Suppose that a coordinate patch has $g_{11} = 1$, $g_{12} = 0$ and $g_{22} = G$ for some smooth function G. Show that the equations for geodesics in this case are

$$c_1'' - \frac{1}{2}\frac{\partial G}{\partial u_1}(c_2')^2 = 0, \tag{7.2.6}$$

$$c_2'' + \frac{1}{G}\frac{\partial G}{\partial u_1}c_1'c_2' + \frac{1}{2G}\frac{\partial G}{\partial u_2}(c_2')^2 = 0. \tag{7.2.7}$$

7.2.4*. Let $c\colon (a, b) \to \mathbb{R}^3$ be a smooth, injective, unit-speed curve. Show that c is a geodesic on the rectifying developable surface generated by it.

7.2.5*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let p be a point in M, let $v \in T_pM$ be a vector and let $c\colon (-\epsilon, \epsilon) \to M$ be a geodesic such that $c(0) = p$ and $c'(0) = v$. If $\lambda > 0$ is any number, define the curve $\widetilde{c}\colon (-\epsilon/\lambda, \epsilon/\lambda) \to M$ by $\widetilde{c}(t) = c(\lambda t)$. Show that $\widetilde{c}$ is the unique geodesic such that $\widetilde{c}(0) = p$ and $\widetilde{c}'(0) = \lambda v$.

7.2.6*. The goal of this exercise is to prove Lemma 7.2.3. The proof is broken down into steps.

(1) The map c has constant non-zero speed by Lemma 7.2.2 (i).

(2) Let $h\colon [0, 1] \to A$ be a homeomorphism (guaranteed by the definition of an arc). Consider the map $h^{-1} \circ c\colon [x, y] \to [0, 1]$. Suppose that $c|[x, y]$ is not injective, so that there are points $u, v \in [x, y]$ with $u < v$ and $c(u) = c(v)$. Then $h^{-1} \circ c(u) = h^{-1} \circ c(v)$. There are now two cases

Case (a): The map $h^{-1} \circ c|[u, v]$ is a constant map. Derive a contradiction.

Case (b): The map $h^{-1} \circ c|[u, v]$ is not a constant map. Use Exercise 1.6.10 to find a point $z \in (u, v)$ such that $h^{-1} \circ c$ is not injective on any open neighborhood of z. Deduce that c is not injective on any open neighborhood of z. Use Exercise 4.2.5 to show that c is injective on some open neighborhood of z, a contradiction. Hence $c|[x, y]$ is not injective.

(3) Now suppose that $c(x)$ is not an endpoint of A. Hence $h^{-1} \circ c(x)$ is neither 0 nor 1. Use the bijectivity of $c|[x, y]\colon [x, y] \to A$ and the Intermediate Value Theorem (Theorem 1.5.4) to derive a contradiction.

7.2.7*. Prove Lemma 7.2.8.

7.3 Shortest Paths

Though a geodesic arc need not be a shortest path, we now show that a shortest path must be a geodesic arc. Intuitively, if a path between two points is the shortest possible between the points, then if we travelled along the path we would not feel ourselves veering to the right or the left, since doing so would presumably increase the length of the path.

Theorem 7.3.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p, q \in M$ be points and let A be a regular arc in M with endpoints p and q that has the shortest length of all regular arcs with endpoints p and q. If $c\colon (d, e) \to M$ is a regular curve such that $A = c([a, b])$ for some closed interval $[a, b] \subset (d, e)$, then $c|(a, b)$ is a geodesic.*

The proof of this theorem will be given after some discussion and a lemma. We follow [DO1]. (An alternate proof is found in Exercise 8.A1.3, making use of some technical tools we develop in Chapter VIII.) In analogy to maximum–minimum problems in Calculus, we will essentially characterize curves of minimal length by taking the derivative of the length function of all the curves between the given endpoints, and then setting the derivative equal to zero. The collection of all such curves cannot be parametrized by a single variable, however, but it will suffice to consider a smaller family of curves that can be parametrized by a single parameter. We start by constructing this family of curves and proving a lemma.

Let $M \subset \mathbb{R}^3$ be a smooth surface, let $x\colon U \to M$ be a coordinate patch and let $c\colon (d, e) \to x(U)$ be a unit speed curve; we let s denote the variable in c. We may assume that c is unit speed, since every curve can be reparametrized as a unit speed curve, and we know that if a curve is a geodesic it must have constant speed. Let $\bar{c}$, c_1 and c_2 be as in Section 5.2. Next, suppose we have two smooth functions $\varphi_1, \varphi_2\colon [x, y] \to \mathbb{R}$ for some $x, y \in (d, e)$ with $x < y$ such that $\varphi_i(x) = 0 = \varphi_i(y)$ for $i = 1, 2$; in the proof of Theorem 7.3.1 we will choose particular functions φ_1, φ_2, but for now it does not matter what these functions are. By definition $\binom{c_1(s)}{c_2(s)} \in U$ for all $s \in (d, e)$; it follows from Exercise 1.6.13 that there is some number $\epsilon > 0$ such that

$$\begin{pmatrix} c_1(s) + t\varphi_1(s) \\ c_2(s) + t\varphi_2(s) \end{pmatrix} \in U$$

for all $\binom{s}{t} \in [x, y] \times (-\epsilon, \epsilon)$. Hence the function $\alpha: [x, y] \times (-\epsilon, \epsilon) \to x(U)$ given by

$$\alpha\left(\begin{pmatrix} s \\ t \end{pmatrix}\right) = x\left(\begin{pmatrix} c_1(s) + t\varphi_1(s) \\ c_2(s) + t\varphi_2(s) \end{pmatrix}\right) \tag{7.3.1}$$

is well-defined.

We can think of the function $\alpha(\binom{s}{t})$ as a parametrized family of curves, with one curve in the variable s for each value of t; this family of curves forms a variation of the restriction of the original curve $c(s)$ to $[x, y]$. Observe that $\alpha(\binom{s}{0}) = c(s)$ for all $s \in [x, y]$, and that $\alpha(\binom{x}{t}) = c(x)$ and $\alpha(\binom{y}{t}) = c(y)$ for all $t \in (-\epsilon, \epsilon)$. See Figure 7.3.1. For later use also note that

$$\frac{\partial \alpha}{\partial s}|_{t=0} = c'(s) \quad \text{and} \quad \frac{\partial \alpha}{\partial t}|_{s=x} = 0 = \frac{\partial \alpha}{\partial t}|_{s=y}. \tag{7.3.2}$$

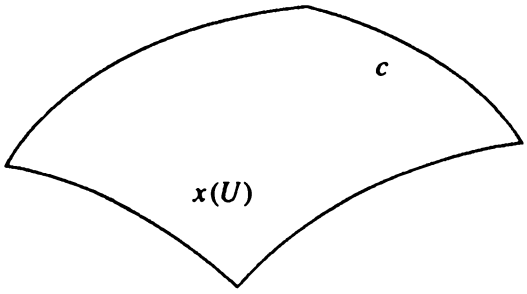


Figure 7.3.1

Next, for each fixed t we can compute the length of the curve $\alpha(\binom{s}{t})$ with respect to the variable s, thus defining a length function $\Lambda: (-\epsilon, \epsilon) \to \mathbb{R}$, where $\Lambda(t)$ is the length of the curve $\alpha(\binom{s}{t})$. More explicitly, we have

$$\Lambda(t) = \int_x^y \left\| \frac{\partial \alpha}{\partial s} \right\| ds. \tag{7.3.3}$$

Observe that $\Lambda(0)$ is simply the length of $c|[x, y]$. The following lemma shows how to differentiate Λ.

Lemma 7.3.2. *Let $\Lambda\colon (-\epsilon, \epsilon) \to \mathbb{R}$ be as above. Then there is a number δ such that $0 < \delta \leq \epsilon$ and the function $\Lambda(t)$ is differentiable on $(-\delta, \delta)$; differentiation can be computed under the integral sign, that is*

$$\Lambda'(t) = \int_a^b \frac{\partial}{\partial t} \| \frac{\partial \alpha}{\partial s} \| \, ds.$$

Proof. Since c is a unit speed curve we see that

$$\| \frac{\partial \alpha}{\partial s}|_{t=0} \| = \|c'(s)\| = 1,$$

for all $s \in [x, y]$. We can thus apply Exercise 1.6.7 to the continuous function $\| \frac{\partial \alpha}{\partial s} \|$ to deduce that there are numbers $M, \delta > 0$ such that $\delta \leq \epsilon$ and $\| \frac{\partial \alpha}{\partial s} \| \geq M$ for all $\binom{s}{t} \in [x, y] \times (-\delta, \delta)$. Using the infinite differentiability of $\frac{\partial \alpha}{\partial s}$ for all t and the differentiability of $\| \ \|$ away from the zero vector, we deduce that $\| \frac{\partial \alpha}{\partial s} \|$ is infinitely differentiable in the variable t for all $t \in (-\delta, \delta)$. The lemma now follows from a standard result of advanced Calculus concerning differentiation under the integral sign (see, for example, [BT, §23]). □

We are now ready to prove our main result.

Proof of Theorem 7.3.1. Assume that $c|(a, b)$ is not a geodesic. There is thus some number $s_0 \in (a, b)$ such that $\frac{Dc'(s)}{ds}(s_0) \neq 0$. Assume that $\frac{Dc'(s)}{ds}(s_0) > 0$; the other case is similar. For later use, note that by smoothness $\frac{Dc'(s)}{ds} > 0$ for all s close enough to s_0. Let $x\colon U \to M$ be a coordinate patch with $c(s_0) \in x(U)$. Since $x(U)$ is an open subset of M by Proposition 5.2.5, and since c is continuous, there is some number $\eta > 0$ such that $(s_0 - \eta, s_0 + \eta)c^{-1}(x(U))$. Let $x = s_0 - \frac{\eta}{2}$ and $y = s_0 + \frac{\eta}{2}$, so that $c([x, y]) \subset x(U)$. If $c|[a, b]$ is the shortest smooth path from $c(a)$ to $c(b)$, then $c|[x, y]$ must be the shortest smooth curve from $c(x)$ to $c(y)$; for otherwise, find a shorter smooth curve from $c(x)$ to $c(y)$ and splice it into the original curve c, yielding a smooth curve from $c(a)$ to $c(b)$ that is shorter than c (we may have to "smooth out" the corners where we splice, but that is not a serious problem). We now restrict our attention to the interval $[x, y]$. Since the rest of the domain of c will have no role from now on, for ease of notation we will simply use c to denote $c|[x, y]$ (though it will still make sense to take derivatives at x and y).

Let $\lambda\colon [x, y] \to \mathbb{R}$ be a smooth function such that $\lambda(x) = \lambda(y) = 0$, that $\lambda(s_0) > 0$ and that $\lambda(s) \geq 0$ for all $s \in [x, y]$ (there are many such functions, so pick one). Note that

$$\lambda(s)\left|\frac{Dc'(s)}{ds}\right|^2 \geq 0 \tag{7.3.4}$$

for all $s \in [x, y]$, and

$$\lambda(s)\left|\frac{Dc'(s)}{ds}\right|^2 > 0 \tag{7.3.5}$$

on some interval containing s_0.

Since $\frac{Dc'(s)}{ds}$ is a tangent vector to M for all $s \in [x, y]$, then so is $\lambda(s)\frac{Dc'(s)}{ds}$. Using an argument similar to the proof of Exercise 5.7.1, there must exist smooth functions $\varphi_1, \varphi_2 : [x, y] \to \mathbb{R}$ such that

$$\lambda(s)\frac{Dc'(s)}{ds} = \sum_{i=1}^{2} \varphi_i(s)\, x_i(\bar{c}(s)) \tag{7.3.6}$$

for all $s \in [x, y]$. Observe that $\varphi_i(x) = 0 = \varphi_i(y)$ for $i = 1, 2$ because of the hypotheses on λ. Using these particular functions φ_1 and φ_2 we can define the function $\alpha : [x, y] \times (-\epsilon, \epsilon) \to x(U)$ as in Equation 7.3.1, where $\epsilon > 0$ is some sufficiently small number so that α is well-defined. In addition to the properties of α mentioned above, we note that by the definition of α, the chain rule and Equation 7.3.6 we have

$$\frac{\partial \alpha}{\partial t}|_{t=0} = \sum_{i=1}^{2} \varphi_i(s)\, x_i(\bar{c}(s)) = \lambda(s)\frac{Dc'(s)}{ds}. \tag{7.3.7}$$

We can define the length function $\Lambda(t)$ just as before, and this function is differentiable on some open interval $(-\delta, \delta) \subset (-\epsilon, \epsilon)$ by Lemma 7.3.2. Since c has the minimal length of all smooth curves from $c(x)$ to $c(y)$ by hypothesis, and since $\Lambda(0)$ equals the length of c, it follows from a standard result in Calculus that $\Lambda'(0) = 0$.

The remainder of the proof will be a calculation that shows that $\Lambda'(0) < 0$, the desired contradiction. From Lemma 7.3.2 we have

$$\Lambda'(t) = \int_x^y \frac{\partial}{\partial t}\left\langle \frac{\partial \alpha}{\partial s}, \frac{\partial \alpha}{\partial s}\right\rangle^{1/2} ds = \int_x^y \frac{1}{2}\frac{\frac{\partial}{\partial t}\left\langle \frac{\partial \alpha}{\partial s}, \frac{\partial \alpha}{\partial s}\right\rangle}{\left\langle \frac{\partial \alpha}{\partial s}, \frac{\partial \alpha}{\partial s}\right\rangle^{1/2}} ds. \tag{7.3.8}$$

Applying Lemmas 5.6.8, 5.7.8 and 5.6.8 again to the numerator inside the final integral in Equation 7.3.8, we obtain

$$\begin{aligned}\frac{\partial}{\partial t}\langle\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial s}\rangle &= 2\langle\frac{D}{\partial t}\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial s}\rangle = 2\langle\frac{D}{\partial s}\frac{\partial\alpha}{\partial t},\frac{\partial\alpha}{\partial s}\rangle \\ &= 2\Big\{\frac{\partial}{\partial s}\langle\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle - \langle\frac{D}{\partial s}\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle\Big\}.\end{aligned} \tag{7.3.9}$$

By Equation 7.3.2 and the fact that c is a unit speed curve we deduce that

$$\langle\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial s}\rangle|_{t=0} = \langle c'(s), c'(s)\rangle = 1. \tag{7.3.10}$$

Plugging the value $t = 0$ into the final integral in Equation 7.3.8, and then using Equations 7.3.9, 7.3.10, 7.3.2 and 7.3.7 we compute

$$\begin{aligned}\Lambda'(0) &= \int_x^y \Big\{\frac{\partial}{\partial s}\langle\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle|_{t=0} - \langle\frac{D}{\partial s}\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle|_{t=0}\Big\}\,ds \\ &= \langle\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle|_{t=0}\,]_{s=x}^{s=y} - \int_x^y \langle\frac{D}{\partial s}\frac{\partial\alpha}{\partial s},\frac{\partial\alpha}{\partial t}\rangle|_{t=0}\,ds \\ &= 0 - \int_x^y \langle\frac{Dc'(s)}{\partial s},\lambda(s)\frac{Dc'(s)}{ds}\rangle\,ds \\ &\qquad\qquad \text{by Equation 7.3.2 and Equation 7.3.7,} \\ &= -\int_x^y \lambda(s)\Big|\frac{Dc'(s)}{ds}\Big|^2\,ds < 0,\end{aligned}$$

where the last inequality holds by Equations 7.3.4 and 7.3.5. We have thus obtained the desired contradiction. □

Exercises

7.3.1. Suppose that a smooth surface $M \subset \mathbb{R}^3$ contains a straight line (for example, a saddle surface, not to mention ruled surfaces). Is the straight line a geodesic?

Endnotes

Notes for Section 7.2

(A) The concept of a tangent vector field along a curve in a surface that stays parallel to itself from the point of view of a creature on the surface is called

parallel transport, and is important in differential geometry. See [DO1, §4-4] or [M-P, §4-6].

(B) For further discussion of geodesics on surfaces of revolution, see [DO1, pp. 255–260] or [SK3 vol. III, pp. 314–319].

(C) The precise idea that a surface has no "holes," needed to guarantee the existence of a length-minimizing geodesic joining any two points on the surface, is the notion of topological completeness (also known as Cauchy completeness); see [MU2, §7-1]. A compact surface (with no boundary, as we are always assuming), is topologically complete.

Notes for Section 7.3

The method of proof used in this section, which is quite standard, has other uses as well; see [DO1, §5-4] and [OS1].

CHAPTER VIII

The Gauss–Bonnet Theorem

8.1 Introduction

Smooth surfaces can be analyzed geometrically (as in Chapters 6 and 7) and topologically (since they are also topological surfaces). The Gauss–Bonnet Theorem, essentially the point toward which this entire book has been aimed, shows that these two approaches are deeply related. The simplicial Gauss–Bonnet Theorem (Theorem 3.7.2) has already shown us one connection between a topological invariant (the Euler characteristic) and a geometric quantity (the angle defect in simplicial surfaces). Although the statement of this theorem was somewhat surprising, the proof was not difficult, since both angle defect and the Euler characteristic were defined in terms of triangulations; indeed, the real surprise was that the Euler characteristic turns out to be a topological invariant, that is, it does not depend upon the choice of triangulation of a given surface.

The statement of the Gauss–Bonnet Theorem (which always refers to the smooth case when unadorned with any adjective such as "simplicial") is very similar to the simplicial version: The total Gaussian curvature of a compact smooth surface (obtained by integrating the curvature function) is related to the Euler characteristic of the surface (thought of as a compact topological surface). Unlike the simplicial case, Gaussian curvature was defined in a way that has absolutely nothing to do with triangulations, so the connection between Gaussian curvature and a triangulation of a smooth surface is much more subtle, and much harder to prove.

The precise statement of the result we have all been waiting for is the following.

Theorem 8.1.1 (Gauss–Bonnet Theorem). *Let $M \subset \mathbb{R}^3$ be a compact smooth surface. Then*

$$\int_M K\,dA = 2\pi\chi(M).$$

Before proceeding with the proof of the Gauss–Bonnet Theorem in Section 8.4, we develop some admittedly technical material, which gives us a special type of coordinate patch with nice metric coefficients.

8.2 The Exponential Map

Consider a smooth surface $M \subset \mathbb{R}^3$ and the tangent plane T_pM at a point $p \in M$. Think of T_pM as being translated in $\mathbb{R}^3$ so that its origin is at p, and suppose that it is made out of flexible material. Then we can imagine wrapping the tangent plane around the surface, as in Figure 8.2.1. Although this wrapping process works nicely near the point p, we might run into trouble away from p, where there might be a hole in the surface, or where the wrapped tangent plane might be forced to intersect itself. The exponential map, a map from an open neighborhood of the origin of T_pM onto an open neighborhood of p in M, is a formalization of this wrapping process. There is one exponential map for each point in the surface. (Do not be misled by the name of this map — it has little to do with the exponential and logarithmic functions one encounters in high school; in one very specific case concerning spaces of matrices one does use the power series for e^x applied to matrices, and the name has stuck for the more general situation.)

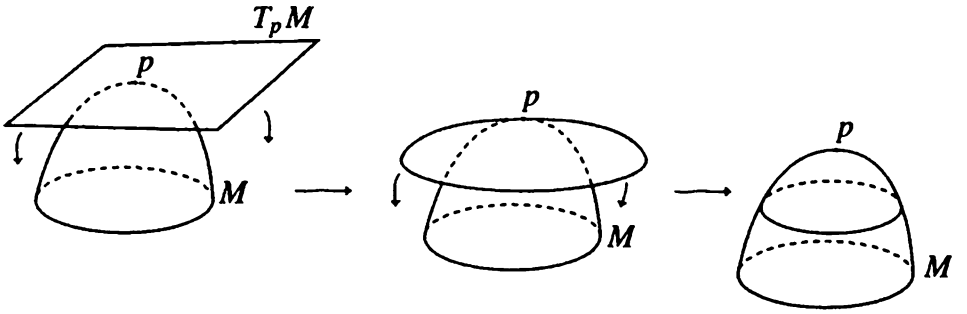


Figure 8.2.1

The exponential map is cleverly defined using geodesics. We use the definition and properties of ρ_v in Section 7.2.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. The set $E_p \subset T_pM$ is defined to be

$$E_p = \{v \in T_pM \mid \rho_v > 1\}. \quad \Diamond$$

Instead of the number 1 in the above definition we could have chosen any positive number, though 1 is both convenient and quite standard. From the

definition of ρ_v it is seen that the origin is always in E_p for any p. The following lemma shows that E_p always contains more than the origin.

Lemma 8.2.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Then the following hold.*

(i) *If $v \in E_p$ and $s \in \mathbb{R}$ is a number, then $sv \in E_p$ iff $-\rho_v < s < \rho_v$.*
(ii) *If $u \in T_pM$ is a unit vector, then*

$$E_p \cap \{su \mid s \in \mathbb{R}\} = \{su \mid -\rho_u < s < \rho_u\}.$$

Proof. Exercise 8.2.1. □

It will be seen in Proposition 8.2.3 that E_p contains an open disk centered at the origin. We cannot derive this fact just using the above lemma, since we don't know whether ρ_v varies continuously with $v \in T_pM$. Even without Proposition 8.2.3 we can make the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. The map $\exp_p: E_p \to M$, called the **exponential map** at p, is defined as follows. For each $v \in E_p$ let $c_v: (-\rho_v, \rho_v) \to M$ be the unique geodesic such that $c_v(0) = p$ and $c_v'(0) = v$; define $\exp_p(v)$ by setting $\exp_p(v) = c_v(1)$. ◊

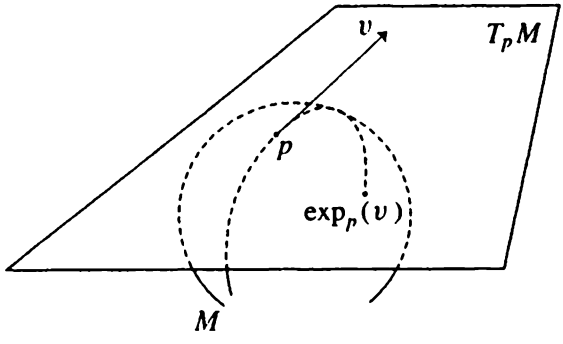


Figure 8.2.2

Note that $\exp_p(0) = p$. The following lemma shows that geodesics in M through p are simply the images under $\exp_p$ of lines through the origin in T_pM.

Lemma 8.2.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let $v \in E_p$ be a vector. Then the map $d\colon (-\rho_v, \rho_v) \to M$ defined by $d(s) = exp_p(sv)$ is the unique geodesic with $d(0) = p$ and $d'(0) = v$. The length of the curve d from p to $exp_p(v)$ equals $\|v\|$.*

Proof. Let $c_v\colon (-\rho_v, \rho_v) \to M$ be the unique geodesic such that $c_v(0) = p$ and $c_v'(0) = v$. Let $s \in (-\rho_v, \rho_v)$ be a fixed non-zero number. Define the curve $\widetilde{c}\colon (-\rho_v/s, \rho_v/s) \to M$ by $\widetilde{c}(t) = c_v(st)$. By Exercise 7.2.5 we know that $\widetilde{c}$ is the unique geodesic such that $\widetilde{c}(0) = p$ and $\widetilde{c}'(0) = sv$. We can thus use $\widetilde{c}$ to compute $\exp_p(sv)$; thus

$$\exp_p(sv) = \widetilde{c}(1) = c_v(s \cdot 1) = c_v(s).$$

Hence $d(s) = c_v(s)$ for all non-zero $s \in (-\rho_v, \rho_v)$. We also know that $d(0) = p = c_v(0)$. Therefore $d = c_v$, and thus d is the unique geodesic with $d(0) = p$ and $d'(0) = v$.

By Lemma 7.2.2 (i) we know that d has constant speed. Since $d'(0) = v$ it follows that $\|d'(s)\| = \|d'(0)\| = \|v\|$ for all s. Using Equation 4.3.1 we see that the length of d from $p = d(0)$ to $\exp_p(v) = d(1)$ is

$$\int_0^1 \|d'(s)\| ds = \|v\|. \quad \square$$

The exponential map need not be injective as defined. Consider, for example, the surface S^2; for each point p the set E_p is in fact all of T_pS^2, since geodesics are great circles and they can be extended indefinitely. However, if we allow arbitrarily large vectors in the domain of $\exp_p$, then the map will not be injective, since all the geodesics starting at p will go through the antipodal point to p if extended far enough. The following proposition shows, among other things, that if we restrict our attention to a small enough neighborhood of the origin then $\exp_p$ will be injective. The proof of this result is rather long and involved, and is in Appendix A8.1. Observe first that any open subset of a plane in $\mathbb{R}^3$ is a smooth surface in $\mathbb{R}^3$; hence it makes sense to discuss the smoothness of the map $\exp_p$.

Proposition 8.2.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Then there exists an open set $W \subset M$ containing p and a number $\delta_p > 0$ such that for every $q \in W$ the following properties hold:*

(i) *the set E_q contains the open disk $O_{\delta_p}(O_3, T_qM)$;*

(ii) *the set* $exp_p(O_{\delta_p}(O_3, T_qM))$ *is open in* M;
(iii) *the map* $exp_q|O_{\delta_p}(O_3, T_qM)$ *is a diffeomorphism from the set* $O_{\delta_p}(O_3, T_qM)$ *onto* $exp_p(O_{\delta_p}(O_3, T_qM))$;
(iv) $exp_q(O_{\delta_p}(O_3, T_qM)) \supset W$.

The following three results can be deduced from the statement and the proof of Proposition 8.2.3; details are left to the reader who has read the proof of this Proposition.

Corollary 8.2.4. *Let* $M \subset \mathbb{R}^3$ *be a smooth surface, let* $p \in M$ *be a point and let* ϵ *be any number such that* $0 < \epsilon \leq \delta_p$. *Then there exists an open set* $V \subset M$ *containing* p *such that the following properties hold:*

(i) *any two points in* V *can be joined by a unique geodesic arc of length less than* ϵ;
(ii) $V \subset exp_q(O_{\delta_p}(O_3, T_qM))$ *for all* $q \in V$.

(If $\epsilon = \delta_p$ *then* $V = W$ *works.)*

Proof. Exercise 8.2.3. □

It is not guaranteed in the above corollary that the geodesic arc between any two points in the set W will lie entirely in W; hence W need not be what would reasonably be called "geodesically convex." Theorem A8.1.2 says that in fact every point in a smooth surface does have a geodesically convex neighborhood; we will use Corollary 8.2.4 in the proof of this stronger result.

Corollary 8.2.5. *Let* $M \subset \mathbb{R}^3$ *be a compact smooth surface. Then there exists a number* $\delta_M > 0$ *such that for each point* $p \in M$

(i) *the set* E_p *contains the open disk* $O_{\delta_M}(O_3, T_pM)$;
(ii) *the set* $exp_p(O_{\delta_M}(O_3, T_pM))$ *is open in* M;
(iii) *the map* $exp_p|O_{\delta_M}(O_3, T_pM)$ *is a diffeomorphism from the set* $O_{\delta_M}(O_3, T_pM)$ *onto* $exp_p(O_{\delta_M}(O_3, T_pM))$;
(iv) *there is an open set* $W \subset M$ *containing* p *such that* $exp_q(O_{\delta_M}(O_3, T_qM)) \supset W$ *for all* $q \in W$.

Proof. Exercise 8.2.5. □

Corollary 8.2.6. *Let* $M \subset \mathbb{R}^3$ *be a compact smooth surface and let* δ_M *be as in Corollary 8.2.5. Then there is a number* $\epsilon_M > 0$ *such that for each point* $p \in M$ *we have*

$$O_{\epsilon_M}(p, M) \subset exp_q(O_{\delta_M}(O_3, T_qM))$$

for every $q \in O_{\epsilon_M}(p, M)$.

Proof. Exercise 8.2.6. □

Our discussion concerning the exponential map should look somewhat familiar: The exponential map is a diffeomorphism from some open subset of a plane (namely an open disk in the tangent plane) to a subset of the surface. This phenomenon is very similar to a coordinate patch, which is a smooth map from an open set $U \subset \mathbb{R}^2$ onto a subset of the surface, subject to certain conditions (which turn out to be equivalent, via the Inverse Function Theorem, to being a diffeomorphism). Thus, the exponential map restricted to a small enough disk in the tangent plane is essentially as good as a coordinate patch, except that it is from the tangent plane instead of from a subset of $\mathbb{R}^2$ (and the latter is really nothing but a particular choice of a plane in $\mathbb{R}^3$ and a choice of coordinate axes for this plane). We remedy this situation as follows.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, let δ_p be as in Proposition 8.2.3 and let $\Upsilon_p: \mathbb{R}^2 \to T_pM$ be any choice of an orthogonal linear map. The **exponential coordinate patch** at p is the map

$$\mathrm{Exp}_p = \exp_p \circ \Upsilon_p: O_{\delta_p}(O_2, \mathbb{R}^2) \to M. \quad \Diamond$$

The choice of orthogonal map in the above definition is arbitrary; assume that such a map has been chosen once and for all for each point in each smooth surface. That Exp_p is indeed a coordinate patch follows immediately from Exercise 5.2.10 and the fact that a non-singular linear map defined on an open subset of a plane in $\mathbb{R}^3$ is a diffeomorphism from its domain onto its image. As seen in the following lemma, the metric coefficients of Exp_p are nicely behaved at the origin for any point on a surface.

Lemma 8.2.7. *Let* $M \subset \mathbb{R}^3$ *be a smooth surface, and let* $p \in M$ *be a point. The metric coefficients of* Exp_p *at* O_2 *are* $E(O_2) = G(O_2) = 1$ *and* $F(O_2) = 0$.

Proof. As usual we let $\{u_1, u_2\}$ denote the standard basis for $\mathbb{R}^2$. We then compute

$$(\mathrm{Exp}_p)_i(O_2) = \frac{d\,\mathrm{Exp}_p(tu_i)}{dt}|_{t=0} = \frac{d\exp_p(\Upsilon_p(tu_i))}{dt}|_{t=0}$$
$$= \frac{d\exp_p(t\Upsilon_p(u_i))}{dt}|_{t=0} = \Upsilon_p(u_i),$$

where the last equality follows from Lemma 8.2.2. Since Υ_p is an orthogonal map, we know that the set $\{\Upsilon_p(u_1), \Upsilon_p(u_2)\}$ is orthonormal, and the lemma follows immediately. □

We conclude this section with the following definition, which is seen to make sense in light of Proposition 8.2.3.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, let δ_p be as in Proposition 8.2.3 and let r be a number such that $0 \leq r < \delta_p$. The **geodesic circle** of radius r centered at p, the **open geodesic ball** of radius r centered at p and the **closed geodesic ball** of radius r centered at p are the images under $\exp_p$ of the circle in T_pM of radius r centered at the origin, of the open ball in T_pM of radius r centered at the origin, and of the closed ball in T_pM of radius r centered at the origin respectively; these three sets are denoted $GS_r(p, M)$, $GO_r(p, M)$ and $\overline{GO}_r(p, M)$. ◊

It follows from Proposition 8.2.3 that $GS_r(p, M)$, $GO_r(p, M)$ and $\overline{GO}_r(p, M)$ are respectively a 1-sphere, the interior of a disk, and a disk. Moreover, combining Proposition 8.2.3 with Theorem 7.2.6, it follows that $GS_r(p, M)$, $GO_r(p, M)$ and $\overline{GO}_r(p, M)$ consist respectively of all points in M that can be joined to p by a geodesic arc of length r, of length less than r, and of length less than or equal to r.

Exercises

8.2.1*. Prove Lemma 8.2.1.

8.2.2*. Let $A \subset \mathbb{R}^2 \times \mathbb{R}^2 = \mathbb{R}^4$ be an open set, and let $f: A \to \mathbb{R}^2 \times \mathbb{R}^2$ be a smooth map such that $f(A)$ is open in $\mathbb{R}^2 \times \mathbb{R}^2$ and f is a diffeomorphism from A onto $f(A)$. Suppose further that f has the form $f((\bar{p}, \bar{q})) = (\bar{p}, h(\bar{p}, \bar{q}))$ for some smooth map $h: A \to \mathbb{R}^2$, where $(\bar{p}, \bar{q}) \in A$. If $\bar{p} \in \mathbb{R}^2$ is a point, let $J_{\bar{p}}: \mathbb{R}^2 \to \mathbb{R}^2 \times \mathbb{R}^2$ denote the map $J_{\bar{p}}(\bar{q}) = (\bar{p}, \bar{q})$, and let $P_2: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ denote projection onto the second factor. Suppose $B \subset \mathbb{R}^2$ is an open set

such that $J_p(B) \subset A$. Show that $P_2 \circ f \circ J_{\bar{p}}(B)$ is open in $\mathbb{R}^2$ and that $P_2 \circ f \circ J_{\bar{p}}|B: B \to \mathbb{R}^2$ is a diffeomorphism from B onto $P_2 \circ f \circ J_{\bar{p}}(B)$.

8.2.3*. Prove Corollary 8.2.4 (this uses the proof of Proposition 8.2.3).

8.2.4. For the unit sphere S^2, what is the largest possible value of δ_{S^2}, the existence of which is guaranteed by Corollary 8.2.5?

8.2.5*. Prove Corollary 8.2.5 (this uses the proof of Proposition 8.2.3).

8.2.6*. Prove Corollary 8.2.6 (this uses the proof of Proposition 8.2.3).

8.3 Geodesic Polar Coordinates

We wish to introduce the idea of polar coordinates on a surface; we do this by introducing such coordinates in $\mathbb{R}^2$ and then using the exponential coordinate patch.

Definition. Let rect: $\mathbb{R}^2 \to \mathbb{R}^2$ be the function given by

$$\text{rect}\left(\begin{pmatrix} R \\ \theta \end{pmatrix}\right) = \begin{pmatrix} R\cos\theta \\ R\sin\theta \end{pmatrix}.$$

If $M \subset \mathbb{R}^3$ is a smooth surface, $p \in M$ is a point and δ_p is as in Proposition 8.2.3, the **geodesic polar coordinate patch** at p is the map

$$D_p = \text{Exp}_p \circ \text{rect}: (0, \delta_p) \times (0, 2\pi) \to M. \quad \Diamond$$

The fact that the map D_p is a coordinate patch follows from Lemma 5.2.6 and the facts that Exp_p is a coordinate patch and $\text{rect}|(0, \delta_p) \times (0, 2\pi)$ is a diffeomorphism from $(0, \delta_p) \times (0, 2\pi)$ onto an open subset of $\mathbb{R}^2$. Although we chose $(0, \delta_p) \times (0, 2\pi)$ as the domain of D_p (in order to insure that rect is injective), we could just as well have chosen any domain of the form $(0, \delta_p) \times (-\lambda, 2\pi - \lambda)$ for $\lambda \in \mathbb{R}$; we will need this flexibility in our choice of domain. We will denote elements of the domain of D_p by $\binom{R}{\theta}$. Observe that by Lemma 8.2.2 the lines in $(0, \delta_p) \times (0, 2\pi)$ of the form $\theta = k$ for any constant k are mapped by D_p to geodesics in M which converge to, but do not contain, the

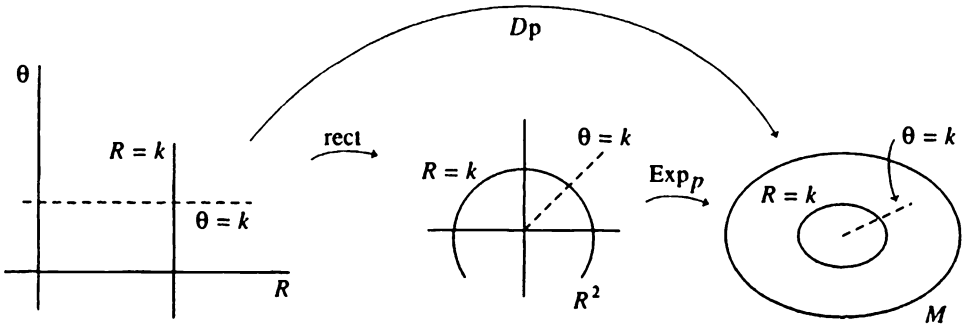


Figure 8.3.1

point p. Each line in $(0, \delta_p) \times (0, 2\pi)$ of the form $R = k$ for a constant k is mapped by D_p to a geodesic circle of radius k. See Figure 8.3.1.

The one major drawback of geodesic polar coordinate patches is that the image of D_p does not contain the point p and a geodesic ray starting at p. (This problem arises from the need to keep rect, and hence D_p, injective and defined on an open set.) The choice of domain of D_p as mentioned above is equivalent to a choice of which geodesic ray starting at p is to be excluded from the image of D_p. We start our discussion of geodesic polar coordinates with the following lemma.

Lemma 8.3.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let δ_p and W be as in Proposition 8.2.3. For any point $q \in W$ such that $q \neq p$, there is a unique number $d > 0$ for which the following criteria all hold.*

(1) *The unique geodesic from p to q has length d.*
(2) *$q \in GS_d(p, M)$.*
(3) *If D_p is a choice of a geodesic polar coordinate patch, the image of which contains q, then $(D_p)^{-1}(q) = \binom{d}{\theta}$ for some number θ.*

Proof. Exercise 8.3.7. □

Using this lemma we can make the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let δ_p and W be as in Proposition 8.2.3. For any point $q \in W$ such that $q \neq p$, the number d as in Lemma 8.3.1 is the ***R*-coordinate** of q with respect to p. ◊

The nice feature of geodesic polar coordinate patches are their metric coefficients (not just at one point, as for Exp_p, but throughout its domain). We will let $(D_p)_1$ denote $\frac{\partial D_p}{\partial R}$ and $(D_p)_2$ denote $\frac{\partial D_p}{\partial \theta}$, and similarly for higher order partial derivatives. For this section and the next we will let E, F and G denote the metric coefficients of D_p, and we will let $\overline{E}$, $\overline{F}$ and $\overline{G}$ denote the metric coefficients of Exp_p.

Proposition 8.3.2 (Gauss' Lemma). *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Then $E(\begin{pmatrix} R \\ \theta \end{pmatrix}) = 1$, and $F(\begin{pmatrix} R \\ \theta \end{pmatrix}) = 0$, and $G(\begin{pmatrix} R \\ \theta \end{pmatrix}) > 0$ for all $\begin{pmatrix} R \\ \theta \end{pmatrix} \in (0, \delta_p) \times (0, 2\pi)$.*

Proof. The proof follows [DO1]. At times we will drop the arguments in the metric coefficients E, F and G. We start out by defining two types of curves in the image of D_p. For fixed $\theta \in (0, 2\pi)$, let $\gamma_\theta\colon (0, \delta_p) \to M$ be defined by $\gamma_\theta(s) = D_p(\begin{pmatrix} s \\ \theta \end{pmatrix})$ for all $s \in (0, \delta_p)$; for fixed $R \in (0, \delta_p)$, let $\alpha_R\colon (0, 2\pi) \to M$ be defined by $\alpha_R(t) = D_p(\begin{pmatrix} R \\ t \end{pmatrix})$ for all $t \in (0, 2\pi)$. As mentioned above γ_θ is a geodesic and α_R is a geodesic circle of radius R. Observe that $\gamma_\theta'(s) = (D_p)_1(\begin{pmatrix} s \\ \theta \end{pmatrix})$ and $\alpha_R'(t) = (D_p)_2(\begin{pmatrix} R \\ t \end{pmatrix})$. The coordinate patch D_p is only defined for $R > 0$, and hence $\gamma_\theta(s)$ is only defined as given for $s > 0$, and $\alpha_R(t)$ is only defined as given for $R > 0$. By extending the domain of D_p slightly (in which case D_p would still be smooth, though not injective), it can be seen that each geodesic $\gamma_\theta(s)$ can be extended smoothly to include $s = 0$. Note that $\lim_{s \to 0} \gamma_\theta'(s)$ exists for all θ. Similarly, the family of geodesic circles $\alpha_R(t)$, one such circle for each $R > 0$, can be extended smoothly to include the circle of radius $R = 0$, namely the constant map $\alpha_0(t) = p$. It can be verified that $\lim_{R \to 0} \alpha_R'(t) = \alpha_0'(t) = 0$ for all t.

Fix $\theta \in (0, 2\pi)$. By the above observations we note that $E(\begin{pmatrix} s \\ \theta \end{pmatrix}) = \|(D_p)_1(\begin{pmatrix} s \\ \theta \end{pmatrix})\|^2 = \|\gamma_\theta'(s)\|^2$ for all appropriate values of s. Observe that

$$\gamma_\theta(s) = D_p(\begin{pmatrix} s \\ \theta \end{pmatrix}) = \text{Exp}_p \circ \text{rect}(\begin{pmatrix} s \\ \theta \end{pmatrix}) = \text{Exp}_p(\begin{pmatrix} s\cos\theta \\ s\sin\theta \end{pmatrix}),$$

and hence

$$\gamma_\theta'(s) = \cos\theta\,(\text{Exp}_p)_1 + \sin\theta\,(\text{Exp}_p)_2,$$

so that

$$\|\gamma_\theta'(s)\|^2 = \cos^2\theta\,\overline{E} + 2\sin\theta\cos\theta\,\overline{F} + \sin^2\theta\,\overline{G},$$

where the partial derivatives $(\mathrm{Exp}_p)_i$ and the metric coefficients $\overline{E}$, $\overline{F}$ and $\overline{G}$ are evaluated at $\binom{s\cos\theta}{s\sin\theta}$. By Lemma 8.2.7 we see that $\|\gamma_\theta'(0)\| = 1$, and by Lemma 7.2.2, (i) it follows that $\|\gamma_\theta'(s)\| = 1$ for all s. Hence E is constantly 1.

We now show that F is independent of R, to be accomplished by showing that $F_1 = \frac{\partial F}{\partial R} = 0$ for all $\binom{R}{\theta}$. As just noted, the curve $\gamma_\theta(s) = D_p(\binom{s}{\theta})$ is a geodesic for any fixed value of θ. This geodesic has coordinate functions $c_1(s) = s$ and $c_2(s) = \theta$. Hence $c_1'(s) = 1$, $c_2'(s) = c_1''(s) = c_2''(s) = 0$. If we plug these coordinate functions into the second part of Equation 7.2.2, we obtain

$$0 + \Gamma_{11}^2 \cdot 1 + 2\Gamma_{12}^2 \cdot 1 \cdot 0 + \Gamma_{22}^2 \cdot 0 = 0.$$

Hence Γ_{11}^2 is constantly 0. Now, applying parts of Equation 5.5.3 to our present situation, and using the fact that E is a constant function, we see that

$$\begin{aligned} \langle (D_p)_{11}, (D_p)_1 \rangle &= \frac{1}{2}E_1 = 0 \\ \langle (D_p)_{11}, (D_p)_2 \rangle &= F_1 - \frac{1}{2}E_2 = F_1. \end{aligned} \tag{8.3.1}$$

Since $(D_p)_1$ and $(D_p)_2$ are in T_pM, it is not hard to see that

$$\langle (D_p)_{11}, (D_p)_1 \rangle = \langle \Pi_{T_pM}\left((D_p)_{11}\right), (D_p)_1 \rangle,$$

and similarly for the second part of Equation 8.3.1. Using an argument similar to the proof of Lemma 5.7.1 it can be verified that

$$\Pi_{T_pM}\left((D_p)_{11}\right) = \Gamma_{11}^1(D_p)_1 + \Gamma_{11}^2(D_p)_2.$$

Combining the above remarks with both parts of Equation 8.3.1, and using the definition of the metric coefficients, yields

$$0 = \langle \Gamma_{11}^1(D_p)_1 + \Gamma_{11}^2(D_p)_2, (D_p)_1 \rangle = \Gamma_{11}^1 E + \Gamma_{11}^2 F, \tag{8.3.2}$$

$$F_1 = \langle \Gamma_{11}^1(D_p)_1 + \Gamma_{11}^2(D_p)_2, (D_p)_2 \rangle = \Gamma_{11}^1 F + \Gamma_{11}^2 G. \tag{8.3.3}$$

Since we saw that $E = 1$ and $\Gamma_{11}^2 = 0$, it now follows from Equation 8.3.2 that $\Gamma_{11}^1 = 0$ as well. Plugging these values for Γ_{11}^1 and Γ_{11}^2 into Equation 8.3.3 we

deduce that $F_1 = 0$. Thus F is independent of R.

Next, we compute

$$\lim_{R\to 0} F\left(\begin{pmatrix} R \\ \theta \end{pmatrix}\right) = \lim_{R\to 0} \langle \gamma'_\theta(R), \alpha'_R(\theta)\rangle = \langle \lim_{R\to 0} \gamma'_\theta(R), \lim_{R\to 0} \alpha'_R(\theta)\rangle = 0,$$

for any $\theta \in (0, 2\pi)$. Since F is independent of R, it follows that F is constantly 0. Finally, since $EG - F^2$ is never zero, and G is always non-negative (for any coordinate patch), using the values for E and F just computed, it follows that $G > 0$ for all (R, θ). □

We need two rather trivial observations. First, the coordinate system D_p has

$$\det(g_{ij}) = G. \tag{8.3.4}$$

Second, the change of variable map from D_p to Exp_p is just the map rect . The Jacobian of rect is the matrix

$$D\,\mathrm{rect} = \begin{pmatrix} \cos\theta & -R\sin\theta \\ \sin\theta & R\cos\theta \end{pmatrix}. \tag{8.3.5}$$

Since geodesic polar coordinates are not defined for $R = 0$, we cannot directly compute G at $R = 0$. We can, however, say something about the behavior of the function $\sqrt{G(\begin{pmatrix} R \\ \theta \end{pmatrix})}$ for fixed θ as R goes to zero.

Lemma 8.3.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Let θ be a fixed number in $(0, 2\pi)$. Then*

$$\lim_{R\to 0} \sqrt{G\left(\begin{pmatrix} R \\ \theta \end{pmatrix}\right)} = 0,$$

$$\lim_{R\to 0} \frac{\partial \sqrt{G}}{\partial R} = 1,$$

$$\lim_{R\to 0} \frac{\partial^2 \sqrt{G}}{\partial R^2} = 0,$$

$$\lim_{R\to 0} \frac{\partial^3 \sqrt{G}}{\partial R^3} = -K(p).$$

Proof. We prove the first two equalities in the lemma, leaving the other two to the reader (Exercise 8.3.1). Using Proposition 8.3.2, Lemma 5.5.3, and Equation 8.3.5, we see that

$$\sqrt{G} = \sqrt{EG - F^2} = \sqrt{\overline{E}\,\overline{G} - \overline{F}^2}\,\det\begin{pmatrix}\cos\theta & -R\sin\theta\\ \sin\theta & R\cos\theta\end{pmatrix} = R\sqrt{\overline{E}\,\overline{G} - \overline{F}^2}, \tag{8.3.6}$$

where E, F, and G are evaluated at $\binom{R}{\theta}$, and $\overline{E}$, $\overline{F}$, and $\overline{G}$ are evaluated at $\binom{R\cos\theta}{R\sin\theta}$. Since $\overline{E}$, $\overline{F}$ and $\overline{G}$ are defined at the origin of $\mathbb{R}^2$, we can think of $\overline{E}(\binom{R\cos\theta}{R\sin\theta})$, etc. as being extended smoothly to $R = 0$. It follows from Lemma 8.2.7 that $\overline{E} = \overline{G} = 1$ and $\overline{F} = 0$ at $R = 0$. Combining these observations with Equation 8.3.6, we have

$$\lim_{R\to 0}\sqrt{G(\binom{R}{\theta})} = \lim_{R\to 0} R\sqrt{\overline{E}\,\overline{G} - \overline{F}^2} = 0$$

and

$$\lim_{R\to 0}\frac{\partial\sqrt{G}}{\partial R} = \lim_{R\to 0}\left\{\sqrt{\overline{E}\,\overline{G} - \overline{F}^2} + R\frac{\partial}{\partial R}\sqrt{\overline{E}\,\overline{G} - \overline{F}^2}\right\} = 1. \quad \square$$

The above lemma can be used to prove the following result.

Lemma 8.3.4. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let $\theta \in (0, 2\pi)$ be some fixed number. Then for each $R \in (0, \delta_p)$ there is some number $C_{R,\theta} \in (0, R)$ such that*

$$\sqrt{G(\binom{R}{\theta})} = R - \frac{R^3}{3!}K(p) - \frac{1}{4!}\frac{\partial(K\circ D_p)}{\partial R}(\binom{C_{R,\theta}}{\theta})\,R^4.$$

Proof. Exercise 8.3.2. $\square$

We need one more lemma making use of Proposition 8.3.2.

Lemma 8.3.5. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let δ_p and W be as in Proposition 8.2.3. Suppose that $x, y \in W$ are two points contained in the image of a geodesic polar coordinate patch D_p; let d_x and d_y*

denote the R-coordinates of x and y respectively. If C is a regular arc contained in the image of D_p and with endpoints x and y, then Length(C) $\geq |d_x - d_y|$.

Proof. By the definition of a regular arc (in Section 7.2) there is a regular curve $c\colon (a, b) \to M$ such that $C = c([u, v])$ for some closed interval $[u, v] \subset (a, b)$. We may assume without loss of generality that the image of c is contained in the image of D_p. Also, assume that $c(u) = x$ and $c(v) = y$. We can define coordinate functions $R, \theta\colon (a, b) \to \mathbb{R}$ for the curve c with respect to the coordinate patch D_x, that is $c(t) = D_x\left(\begin{pmatrix} R(t) \\ \theta(t) \end{pmatrix}\right)$ for all $t \in (a, b)$. Clearly $R(u) = d_x$ and $R(v) = d_y$. Using Lemma 5.8.1 and Proposition 8.3.2, we compute

$$\begin{aligned} \text{Length}(C) &= \int_u^v \sqrt{(R'(t))^2 + (\theta'(t))^2\, G(\tilde{c}(t))}\; dt \geq \int_u^v \sqrt{(R'(t))^2}\; dt \\ &\geq \int_u^v R'(t)\; dt = R(v) - R(u) = d_x - d_y. \end{aligned}$$

By integrating in the other direction, it is similarly seen that Length$(C) \geq d_y - d_x$. □

As an application of geodesic polar coordinates we show that our definitions of simplicial curvature (for simplicial surfaces) and Gaussian curvature (for smooth surfaces) are more closely analogous than they first appear. Simplicial curvature was defined using the angle defect; here we define an angle defect in the smooth case and show that it equals Gaussian curvature. To see where angles come into play in smooth surfaces (in which there are no natural triangles to make use of), recall from planar geometry that in a circle of radius R, the length S of an arc subtended by a central angle of ϕ radians is given by the formula $S = R\phi$. See Figure 8.3.2. Conversely, we can measure central angles in circles by $\phi = \frac{S}{R}$.

Now consider a point p on a smooth surface $M \subset \mathbb{R}^3$. For all small enough numbers of $r > 0$ there is a geodesic circle $\alpha_r(t)$ of radius r. Let L_r denote the length of this geodesic circle. It would be tempting to define the total angle around the point p to be $\frac{L_r}{r}$, and the angle defect at p to be $2\pi - \frac{L_r}{r}$. Unfortunately, in arbitrary surfaces the number $\frac{L_r}{r}$ depends upon the choice of r, so we take the limit as $r \to 0$, though we first need the following modification to make the limit work. Recall Exercise 3.7.4, where in order to form a better analog in the simplicial case of the smooth Gauss–Bonnet Theorem we modified

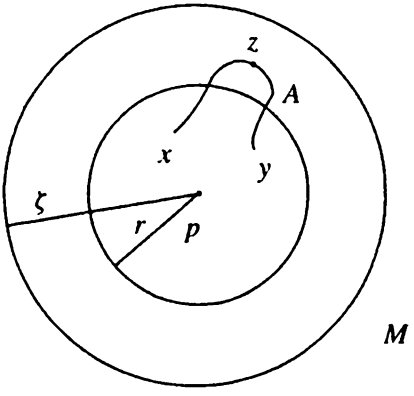


Figure 8.3.2

the angle defect for simplicial surfaces by dividing it by one third of the area of the star of the vertex prior; we make the analogous modification here. There is no "star" in the smooth case, but we take as our analog the region bounded by the geodesic circle we are using. We will use the formula $\pi r^2/3$ for one third of this area. (This formula is not quite accurate, since the surface need not be planar, but it is close to the correct value for small enough r, and it works in the limit.) The following proposition, due to Bertrand and Puiseux, shows that with this modification everything works out as desired.

Proposition 8.3.6. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let L_r be as above. Then*

$$lim_{r\to 0} \frac{2\pi - \frac{L_r}{r}}{\frac{\pi r^2}{3}} = K(p).$$

Proof. Let α_r be as in the proof of Proposition 8.3.2; recall that $\alpha_r'(t) = (D_p)_2(\binom{r}{t})$. Hence $\|\alpha_r'(t)\| = \sqrt{G(\binom{r}{t})}$. Using the formula for arc-length given in Equation 4.3.1, we compute

$$L_r = \lim_{\epsilon\to 0} \int_{\epsilon}^{2\pi-\epsilon} \|\alpha_r'(t)\|\, dt. = \lim_{\epsilon\to 0} \int_{\epsilon}^{2\pi-\epsilon} \sqrt{G(\binom{r}{t})}\, dt$$

where the integral is from ϵ to $2\pi - \epsilon$ since $\alpha_r(t)$ is only defined on $(0, 2\pi)$,

$$= \lim_{\epsilon \to 0} \int_{\epsilon}^{2\pi - \epsilon} \left\{ r - \frac{r^3}{3!} K(p) - \frac{1}{4!} \frac{\partial (K \circ D_p)}{\partial R} \left(\begin{pmatrix} C_{r,t} \\ t \end{pmatrix} \right) r^4 \right\} dt$$

by Lemma 8.3.4, where each $C_{r,t}$ is a number in $(0, r)$,

$$\begin{aligned} &= \lim_{\epsilon \to 0} \left\{ r - \frac{r^3}{3!} K(p) \right\} [(2\pi - \epsilon) - \epsilon] \\ &\quad - \lim_{\epsilon \to 0} \int_{\epsilon}^{2\pi - \epsilon} \frac{1}{4!} \frac{\partial (K \circ D_p)}{\partial R} \left(\begin{pmatrix} C_{r,t} \\ t \end{pmatrix} \right) r^4 \, dt \\ &= 2\pi \left\{ r - \frac{r^3}{3!} K(p) \right\} - \mathcal{E}(r), \end{aligned} \tag{8.3.7}$$

where

$$\mathcal{E}(r) = \lim_{\epsilon \to 0} \int_{\epsilon}^{2\pi - \epsilon} \frac{1}{4!} \frac{\partial (K \circ D_p)}{\partial R} \left(\begin{pmatrix} C_{r,t} \\ t \end{pmatrix} \right) r^4 \, dt.$$

Solving for $K(p)$ in Equation 8.3.7 and rearranging a bit, we obtain

$$K(p) = \frac{2\pi - \frac{L_r}{r}}{\frac{\pi r^2}{3}} - \frac{3}{\pi} \frac{\mathcal{E}(r)}{r^3}. \tag{8.3.8}$$

We now show that

$$\lim_{r \to 0} \frac{\mathcal{E}(r)}{r^3} = 0; \tag{8.3.9}$$

the conclusion of the theorem will then follow easily, observing that $K(p)$ does not depend upon r. Consider the function $\frac{\partial (K \circ D_p)}{\partial R}$; we want to find a maximal absolute value for this function for small values of R. Unfortunately, the function as given is defined on $(0, \delta_p) \times (0, 2\pi)$, a non-compact set. Note, however, that the definition of D_p implies that

$$\frac{\partial (K \circ D_p)}{\partial R} = \nabla (K \circ \operatorname{Exp}_p) \left(\begin{pmatrix} R\cos\theta \\ R\sin\theta \end{pmatrix} \right) \cdot \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix},$$

where ∇ in this case denotes the gradient and $\cdot$ denotes matrix multiplication. The right hand side of this equation is defined and continuous for all $\begin{pmatrix} R \\ \theta \end{pmatrix} \in [0, \delta'] \times [0, 2\pi]$, where δ' is any number such that $0 < \delta' < \delta_p$. Observing that $[0, \delta'] \times [0, 2\pi]$ is compact, it follows from Proposition 1.6.12 that $\nabla (K \circ$

$\mathrm{Exp}_p)(\begin{pmatrix} R\cos\theta \\ R\sin\theta \end{pmatrix}) \cdot \begin{pmatrix} \cos\theta \\ \sin\theta \end{pmatrix}$ has an absolute maximum and an absolute minimum on $[0, \delta'] \times [0, 2\pi]$. Hence there is some number D such that $|\frac{\partial(K \circ D_p)}{\partial R}| \leq D$ for all $\begin{pmatrix} R \\ \theta \end{pmatrix} \in (0, \delta') \times (0, 2\pi)$. We then compute

$$\begin{aligned}
\lim_{r\to 0} |\frac{\mathcal{E}(r)}{r^3}| &= \lim_{r\to 0} \frac{1}{r^3} \lim_{\epsilon\to 0} \left| \int_\epsilon^{2\pi-\epsilon} \frac{1}{4!} \frac{\partial(K \circ D_p)}{\partial R}(\begin{pmatrix} C_{r,t} \\ t \end{pmatrix}) r^4 \, dt \right| \\
&\leq \lim_{r\to 0} \frac{1}{r^3} \lim_{\epsilon\to 0} \int_\epsilon^{2\pi-\epsilon} \frac{1}{4!} \left| \frac{\partial(K \circ D_p)}{\partial R}(\begin{pmatrix} C_{r,t} \\ t \end{pmatrix}) \right| r^4 \, dt \\
&\leq \lim_{r\to 0} \frac{r}{4!} \lim_{\epsilon\to 0} \int_\epsilon^{2\pi-\epsilon} D \, dt = \lim_{r\to 0} \frac{r}{4!} 2\pi D = 0. \qquad \square
\end{aligned}$$

Exercises

8.3.1*. Prove the third and fourth equalities of Lemma 8.3.3.

8.3.2*. Prove Lemma 8.3.4.

8.3.3*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Show that p is contained in the image of D_q for some $q \in M$.

8.3.4*. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Suppose that c: $(a, b) \to M$ is a geodesic, the image of which is contained in the image of D_p. Show that the curve $(D_p)^{-1} \circ c$ has nowhere zero speed, and that its image is either contained in a line of the form $\theta = k$ for some constant k or it has no horizontal tangent vectors.

8.3.5*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point, and let δ_p be as in Proposition 8.2.3. Show that for any number δ' such that $0 < \delta' < \delta_p$, there is a number D such that

$$|\frac{\partial\sqrt{G}}{\partial R}| \leq D$$

for all $\begin{pmatrix} R \\ \theta \end{pmatrix} \in (0, \delta') \times (0, 2\pi)$.

8.3.6. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point and let A_r be the area bounded by the geodesic circle of radius r centered at p for all sufficiently small $r > 0$. Show that

$$\lim_{r \to 0} \frac{\pi - \frac{A_t}{r^2}}{\frac{\pi r^2}{12}} = K(p).$$

8.3.7*. Prove Lemma 8.3.1.

8.4 Proof of the Gauss–Bonnet Theorem

To prove the Gauss–Bonnet Theorem (Theorem 8.1.1) we need to compute the Euler characteristic of a compact smooth surface. A smooth surface is by definition a topological surface, so for any compact smooth surface we can simply forget that it is smooth and proceed to compute the Euler characteristic as for topological surfaces, namely by using triangulations. An arbitrary triangulation of a smooth surface might ignore the smooth nature of the smooth surface, however, and will turn out to be of no technical use to us. It turns out that compact smooth surfaces can always be triangulated in a particularly nice way that avoids this problem, as seen in Theorem 8.4.2 below.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface and let $x, y, z \in M$ be points. A subset $T \subset M$ is a **geodesic triangle** with vertices x, y and z, denoted Δxyz, if it is a disk, and if ∂T is the union of three geodesic arcs in M, denoted $\overline{xy}$, $\overline{xz}$ and $\overline{yz}$, where $\overline{xy}$ has endpoints x and y and similarly for the other two geodesic arcs; these geodesic arcs are called the **edges** of the geodesic triangle. We let $\angle x$, $\angle y$, $\angle z$ denote the angles between the edges of the geodesic triangle (that is, the angles between the tangent vectors to the edges at their points of intersection). See Figure 8.4.1. A triangulation $t: |K| \to M$ of M (for some simplicial surface K) is a **geodesic triangulation** if $t(\sigma)$ is a geodesic triangle in M for each 2-simplex σ of K. $\Diamond$

A geodesic triangle is a disk, and the union of the three edges of the geodesic triangle forms the boundary of the disk.

Example 8.4.1. A geodesic triangulation of S^2 is obtained by placing a small regular tetrahedron inside S^2 so that the center of mass of the tetrahedron is at the origin, and then projecting outward radially from the tetrahedron onto S^2. It is not hard to see that the images of the 1-simplices of the tetrahedron are parts of great circles on S^2, and hence are geodesic arcs. See Figure 3.4.2. $\Diamond$

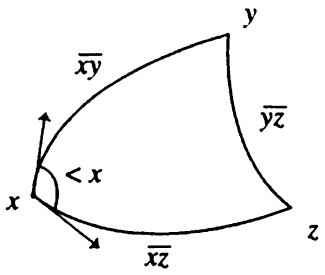


Figure 8.4.1

Do all compact smooth surfaces have geodesic triangulations, and if so are these triangulations unique (up to subdivision)? The following theorem answers the first part of this question.

Theorem 8.4.2. *Let $M \subset \mathbb{R}^3$ be a compact smooth surface. For every number $\epsilon > 0$ there is a geodesic triangulation $t: |K| \to M$ for some simplicial surface K such that for each 2-simplex σ of K the geodesic triangle $t(\sigma)$ is contained in an open ball of the form $O_\epsilon(p, M)$ for some point $p \in M$.*

The proof of this theorem is rather lengthy, and is given in Appendix A8.2. The second part of the question is easy to settle. Since compact smooth surfaces are compact topological surfaces and geodesic triangulations are triangulations, it follows from Theorem 3.4.5 that if a smooth surface can be geodesically triangulated by two different simplicial complexes K_1 and K_2, then K_1 and K_2 have simplicially isomorphic subdivisions.

Now to the proof of the Gauss–Bonnet Theorem. We need to compute a certain integral over a surface. Rather than attempting to evaluate the whole integral at once it is much easier to break up the surface into geodesic triangles, evaluate the integral over each geodesic triangle, and then piece the results together. The following theorem, in which the Gaussian curvature is integrated over a single geodesic triangle, is due to Gauss. Though we are stating this theorem as a preliminary to the Gauss–Bonnet theorem for convenience, the proof of this theorem contains the essence (as well as the difficulties) of the matter. We use the notation of the previous section.

Theorem 8.4.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface. If Δxyz is a geodesic triangle in M contained in the set $\exp_x(O_{\delta_x}(O_3, T_xM))$, then*

$$\int_{\Delta xyz} K dA = \angle x + \angle y + \angle z - \pi.$$

Proof. We follow (with minor modification) the proof in [SK3 vol. II], which is a modernized version of Gauss' original proof. The initial setup may appear somewhat unmotivated until the end of the proof. In the construction of the exponential coordinate patch Exp_x from the map $\exp_x$, an arbitrary orthogonal linear map $\Upsilon_x: \mathbb{R}^2 \to T_xM$ is chosen. In the present case we choose the map Υ_x so that the image of the positive x-axis under Υ_x contains line segment $(\exp_x)^{-1}(\overline{xy})$, which is possible since $\Delta xyz \subset \exp_x(O_{\delta_x}(O_3, T_xM))$ by hypothesis and $\overline{xy}$ is a geodesic. We now define a geodesic polar coordinate patch $D_x: (0, \delta_x) \times (-\lambda, 2\pi - \lambda) \to M$, where $\lambda > 0$ is some small enough number so that $\Delta xyz - \{x\}$ is contained in the image of D_x. Observe that using the interval $(-\lambda, 2\pi - \lambda)$ instead of $(0, 2\pi)$ changes none of the properties of D_x that were discussed in the previous section. Let $\hat{y} = D_x^{-1}(y)$ and $\hat{z} = D_x^{-1}(z)$. Observe that $\hat{y}$ has R-θ coordinates $\binom{y_1}{0}$, where $y_1 = \mathrm{Length}(\overline{xy})$, and $\hat{z}$ has R-θ coordinates $\binom{z_1}{\angle x}$, where $z_1 = \mathrm{Length}(\overline{xz})$. Also,

$$D_x^{-1}(\overline{xy}) = \{\begin{pmatrix} R \\ 0 \end{pmatrix} \mid 0 < R \leq y_1\} \quad \text{and} \quad D_x^{-1}(\overline{xz}) = \{\begin{pmatrix} R \\ \angle x \end{pmatrix} \mid 0 < R \leq z_1\}.$$

See Figure 8.4.2.

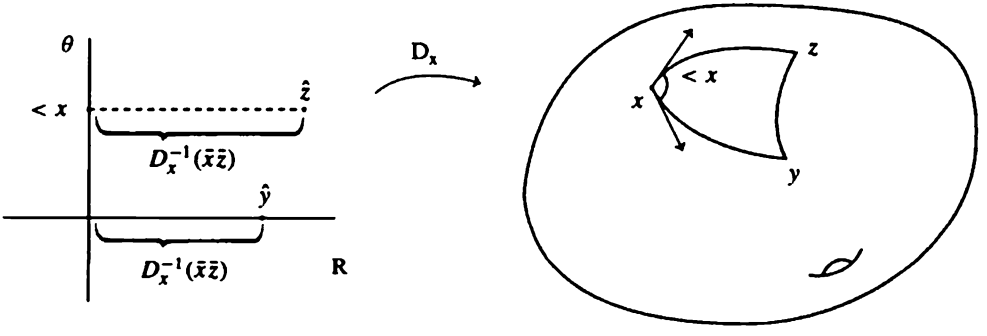


Figure 8.4.2

Since $\overline{yz}$ is a geodesic arc, there is by definition a geodesic $c: (p, q) \to M$ such that $\overline{yz} = c([m, n])$ for some closed interval $[m, n] \subset (p, q)$. By

possibly choosing new values for p and q (making them very close to m and n, respectively) we may assume that the image of c is contained in the image of D_x. It follows from Lemma 7.2.3 that $c(m)$ and $c(n)$ are the endpoints of $\overline{yz}$; without loss of generality we may assume that $c(m) = y$ and $c(n) = z$. We will let $\tilde{c}$ denote the curve $\tilde{c} = (D_x)^{-1} \circ c\colon (p, q) \to \mathbb{R}^2$. We think of the axes of $\mathbb{R}^2$ as corresponding to the variables R and θ. Observe that since $y \neq z$ the image of the curve $\tilde{c}$ cannot be contained in a line of the form $\theta =$ constant. We now use Exercises 8.3.4 and 4.2.4 to deduce that the image of $\tilde{c}$ is the graph of a function of the form $R = f(\theta)$ for some smooth function $f\colon (u, v) \to \mathbb{R}$, where (u, v) is some open interval in $\mathbb{R}$.

Let $\beta\colon (u, v) \to M$ be defined by $\beta(\theta) = D_x(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix})$. Observe that β is injective and is a parametrization of the image of c. It can be verified that $\beta^{-1}(y) = 0$ and $\beta^{-1}(z) = \angle x$. Further, we see that

$$\beta'(\theta) = f'(\theta)(D_x)_1(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix}) + (D_x)_2(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix}).$$

Using Proposition 8.3.2, a simple computation shows that $\|\beta'(\theta)\| \neq 0$ for all θ. Thus β is a regular curve.

By Proposition 4.3.4 there exists a diffeomorphism $h\colon (u', v') \to (u, v)$ for some open interval (u', v') such that $\beta \circ h$ is unit speed. We write $\theta = h(s)$. Using Exercise 4.3.3 we may assume without loss of generality that $h'(s) > 0$ for all $s \in (u', v')$. Since $y \neq z$, it follows from Lemma 7.2.2 (i) that c has non-zero constant speed. From Lemma 7.2.3 we deduce that $c|[m, n]$ is injective. By Exercise 4.3.10 we can find a number $\epsilon > 0$ such that $c|(m - \epsilon, n + \epsilon)$ is a homeomorphism from $(m - \epsilon, n + \epsilon)$ to $c((m - \epsilon, n + \epsilon))$. To avoid cumbersome notation we will simply assume that p and q are $m - \epsilon$ and $n + \epsilon$ respectively. Using Lemma 7.2.2 (iii) we deduce that $\beta \circ h$ is a geodesic. Note that the coordinate functions of $\beta \circ h$ with respect to the coordinate patch D_x are $(\beta \circ h)_1 = f \circ h$ and $(\beta \circ h)_2 = h$. Proposition 8.3.2 and Exercise 7.2.3 together allow us to apply Equation 7.2.6 to $\beta \circ h$, which yields

$$(f \circ h)''(s) - \frac{1}{2}\frac{\partial G}{\partial R}(h'(s))^2 = 0, \tag{8.4.1}$$

where G is evaluated at $\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}$.

We now define an angle function $\phi\colon (u, v) \to \mathbb{R}$ along the curve β as follows. For each $\theta \in (u, v)$, define $\phi(\theta)$ to be the angle from the tangent vector $(D_x)_1(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix})$ to the tangent vector $\beta'(\theta)$; observe that both these vectors are

in $T_{\beta(\theta)}$. See Figure 8.4.3. It is seen that

$$\phi(0) = \pi - \angle y \quad \text{and} \quad \phi(\angle x) = \angle z. \tag{8.4.2}$$

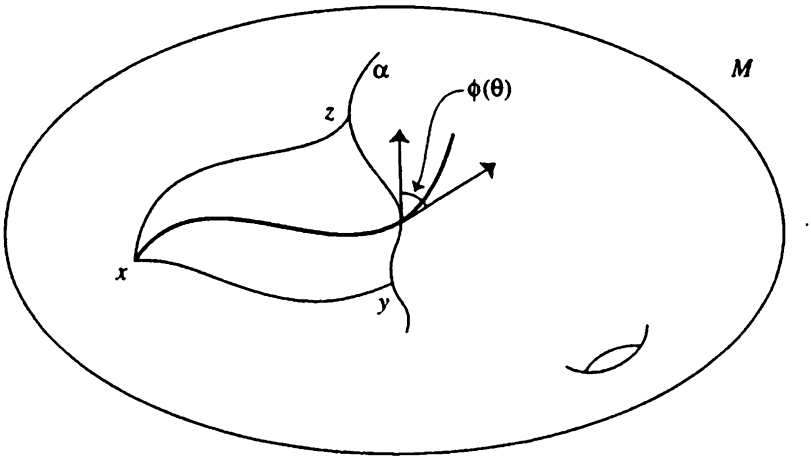


Figure 8.4.3

The function $\phi \circ h$ is the angle from the vector $(D_x)_1\left(\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}\right)$ to the vector $\beta'(h(s))$. This last vector is a positive multiple of the vector $(\beta \circ h)'(s)$ by the chain rule, so $\phi \circ h$ is also the angle from the vector $(D_x)_1\left(\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}\right)$ to the vector $(\beta \circ h)'(s)$; these two vectors are both unit vectors. Note that

$$(\beta \circ h)'(s) = (f \circ h)'(s)(D_x)_1 + h'(s)(D_x)_2,$$

where the partial derivatives of D_x are evaluated at $\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}$. Hence

$$\begin{aligned} \cos(\phi \circ h(s)) &= \langle (D_x)_1, (\beta \circ h)'(s) \rangle \\ &= \langle (D_x)_1, (f \circ h)'(s)(D_x)_1 + h'(s)(D_x)_2 \rangle = (f \circ h)'(s), \end{aligned} \tag{8.4.3}$$

where the last equality holds by Proposition 8.3.2. If we use the cross product instead of the inner product, we obtain

$$\begin{aligned}\sin(\phi \circ h(s)) &= \|(D_x)_1 \times (\beta \circ h)'(s)\| \\ &= \|(D_x)_1 \times \{(f \circ h)'(s)(D_x)_1 + h'(s)(D_x)_2\}\| \\ &= |h'(s)|\,\|(D_x)_1 \times (D_x)_2\| = |h'(s)| \sqrt{G\left(\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}\right)},\end{aligned} \tag{8.4.4}$$

where the last equality holds by Lemma 5.5.2 (i) and Equation 8.3.4. Since $h'(s) > 0$ for all $s \in (u', v')$, we can drop the absolute value in the last term in Equation 8.4.4. Combining Equations 8.4.1, 8.4.3, and 8.4.4 in that order, we obtain

$$\begin{aligned}\frac{1}{2}\frac{\partial G}{\partial R}(h'(s))^2 = (f \circ h)''(s) &= [\cos(\phi \circ h(s))]' = -\sin(\phi \circ h(s))\,\phi'(h(s))\,h'(s) \\ &= -h'(s)\sqrt{G\left(\begin{pmatrix} f(h(s)) \\ h(s) \end{pmatrix}\right)}\,\phi'(h(s))\,h'(s).\end{aligned}$$

Cancelling, isolating ϕ' (making use of Proposition 8.3.2), inserting the arguments in the derivative of G, and substituting $h(s) = \theta$, we deduce

$$\phi'(\theta) = -\frac{1}{2\sqrt{G\left(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix}\right)}}\frac{\partial G}{\partial R}\left(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix}\right) = -\frac{\partial \sqrt{G}}{\partial R}\left(\begin{pmatrix} f(\theta) \\ \theta \end{pmatrix}\right). \tag{8.4.5}$$

We have one more preliminary issue, which is that the point x is not in the image of the coordinate patch D_x (though every other point in Δxyz is contained in the image of D_x). For small enough $\mu \geq 0$, let $T_\mu \subset (0, \delta_x) \times (-\lambda, 2\pi - \lambda)$ be the region bounded by the lines $\theta = 0$, $\theta = \angle x$, and $R = \mu$ and by the curve $R = f(\theta)$. The set $D_x(T_\mu)$ is Δxyz with a geodesic disk of radius μ removed. See Figure 8.4.4. As μ goes to zero, the sets $D_x(T_\mu)$ converge to all of Δxyz.

Everything is now in place, and we compute

$$\begin{aligned}\int_{\Delta xyz} K\,dA &= \lim_{\mu \to 0} \int_{D_x(T_\mu)} K\,dA = \lim_{\mu \to 0} \iint_{T_\mu} K(D_x(\begin{pmatrix} R \\ \theta \end{pmatrix}))\sqrt{\det(g_{ij})}\,dR\,d\theta \\ &= \lim_{\mu \to 0} \iint_{T_\mu} -\frac{1}{\sqrt{G}}\frac{\partial^2 \sqrt{G}}{\partial R^2}\sqrt{G}\,dR\,d\theta\end{aligned}$$

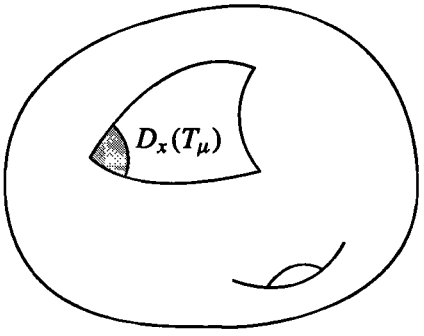


Figure 8.4.4

by Exercise 6.5.1 and Equation 8.3.4,

$$= -\lim_{\mu\to 0}\int_0^{\angle x}\int_\mu^{f(\theta)} \frac{\partial^2\sqrt{G}}{\partial R^2}\,dR\,d\theta$$

$$= -\lim_{\mu\to 0}\int_0^{\angle x}\left[\frac{\partial\sqrt{G}}{\partial R}\left(\begin{pmatrix} f(\theta)\\ \theta\end{pmatrix}\right) - \frac{\partial\sqrt{G}}{\partial R}\left(\begin{pmatrix}\mu\\ \theta\end{pmatrix}\right)\right]d\theta$$

$$= \int_0^{\angle x}\left[1 - \frac{\partial\sqrt{G}}{\partial R}\left(\begin{pmatrix} f(\theta)\\ \theta\end{pmatrix}\right)\right]d\theta,$$

by taking the limit and the minus sign inside the integral (to be justified below), and using Lemma 8.3.3,

$$= \int_0^{\angle x}\left[1 + \frac{d\phi}{d\theta}\right]d\theta$$

by Equation 8.4.5,

$$= \angle x + \phi(\angle x) - \phi(0) = \angle x + \angle z + \angle y - \pi,$$

by Equation 8.4.2. To justify taking the limit under the integral sign, we note that all the functions under consideration are continuous (hence integrable), and we appeal to the Bounded Convergence Theorem for Riemann integrals (see for example [BT, §22]), which can be used in our situation because of Exercise 8.3.5. □

The proof of the Gauss–Bonnet Theorem is now easy.

Proof of Theorem 8.1.1. Let δ_M and ϵ_M be as in Corollaries 8.2.5 and 8.2.6. By Theorem 8.4.2 we can find a geodesic triangulation $t: |K| \to M$ such that for each 2-simplex σ of K the geodesic triangle $t(\sigma)$ is contained in an open ball of the form $O_{\epsilon_M}(p, M)$ for some point $p \in M$. It follows from Corollary 8.2.6 that each geodesic triangle $t(\sigma)$ satisfies the hypothesis of Theorem 8.4.3 with respect to any of the vertices of $t(\sigma)$, and with δ_M replacing δ_τ (a change that has no effect on the outcome of Theorem 8.4.3). If $\sigma \in K$ is a 2-simplex and v is a vertex of σ, let $\angle(t(v), t(\sigma))$ denote the angle in $t(\sigma)$ at $t(v)$. We then use the conclusion of Theorem 8.4.3 to compute

$$\begin{aligned}
\int_M K dA &= \sum_{\sigma \in K^{((2))}} \int_{t(\sigma)} K dA \\
&= \sum_{\sigma \in K^{((2))}} \left[\sum_{v \in \sigma} \angle(t(v), t(\sigma)) - \pi \right] \\
&\qquad \text{where the inner summation is over the three vertices of } \sigma. \\
&= \sum_{\sigma \in K^{((2))}} \sum_{v \in \sigma} \angle(t(v), t(\sigma)) - \sum_{\sigma \in K^{((2))}} \pi \\
&= \sum_{v \in K^{((0))}} \sum_{\sigma v} \angle(t(v), t(\sigma)) - \pi f_2(K) \\
&= \sum_{v \in K^{((0))}} 2\pi - \pi f_2(K) \\
&\qquad \text{since the sum of all the angles around a vertex is } 2\pi, \\
&= 2\pi f_0(K) - \pi f_2(K) = \cdots = 2\pi \chi(K) = 2\pi \chi(M),
\end{aligned}$$

where the $= \cdots =$ uses the same argument as in the proof of Theorem 3.7.2, and the final equality is by the definition of the Euler characteristic of a topological surface. □

A typical application of the Gauss–Bonnet Theorem is the following result.

Proposition 8.4.4. *Let $T \subset \mathbb{R}^3$ be a smooth surface homeomorphic to T^2. Then the Gaussian curvature of T is positive on a non-empty open subset of T and negative on a non-empty open subset of T.*

Proof. Recall that $\chi(T^2) = 0$, and hence $\chi(T) = 0$ by Exercise 3.5.3. Since T^2 is compact, so is T. By the Gauss–Bonnet Theorem (Theorem 8.1.1), we know

$$\int_T K dA = 0.$$

Using the compactness of T and Exercise 6.4.11, it follows that K is positive on some non-empty open subset of T. Hence K must be negative somewhere on T, and the continuity of K implies that K must be negative on some non-empty open subset of T. □

Exercises

8.4.1. Find a formula for the area of a geodesic triangle contained in a hemisphere on a sphere of radius R in terms of the angles of the geodesic triangle.

8.4.2. Let $T_n \subset \mathbb{R}^3$ be a smooth surface homeomorphic to $T^2 \# \cdots \# T^2$, where there are n summands. Show that if $n > 1$ then the Gaussian curvature of T_n must be negative on a non-empty open subset of T_n.

8.5 Non-Euclidean Geometry

This last section in the book, which uses the machinery we have built up for smooth surfaces, takes us back to the Greek origins of the rigorous study of geometry, showing how to connect the modern approach to surfaces with Euclidean and non-Euclidean geometry. We will by necessity be somewhat sketchy; see the various references given below for more details.

There are two fundamental approaches to Euclidean geometry: synthetic and analytic. In the former approach we start with a few axioms concerning undefined objects such as points and lines, and logically deduce various results from these axioms. In the analytic approach the Euclidean plane $\mathbb{R}^2$ is viewed as the set of ordered pairs of real numbers, and lines are defined to be solutions to equations of the form $y = mx + b$ or $x = a$; we then use these equations to prove various geometrical results — the same results we would obtain by the synthetic method, though proved quite differently. Even in the analytic method we are making use of axioms, in this case the properties of the real numbers and set theory, but the assumptions are pushed back out of the realm of geometry. (A reference that incorporates both approaches is [CE]).

Our goal here is to use differential geometry to attempt to give an analytic approach to classical non-Euclidean geometry, which was first given synthetically, just as the Cartesian plane $\mathbb{R}^2$ is an analytic model for Euclidean geometry. As we will see later, we cannot completely fulfill this goal using the tools in this book, but we can get fairly close. For details on non-Euclidean geometry, see [GE], [MCL], [CE] and [TU].

We start with a very brief discussion of the synthetic approach to non-Euclidean geometry, which in turn needs a review of Euclidean geometry. Euclid's system of geometry, found in [EU], was for many years taken almost as gospel; his logic was viewed as a model of deriving absolute knowledge, and his postulates were viewed as a necessarily true description of our physical world. As late as the 18th Century, the philosopher Immanuel Kant said that "the concept of [Euclidean] space is by no means of empirical origin, but is an inevitable necessity of thought"; see [GE] for more discussion. This blind faith in Euclid started to unravel in the early 19th Century with the discovery of non-Euclidean geometry by Gauss, Bolyai and Lobachevsky, and even more so with the work of Riemann, which led, among other things, to the separation of mathematical space from physical space. See [GE] and [MCL] for a discussion of the development of non-Euclidean geometry, and see [SK3, vol. II] and [MCL] for a discussion of Riemann's work. We can now make use of much work in the 19th and 20th Centuries, and see that there are three categories of flaws in Euclid's work (though such criticisms are in no way intended to deny the overwhelming importance of that work).

To examine Euclid's work, we start with the first four of his definitions and his five postulates from Book 1 of the Elements, as stated in [EU].

Definitions:

1. A point is that which has no part.
2. A line is breadthless length. A straight line is a line which lies evenly within itself.
3. The extremities of a line are points.
4. A straight line is a line which lies evenly with the points on itself.

Postulates:

1. To draw a straight line from any point to any point.
2. To produce a finite straight line continuously in a straight line.
3. To describe a circle with any centre and distance.
4. That all right angles are equal to one another.

5. That, if a straight line falling on two straight lines makes the interior angles on the same side less than two right angles, the two straight lines, if produced indefinitely, meet on that side on which are the angles less than the two right angles.

The first problem in Euclid concerns his definitions. Consider, for example, Definition 2 given above. Though this definition does seem to correspond to our intuition concerning straight lines, by modern standards this definition is utterly useless. What is a "length," breadthless or not, and what is "breadth"? What does it mean for something to "lie evenly with itself"? The problem with Euclid's approach is that he tried to define every concept that he was using. What Euclid appeared to miss is that, just as it is necessary to start with some unproved axioms, it is also necessary to start with some undefined concepts upon which all our subsequent definitions will be based. What makes these undefined objects behave as our intuition tells us they should are the various properties of the objects mandated by our axioms. It is the axiomatic properties of the objects, not the objects themselves, that count.

Another problem with Euclid's method is that his axioms do not quite suffice for rigorous proofs of the theorems Euclid states. A simple example is that Euclid never hypothesizes the uniqueness of a straight containing two distinct points, though this property is needed. A more subtle issue concerns Euclid's lack of attention to the issue of certain points on a line being between other points; it may seem obvious from the diagrams used in classical geometry that certain points are between others, but axioms are needed to make this notion rigorous. see [GE, Chapter 3] for a discussion of this matter. These problems can all be remedied, however, by more complete sets of axioms. Two such axiom systems were given by Hilbert and Birkhoff (see [CE]).

Whereas Euclid's problem with definitions and missing axioms can be remedied by using better axiom schemes, the third problem with Euclid, and, in fact, the one that led to the recognition of the other two problems, is not so simple. The issue here is not to find a more complete set of axioms that do rigorously that which Euclid was trying to do, but rather to ask more fundamentally of what planar geometry consists. More specifically, the problem concerns Euclid's fifth postulate. Ever since Euclid, observers have been suspicious of the fifth postulate on the grounds that it does not have the simple and obvious nature of the first four postulates; many mathematicians have attempted to deduce the fifth postulate from its predecessors, thus making it unnecessary. It turns out

that the fifth postulate cannot, in fact, be derived from its predecessors, and that it is possible to have a geometry based on the first four postulates but having an alternative to the fifth postulate. To appreciate these alternatives, we note that it can be shown that the fifth postulate is equivalent to the following statements:

Playfair's Axiom. Let l be a line on the plane, and let P be a point in the plane not contained in l. Then there is one and only one line containing P and parallel to l.

Euclidean Angle-Sum Axiom. The sum of the angles in a triangle in the plane is 180°.

The former of these two axioms is often mistakenly thought to be Euclid's original fifth postulate.

We define hyperbolic geometry to be the geometry derived from Euclid's first four postulates, together with the following postulate.

Hyperbolic Axiom. Let l be a line on the plane, and let P be a point in the plane not contained in l. Then there is more than one line containing P and parallel to l.

Similarly, we define elliptic geometry to be the geometry derived from Euclid's first four postulates, together with the following postulate.

Elliptic Axiom. Let l be a line on the plane, and let P be a point in the plane not contained in l. Then there are no lines containing P and parallel to l.

It can be shown that the sum of the angles in a triangle is less than 180° in hyperbolic geometry and is greater than 180° in elliptic geometry.

Euclidean and non-Euclidean geometry are both two-dimensional geometries. The former can be modeled using the plane $\mathbb{R}^2$, by which we mean that the points and straight lines in $\mathbb{R}^2$ (given by equations as above) satisfy all of Euclid's postulates. Since straight lines in $\mathbb{R}^2$ are geodesics, it would be reasonable to try to model non-Euclidean geometry as a smooth surface in $\mathbb{R}^3$, with geodesics playing the role of lines, which must satisfy appropriate analogs of the first four of Euclid's postulates, together with the analog of the appropriate fifth axiom. A very thorough treatment of the analogs for smooth surfaces of Euclid's postulates can be found in [MCL]; we proceed informally here.

We start with the useful observation about Euclidean space that any point looks like any other point from the point of view of geometric constructions. For

example, a circle of radius 1 centered about any one point is congruent to a circle of the same radius centered about any other point. We want such a property to hold for any surface that models classical geometry. The following definition is a somewhat weaker version of this notion, stated in the more general setting of smooth surfaces.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. We say M is **locally homogeneous** if for any two points $p, q \in M$ there is an isometry from an open neighborhood of p in M to an open neighborhood of q in M. $\Diamond$

Using the Theorema Egregium (Theorem 6.5.3), we see that if a smooth surface has two points p and q with different curvature (as always Gaussian), then the surface could not be locally homogeneous. Thus, if a surface is to model classical geometry, it must have constant curvature.

Suppose $M \subset \mathbb{R}^3$ is a smooth surface with constant curvature K_0. If Δxyz is a small enough geodesic triangle in M, so that it satisfies the hypothesis of Theorem 8.4.3, then

$$\text{Area}(\Delta xyz)\, K_0 = \int_{\Delta xyz} K_0 dA = \angle x + \angle y + \angle z - \pi.$$

If $K_0 = 0$, then it follows that $\angle x + \angle y + \angle z = \pi$, as is the case in Euclidean geometry; if $K_0 > 0$, then $\angle x + \angle y + \angle z > \pi$, as is the case in elliptic geometry; if $K_0 < 0$, then $\angle x + \angle y + \angle z < \pi$, as is the case in hyperbolic geometry.

Another familiar property of the Euclidean plane is that lines can be extended indefinitely. The corresponding notion for smooth surfaces is given as property (3) in the following theorem, which is a version of the well-known Hopf–Rinow Theorem (see [KL] or [MCL] for proofs). We mentioned in Section 7.2 that there is a relation between extending geodesics and the existence of "holes" in the surface. The precise notion of not having holes is known as *completeness*. A precise definition is outside the scope of this book, and can be found in [H-Y, §2-13] or [JA, Chapter IV]; intuitively the idea is that if a sequence of points in the surface get closer and closer to one another, then the sequence in fact converges to something in the surface. This definition depends upon the metric used on the surface, and not just the topology of the surface; we use the standard metric on $\mathbb{R}^3$, which the surface inherits, though an intrinsic metric formed using lengths of curves could also be used.

Theorem 8.5.1 (Hopf–Rinow Theorem). *Let $M \subset \mathbb{R}^3$ be a smooth surface.*

(i) The following are equivalent:

(1) *The surface M is topologically complete.*
(2) *Let $p \in M$ be a point, and let $v \in T_pM$ be a vector. Then there is a geodesic $c\colon (-\infty, \infty) \to M$ such that $c(0) = p$ and $c'(0) = v$.*
(3) *Let $c\colon (a, b) \to M$ be a geodesic. Then there is a geodesic $\hat{c}\colon \mathbb{R} \to M$ such that $\hat{c}|(a, b) = c$.*
(4) *Let $p \in M$ be a point. The exponential map $\exp_p$ is defined on all of T_pM.*

(ii) Let $p, q \in M$ be distinct points. If any of the above conditions hold, then there is a geodesic arc in M with endpoints p and q; this geodesic arc has the shortest length of any curve joining p and q.

We can now make the following definition.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. We say M is **geodesically complete** if any of the four properties in part (i) of the above theorem holds. ◊

Putting all our observations together, we see that for a smooth surface to model classical geometry it must at minimum be geodesically complete and locally homogeneous, have constant curvature, and must satisfy the appropriate analogs of Euclid's first four postulates. Then, depending upon whether the constant curvature is zero, positive or negative, we will obtain a model for Euclidean, elliptic or hyperbolic geometry, respectively. It turns out not to be too hard to find surfaces; in fact, any geodesically complete smooth surface of constant curvature is locally homogeneous and satisfies the appropriate analogs of the first four postulates. That such a surface is locally homogeneous is seen by the following theorem, due to Minding. A geodesically complete surface satisfies the analogs of the first two postulates of Euclid by parts (ii) and (i)(3) of the Hopf–Rinow Theorem (Theorem 8.5.1). The analog of the third postulate holds using part (i)(4) of the Hopf–Rinow Theorem, which implies that geodesic circles of any radius centered at any point can always be defined (see Section 8.2 for the definition of geodesic circles). The analog of the fourth postulate, which requires the existence of isometries of the surface taking one right angle to another, can be deduced, with one caveat, from the proof of the following theorem, by choosing the appropriate geodesic coordinate patch for each surface; the caveat is that we only obtain a local isometry from the theorem, rather than a global isometry, but that is all we can do with our tools.

Theorem 8.5.2. *Let $M, N \subset \mathbb{R}^3$ be smooth surfaces with constant curvatures K_M and K_N respectively. Let $p \in M$ and $q \in N$ be points. Then there is an isometry from an open neighborhood of p in M to an open neighborhood of q in N iff $K_M = K_N$.*

Proof. We follow [MCL]. If there is an isometry from an open neighborhood of p in M to an open neighborhood of q in N, then it follows from the Theorema Egregium (Theorem 6.5.3) that $K_M = K_N$. Now suppose conversely that $K_M = K_N$. Let δ_p and δ_q be as in Proposition 8.2.3, and let $\eta = \min\{\delta_p, \delta_q\}$. We can then define geodesic polar coordinate patches $D_p\colon (0, \eta) \times (0, 2\pi) \to M$ and $D_q\colon (0, \eta) \times (0, 2\pi) \to N$. Let U be the image of D_p and let V be the image of D_q. It follows from Proposition 8.2.3 that U is an open subset of M and V is an open subset of N.

Let Exp_p and Exp_q be exponential coordinate patches (defined in Section 8.2) for p and q respectively, though take $O_\eta(O_2, \mathbb{R}^2)$ as the domains of both functions. We now define $f\colon U \to V$ to be $f = \mathrm{Exp}_q \circ (\mathrm{Exp}_p)^{-1}$. It follows straightforwardly that $f \circ D_p = D_q$. If we can show that D_p and D_q have the same metric coefficients, then it will follow from Proposition 5.9.4 (4) that f is a local isometry, which suffices to imply what we are trying to prove.

Recall from Proposition 8.3.2 the simple form of the metric coefficients of D_p and D_q; let G_p and G_q denote the appropriate metric coefficients, using the notation of that proposition. It will suffice to prove that $G_p = G_q$. Using Exercise 6.5.1 we see that

$$\begin{aligned} K_M \circ D_p &= -\frac{1}{\sqrt{G_p}} \frac{\partial^2 \sqrt{G_p}}{\partial R^2} \\ K_N \circ D_q &= -\frac{1}{\sqrt{G_q}} \frac{\partial^2 \sqrt{G_q}}{\partial R^2}, \end{aligned} \tag{8.5.1}$$

where G_p and G_q are the appropriate metric coefficients of D_p and D_q respectively. The functions $K_M \circ D_p$ and $K_N \circ D_q$ are the same constant function; let us denote this function $(0, \eta) \times (0, 2\pi) \to \mathbb{R}$ by Z. Let $\theta \in (0, 2\pi)$ be fixed. Then we can think of $\sqrt{G_p}$, $\sqrt{G_q}$ and Z as functions of R only. By Equation 8.5.1 we then see that $\sqrt{G_p}$ and $\sqrt{G_q}$ both satisfy the ordinary differential equation

$$\frac{d^2 x}{d R^2} + Zx = 0.$$

By Lemma 8.3.3 we see that $\sqrt{G_p}$ and $\sqrt{G_q}$ both satisfy the initial conditions

$$\lim_{R\to 0} x = 0 \quad \text{and} \quad \lim_{R\to 0} \frac{dx}{dR} = 1.$$

By the existence and uniqueness of solutions of ordinary differential equations (Theorem 4.2.4), we see that $\sqrt{G_p} = \sqrt{G_q}$ for all small enough values of R, with the given θ. Since the choice of θ was arbitrary, we deduce that $G_p = G_q$. □

For more details concerning the analogs of the first four postulates of Euclid for smooth surfaces, see [MCL].

To model Euclidean geometry we simply use $\mathbb{R}^2$, as mentioned. To model elliptic geometry, we simply use the unit sphere S^2 in $\mathbb{R}^3$. We saw in Example 6.3.1 (2) that S^2 has constant curvature 1. Since the geodesics on S^2 are great circles, which can be extended indefinitely, we see that S^2 is geodesically complete. It can be seen directly that the postulate for elliptic surfaces is satisfied; also, since the curvature of S^2 is constantly 1, then the sum of the angle in a geodesic triangle is always greater than 180°; for example, consider a triangle with one vertex at the North pole, and the other two vertices on the equator. We thus see that S^2 is a model for elliptic geometry. It is easy to see from Exercise 6.4.11 that any compact smooth surface in $\mathbb{R}^3$ with constant curvature has positive curvature; a harder proof shows the only compact, connected smooth surface in $\mathbb{R}^3$ with constant curvature is a sphere (see [MCL, p. 193]).

A good attempt at constructing a model for hyperbolic geometry is the surface of revolution parametrized by the curve

$$c(t) = \begin{pmatrix} \sin t \\ \ln \tan \frac{t}{2} + \cos t \end{pmatrix}$$

for $t \in (0, \pi/2)$. See Figure 8.5.1. The domain of c cannot be expanded to a larger interval; the curve c is not defined at $t = 0$, and at $t = \pi/2$ the curve is defined but is not regular. This curve is known as the tractrix, and it is the path taken by a weight at the end of a taut rope pulled by a person walking in a straight line (the weight is not initially on this straight line). The resulting surface of revolution is called the pseudosphere; see Figure 8.5.2. Using the formula for Gaussian curvature for surfaces of revolution given in Exercise 6.4.3, it is seen that the pseudosphere has constant curvature -1.

The one problem with the pseudosphere is that it is not geodesically complete. No geodesic can be extended above the x-y plane. Locally this surface does behave just as hyperbolic geometry should (for example, the sum of the

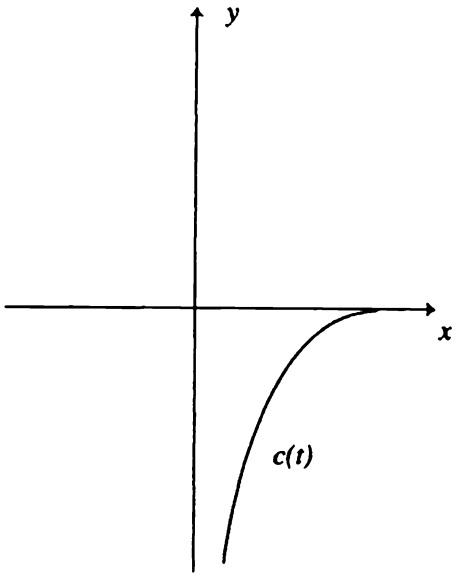


Figure 8.5.1

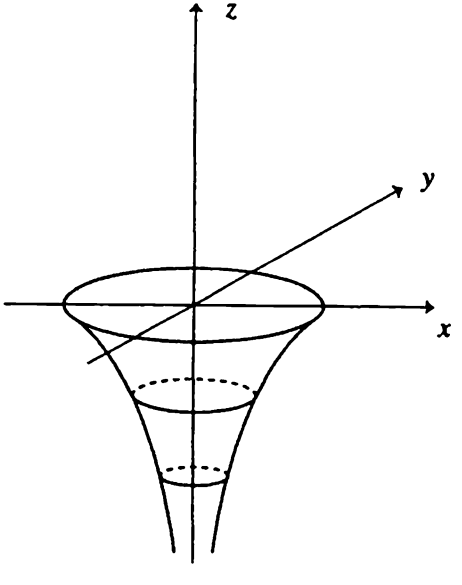


Figure 8.5.2

angles in a geodesic triangle contained in a disk in this surface will be less than 180°); globally, however, this is not a model for hyperbolic geometry.

The following result of Hilbert shows that any attempt to model hyperbolic geometry using a surface in $\mathbb{R}^3$ is doomed to fail.

Theorem 8.5.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface with constant negative curvature. Then M is not geodesically complete.*

Proofs of this theorem can be found in [SK3, vol. III Chapter 5] and [MCL, Chapter 14]; the former reference has a very thorough discussion of complete surfaces of constant curvature, whereas the latter reference has a more succinct treatment of Hilbert's theorem (beware that the statement of the theorem in this reference is missing the crucial phrase "geodesically complete," though the proof uses this concept). It is still possible to find a concrete model for hyperbolic geometry, though such a model necessitates the use of abstract surfaces rather than surfaces in $\mathbb{R}^3$. See [ST] or [MCL, Chapter 15].

Appendix A8.1 Geodesic Convexity

We start with the proof of Proposition 8.2.3.

Proof of Proposition 8.2.3. We follow [BO]. All four parts of the proposition will be proved together. Let $x: U \to M$ be a coordinate patch such that $p \in x(U)$, and as usual let $\bar{p} = \binom{p_1}{p_2} = x^{-1}(p) \in U$. A typical point in $U \times \mathbb{R}^2$ will be denoted $(\bar{q}, \bar{v})$, where $\bar{q} \in U$ and $\bar{v} \in \mathbb{R}^2$. We will identify $\mathbb{R}^2 \times \mathbb{R}^2$ with $\mathbb{R}^4$; if $\bar{q} = \binom{q_1}{q_2}$ and $\bar{v} = \binom{v_1}{v_2}$, we will identify $(\bar{q}, \bar{v}) \in U \times \mathbb{R}^2$ with $\begin{pmatrix} q_1 \\ q_2 \\ v_1 \\ v_2 \end{pmatrix} \in \mathbb{R}^4$.

We start out similarly to the proof of Theorem 7.2.6, using the system of ordinary differential equations given in Equation 7.2.5. However, rather than fixing the initial conditions and using Theorem 4.2.4, we use the more powerful Theorem 4.2.5. Let $p_0 \in U \times \mathbb{R}^2$ denote the point $p_0 = (\bar{p}, O_2) = \begin{pmatrix} p_1 \\ p_2 \\ 0 \\ 0 \end{pmatrix} \in \mathbb{R}^4$. Applying Theorem 4.2.5 to the system of differential equations in Equation 7.2.5 and the given point p_0, it follows that there is a number $\epsilon > 0$, an open subset $Y \subset U \times \mathbb{R}^2$ containing p_0 and a smooth function

$$C: (-\epsilon, \epsilon) \times Y \to U \times \mathbb{R}^2,$$

written

$$C(t, \bar{q}, \bar{v}) = \begin{pmatrix} c_1(t, \bar{q}, \bar{v}) \\ c_2(t, \bar{q}, \bar{v}) \\ d_1(t, \bar{q}, \bar{v}) \\ d_2(t, \bar{q}, \bar{v}) \end{pmatrix},$$

such that the functions $c_1(t, \bar{q}, \bar{v})$, $c_2(t, \bar{q}, \bar{v})$, $d_1(t, \bar{q}, \bar{v})$ and $d_2(t, \bar{q}, \bar{v})$ satisfy our system of differential equations, and

$$\begin{pmatrix} c_1(0, \bar{q}, \bar{v}) \\ c_2(0, \bar{q}, \bar{v}) \\ d_1(0, \bar{q}, \bar{v}) \\ d_2(0, \bar{q}, \bar{v}) \end{pmatrix} = (\bar{q}, \bar{v}) = \begin{pmatrix} q_1 \\ q_2 \\ v_1 \\ v_2 \end{pmatrix}$$

for all $(\bar{q}, \bar{v}) \in Y$. Hence, for each $(\bar{q}, \bar{v}) \in Y$, the curve $c_{\bar{q},\bar{v}}\colon (-\epsilon, \epsilon) \to M$ defined by

$$c_{\bar{q},\bar{v}}(t) = x(\begin{pmatrix} c_1(t, \bar{q}, \bar{v}) \\ c_2(t, \bar{q}, \bar{v}) \end{pmatrix})$$

is the unique geodesic in M with $c_{\bar{q},\bar{v}}(0) = x(\bar{q})$ and $c'_{\bar{q},\bar{v}}(0) = dx_{\bar{q}}(\bar{v}) \in T_{x(\bar{q})}M$, where the latter equality uses Exercise 5.9.7.

Applying Lemma 1.2.9 to the set Y, we see that there are numbers $\epsilon_1, \epsilon_2 > 0$ such that

$$O_{\epsilon_1}(\bar{p}, U) \times O_{\epsilon_2}(O_2, \mathbb{R}^2) \subset Y.$$

For each $\bar{q} \in O_{\epsilon_1}(\bar{p}, U)$ and $\bar{v} \in O_{\epsilon\epsilon_2/2}(O_2, \mathbb{R}^2)$ (note the change in the radius of the second ball) we can define the function $\widehat{c}_{\bar{q},\bar{v}}\colon (-2, 2) \to M$ by

$$\widehat{c}_{\bar{q},\bar{v}}(t) = c_{\bar{q},\frac{2\bar{v}}{\epsilon}}\left(\frac{\epsilon t}{2}\right) = x(\begin{pmatrix} c_1(\frac{\epsilon t}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon}) \\ c_2(\frac{\epsilon t}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon}) \end{pmatrix}).$$

Using Exercise 7.2.5 we know that $\widehat{c}_{\bar{q},\bar{v}}$ is a geodesic, and it is straightforward to verify that $\widehat{c}_{\bar{q},\bar{v}}(0) = x(\bar{q})$ and $\widehat{c}'_{\bar{q},\bar{v}}(0) = dx_{\bar{q}}(\bar{v})$. By the definition of the exponential map we have

$$\exp_{x(\bar{q})}(dx_{\bar{q}}(\bar{v})) = \widehat{c}_{\bar{q},\bar{v}}(1). \tag{A8.1.1}$$

We now define a number of useful maps. Let

$$\Phi\colon O_{\epsilon_1}(\bar{p}, U) \times O_{\epsilon\epsilon_2/2}(O_2, \mathbb{R}^2) \to U \times U$$

be defined by

$$\Phi(\bar{q}, \bar{v}) = (\bar{q}, x^{-1}(\widehat{c}_{\bar{q},\bar{v}}(1))).$$

It can be verified that Φ is a smooth map from an open subset of $\mathbb{R}^2 \times \mathbb{R}^2$ into $\mathbb{R}^4$. It can be further seen that the $\Phi(p_0) = \Phi((\bar{p}, O_2)) = (\bar{p}, \bar{p})$. Further, it is shown in Exercise A8.1.1 that the Jacobian matrix $D\Phi(p_0)$ is non-singular.

Next, let $Tx(U)$ and $\Psi: U \times \mathbb{R}^2 \to Tx(U) \subset \mathbb{R}^6$ be defined as in Exercise 5.9.5; note that Ψ is a homeomorphism. Let $\Delta: U \times U \to M \times M \subset \mathbb{R}^3 \times \mathbb{R}^3$ be defined by $\Delta(a, b) = (x(a), x(b))$. It can be verified that $\Delta(U \times U)$ is open in M, and that Δ is a homeomorphism from $U \times U$ onto $\Delta(U \times U)$. For each point $q \in x(U)$, define $I_q: T_qM \to \mathbb{R}^3 \times \mathbb{R}^3$ by $I_q(v) = (q, v)$, and we define $J_{\bar{q}}: \mathbb{R}^2 \to U \times \mathbb{R}^2$ by $J_{\bar{q}}(w) = (\bar{q}, w)$, where as usual $\bar{q} = x^{-1}(q)$. Finally, let $P_2: \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}^2$ and $P_3: \mathbb{R}^3 \times \mathbb{R}^3 \to \mathbb{R}^3$ both be projections onto the second factor.

The non-singularity of the Jacobian matrix $D\Phi(p_0)$ allows us to apply the Inverse Function Theorem (Theorem 4.2.1) to conclude that there is an open set

$$Z \subset O_{\epsilon_1}(\bar{p}, U) \times O_{\epsilon\epsilon_2/2}(O_2, \mathbb{R}^2) \subset U \times \mathbb{R}^2 \tag{A8.1.2}$$

containing p_0 such that $\Phi(Z)$ is open in $U \times U$ (and hence in $\mathbb{R}^4$) and Φ is a diffeomorphism from Z onto $\Phi(Z)$. From Exercise 5.9.5 we know that $\Psi(Z)$ is an open subset of $Tx(U) \subset \mathbb{R}^3 \times \mathbb{R}^3 = \mathbb{R}^6$. Note that $\Psi(p_0) = (p, O_3)$. By Lemma 1.2.5 there is an open subset $Q \subset \mathbb{R}^6$ such that $\Psi(Z) = Q \cap Tx(U)$. Applying Lemma 1.2.9 to the set Q we see that there are numbers $\epsilon_3, \delta_p > 0$ such that

$$O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3) \subset Q.$$

Hence

$$\left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U) \subset \Psi(Z). \tag{A8.1.3}$$

Since $\left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)$ is an open subset of $Tx(U)$, and since Ψ, $\Phi|Z$ and Δ are homeomorphisms, it follows that

$$\Delta \circ \Phi|Z \circ \Psi^{-1}\left(\left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)\right)$$

is an open subset of $M \times M$ that contains the point (p, p). The reader is asked to verify that an open set $W \subset M$ containing p can be found such that

$$W \times W \subset \Delta \circ \Phi | Z \circ \Psi^{-1}\left(\left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)\right). \quad \text{(A8.1.4)}$$

We observe that for $(q, v) \in \left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)$ we have $\Delta \circ \Phi | Z \circ \Psi^{-1}(q, v) = (q, w)$ for some appropriate w; the same property holds for the map $\Psi \circ (\Phi | Z)^{-1} \circ \Delta^{-1}$. It follows that

$$\Delta \circ \Phi | Z \circ \Psi^{-1}(\{q\} \times O_{\delta_p}(O_3, T_q M)) \supset \{q\} \times W \quad \text{(A8.1.5)}$$

for all $q \in W$.

We can now verify that the set W and number δ_p are what we are looking for. Let $q \in W$ be fixed, and let $\bar{q} = x^{-1}(q)$. To prove part (i) of the proposition, let $v \in O_{\delta_p}(O_3, T_q M)$ be a vector, and let $\bar{v} = (dx_{\bar{q}})^{-1}(v)$. Since $(q, q) \in W \times W$, it follows from Equation A8.1.4 that

$$\Psi \circ (\Phi | Z)^{-1} \circ \Delta^{-1}(q, q) \in \left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U).$$

Using the remark made right after Equation A8.1.4 we deduce that $(q, w) \in \left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)$ for some appropriate w; it then follows that $q \in O_{\epsilon_3}(p, M)$. Since $v \in O_{\delta_p}(O_3, T_q M)$ by hypothesis, it follows that $(q, v) \in \left\{O_{\epsilon_3}(p, \mathbb{R}^3) \times O_{\delta_p}(O_3, \mathbb{R}^3)\right\} \cap Tx(U)$. Using Equations A8.1.3 and A8.1.2 we deduce that

$$(\bar{q}, \bar{v}) = \Psi^{-1}(q, v) \in O_{\epsilon_1}(\bar{p}, U) \times O_{\epsilon\epsilon_2/2}(O_2, \mathbb{R}^2).$$

Hence the geodesic $\widehat{c}_{\bar{q}, \bar{v}}: (-2, 2) \to M$ is defined, and it has the properties $\widehat{c}_{\bar{q}, \bar{v}}(0) = x(\bar{q}) = q$ and $\widehat{c}'_{\bar{q}, \bar{v}}(0) = dx_{\bar{q}}(\bar{v}) = v$. It follows that $\rho_v \geq 2$, so $v \in E_q$. We have thus proved that $O_{\delta_p}(O_3, T_q M) \subset E_q$, which is part (i) of the proposition.

Now define a map

$$F_q = P_3 \circ \Delta \circ \Phi | Z \circ \Psi^{-1} \circ I_q | O_{\delta_p}(O_3, T_q M): O_{\delta_p}(O_3, T_q M) \to M.$$

Tracing through the effect of this map on the vector v, we see that

$$v \overset{I_q}{\longmapsto} (q, v) \overset{\Psi^{-1}}{\longmapsto} (\bar{q}, \bar{v}) \overset{\Phi}{\longmapsto} (\bar{q}, x^{-1}(\widehat{c}_{\bar{q}, \bar{v}}(1))) \overset{\Delta}{\longmapsto} (q, \widehat{c}_{\bar{q}, \bar{v}}(1)) \overset{P_3}{\longmapsto} \widehat{c}_{\bar{q}, \bar{v}}(1).$$

Thus $F_q(v) = \widehat{c}_{\bar{q}, \bar{v}}(1)$. Using Equation A8.1.1 it follows that

$$F_q(v) = \exp_{x(\tilde{q})}(dx_{\tilde{q}}(\tilde{v})) = \exp_q(v).$$

To prove parts (ii)–(iv) of the proposition we need to verify that the set $F_q(O_{\delta_p}(O_3,$
$T_qM))$ is open in M, that F_q is a diffeomorphism from $O_{\delta_p}(O_3, T_qM)$ onto $F_q(O_{\delta_p}(O_3, T_qM))$, and that $F_q(O_{\delta_p}(O_3, T_qM)) \supset W$. First, we can rewrite F_q as

$$F_q = x \circ P_2 \circ \Phi|Z \circ J_{\tilde{q}} \circ P_2 \circ \Psi^{-1} \circ I_q|O_{\delta_p}(O_3, T_qM);$$

to see that the right hand side equals F_q simply see what it does to any vector $v \in O_{\delta_p}(O_3, T_qM)$. It can be verified that the composition $P_2 \circ \Psi^{-1} \circ I_q|O_{\delta_p}(O_3, T_qM)$ is simply the map $(dx_{\tilde{q}})^{-1}$ restricted to an open subset of T_qM. Thus we deduce that the set $P_2 \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM))$ is open in $\mathbb{R}^2$ and $P_2 \circ \Psi^{-1} \circ I_q|O_{\delta_p}(O_3, T_qM)$ is a diffeomorphism from its domain onto its image. Next, by using Exercise 8.2.2 and the choice of Z, it is seen that $P_2 \circ \Phi|Z \circ J_{\tilde{q}}(P_2 \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM)))$ is open in $\mathbb{R}^2$ and $P_2 \circ \Phi|Z \circ J_{\tilde{q}}|P_2 \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM))$ is a diffeomorphism from its domain onto its image. By Proposition 5.2.5 (i), it follows that $x(P_2 \circ \Phi|Z \circ J_{\tilde{q}} \circ P_2 \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM)))$ is open in M, and from Exercise 5.2.9 it follows that $x|P_2 \circ \Phi|Z \circ J_{\tilde{q}} \circ P_2 \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM))$ is a diffeomorphism from its domain onto its image. Putting these three observations together it follows that the image of F_q is open in M and that F_q is a diffeomorphism from its domain onto its image. Finally, we have

$$\begin{aligned} F_q(O_{\delta_p}(O_3, T_qM)) &= P_3 \circ \Delta \circ \Phi|Z \circ \Psi^{-1} \circ I_q(O_{\delta_p}(O_3, T_qM)) \\ &= P_3 \circ \Delta \circ \Phi|Z \circ \Psi^{-1}(\{q\} \times O_{\delta_p}(O_3, T_qM)) \\ &\supset P_3(\{q\} \times W) = W. \qquad \square \end{aligned}$$

The following lemma, which uses the ideas of the above proof, is necessary for the proof of Theorem A8.1.2. The analog of this lemma for straight lines in the plane is straightforwardly true.

Lemma A8.1.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $p \in M$ be a point. Then there is some number $\zeta_p > 0$ for which the following property holds. Let r be a number such that $0 \leq r < \zeta_p$, and suppose that $c\colon (-\delta, \delta) \to M$ is a*

geodesic such $c(0) \in GS_r(p, M)$ and $c'(0)$ is tangent to $GS_r(p, M)$; then there is some number $\eta > 0$ such that $c(t) \notin \overline{GO}_r(p, M)$ for all $t \in (-\eta, 0) \cup (0, \eta)$.

Proof. See Figure A8.1.1 for an illustration of the lemma. Let W be as in Proposition 8.2.3. To begin with let r be any positive number such that $\overline{GO}_r(p, M) \subset W$; we will choose the number ζ_p later. Without loss of generality we may assume that δ is small enough so that the image of c is contained in W. Let $\widehat{c}\colon (-\delta, \delta) \to T_pM$ be defined by $\widehat{c} = (\exp_p)^{-1} \circ c$. If we let C_r denote the circle in T_pM of radius r centered at the origin, then $\widehat{c}(0) \in C_r$ and $\widehat{c}'(0)$ is tangent to C_r. We now define a function $D\colon (-\delta, \delta) \to \mathbb{R}$ by $D(t) = \|\widehat{c}(t)\|^2$. Observe that the function D is smooth, and that $D(0) = r^2$. Suppose we could show that $D'(0) = 0$ and $D''(0) > 0$. It would then follow from standard results in Calculus that there is some number $\eta > 0$ such that $D(t) > r^2$ for all $t \in (-\eta, 0) \cup (0, \eta)$; in other words, we would know that $\widehat{c}(t)$ is outside the disk bounded by C_r for all $t \in (-\eta, 0) \cup (0, \eta)$, and the lemma would then follow by looking at the image of C_r and $\widehat{c}$ under the map $\exp_p$.

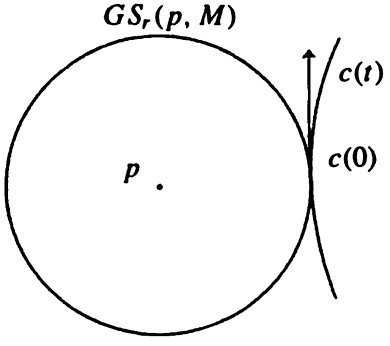


Figure A8.1.1

From the definition of D it follows that $D'(t) = 2\langle \widehat{c}(t), \widehat{c}'(t) \rangle$. Since C_r is a circle, and since $\widehat{c}'(0)$ is tangent to C_r, it follows from Euclidean geometry that $\widehat{c}'(0)$ is perpendicular to $\widehat{c}(0)$. Thus $D'(0) = 0$. We now show that $D''(0) > 0$ for small enough values of r (which is where the number ζ_p in the statement of the lemma shows up). We will make use of the notation of the first two paragraphs of the proof of Proposition 8.2.3.

Let $x\colon U \to M$ be a coordinate patch such that $p \in x(U)$. By shrinking the set U if necessary we may assume without loss of generality that $x(U) \subset W$. By

choosing r small enough we may assume that $\overline{GO}_r(p, M) \subset W$. Let $a = c(0)$, and let $\bar{a} = x^{-1}(a)$. Further, let $\bar{p}$, p_0, Y, $c_{\bar{q},\bar{v}}(t)$, ϵ_1 and ϵ_2 be as in the proof of Proposition 8.2.3. Let $\bar{u} \in O_{\epsilon_2}(O_2, \mathbb{R}^2)$ be a non-zero vector which is a multiple of $(dx_{\bar{a}})^{-1}(c'(0))$; it does not matter which such vector is chosen. Thinking of $dx_{\bar{q}}(\bar{u})$ as a function of $\bar{q}$, it follows from Exercise 5.9.5 that this function is continuous. Hence, if r is chosen small enough then $dx_{\bar{p}}(\bar{u})$ will be non-zero. We will assume that r is so chosen.

We now define a function

$$L\colon (-\epsilon, \epsilon) \times O_{\epsilon_1}(\bar{p}, U) \to \mathbb{R}$$

by

$$L(t, \bar{q}) = \|(\exp_p)^{-1}(c_{\bar{q},\bar{u}}(t))\|^2.$$

Next define a function $H\colon O_{\epsilon_1}(\bar{p}, U) \to \mathbb{R}$ by

$$H(\bar{q}) = \frac{\partial^2 L(t, \bar{q})}{\partial t^2}|_{t=0}.$$

As discussed in the proof of Proposition 8.2.3, the curve $c_{\bar{p},\bar{u}}(t)$ is the unique geodesic in M with $c_{\bar{p},\bar{u}}(0) = p$ and $c'_{\bar{p},\bar{u}}(0) = dx_{\bar{p}}(\bar{u})$. Using Lemma 8.2.2 we see that $(\exp_p)^{-1}(c_{\bar{p},\bar{u}}(t)) = t\, dx_{\bar{p}}(\bar{u})$ for small values of t. Thus $L(t, \bar{p}) = t^2 \|dx_{\bar{p}}(\bar{u})\|^2$. A simple calculation shows that $H(\bar{p}) = 2\|dx_{\bar{p}}(\bar{u})\|^2$, and this vector is strictly positive by hypothesis on r, as mentioned in the previous paragraph. The function H is certainly continuous, thus $H(\bar{q}) > 0$ for all $\bar{q}$ sufficiently near $\bar{p}$.

We now return to our proof that $D''(0) > 0$. Observe that by construction $c_{\bar{a},\bar{u}}(0) = c(0)$ and $c'_{\bar{a},\bar{u}}(0)$ is a non-zero multiple of $c'(0)$; hence by the uniqueness of geodesics we know that $c_{\bar{a},\bar{u}}$ is just a reparametrization of our original geodesic c. Since we are only interested in the image of c we might as well assume that in fact $c(t) = c_{\bar{a},\bar{u}}(t)$ for all small values of t. We thus see that $L(t, \bar{a}) = D(t)$, and hence $H(\bar{a}) = D''(0)$. It follows from the final sentence of the previous paragraph that if r is small enough then $D''(0) > 0$. Combining all the places where we required r to be small yields the desired ζ_p as in the statement of the lemma. □

The following theorem, which is stronger than Proposition 8.2.3, shows the existence of geodesically convex neighborhoods.

Theorem A8.1.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, and let $p \in M$ be a point. Then there is some number γ_p with $0 < \gamma_p < \delta_p$ such that for all numbers r such that $0 < r \leq \gamma_p$ the set $GO_r(p, M)$ has the following properties:*

(i) *for any two points in $x, y \in GO_r(p, M)$ there is a unique geodesic arc contained in $GO_r(p, M)$ with endpoints x and y, and this geodesic arc has the minimal length of all regular arcs in M with endpoints x and y;*

(ii) *$GO_r(p, M) \subset \exp_q(O_{\delta_p}(O_3, T_qM))$ for all $q \in GO_r(p, M)$.*

Proof. Let ζ_p be as in Lemma A8.1.1. Choose some number ϵ such that $0 < \epsilon < (1/2)\min\{\delta_p, \zeta_p\}$. Let V be as in Corollary 8.2.4 using this ϵ. We define γ_p to be some positive number small enough so that $GO_{\gamma_p}(p, M) \subset V$ (using the fact that V is open in M) and so that $\gamma_p \leq \epsilon$.

(ii). This part follows immediately from Corollary 8.2.4.

(i). Let r be any number such that $0 < r \leq \gamma_p$. Let $x, y \in GO_r(p, M)$ be any two points. It follows from property (i) of Corollary 8.2.4 that there is a unique geodesic arc A of length less than ϵ in M joining x and y. Suppose that A is not contained in $GO_r(p, M)$. Take the geodesic arc from p to x, and combine it end-to-end with the arc A, yielding an arc A' from p to y. The arc A' has length less than 2ϵ; it is not necessarily smooth at the point x, but it can be smoothed off at x by some modification in an arbitrarily small neighborhood of x (we will omit the details). It now follows from Exercise 8.3.7 that A' is contained in $GO_{2\epsilon}(p, M)$. Thus the original arc A is contained in $GO_{2\epsilon}(p, M)$, except possibly for a small neighborhood of the point x (since we had to modify A' to make it smooth). However, since $x \in GO_{2\epsilon}(p, M)$ by hypothesis on x, it follows that some small neighborhood of x is contained in $GO_{2\epsilon}(p, M)$. Hence A is entirely contained in $GO_{2\epsilon}(p, M)$. From the definition of ϵ we deduce that A is contained in $GO_\zeta(p, M)$. Since we are assuming that A is not contained in $GO_r(p, M)$, and since both its endpoints are in $GO_r(p, M)$, by the compactness of A there must be some point z in the interior of A that has maximal distance from p. It is then seen that the geodesic circle centered at p and passing through z bounds a closed geodesic ball that entirely contains A. See Figure A8.1.2. Observing that the geodesic arc A must be tangent at z to this geodesic circle, we obtain a contradiction to Lemma A8.1.1.

We now show that A has the minimal length of all regular arcs in M with endpoints x and y. Let C be any other regular arc in M with endpoints x and

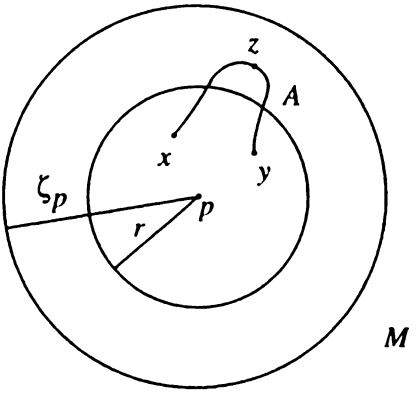


Figure A8.1.2

y. Suppose that A has length d, where $d < \epsilon$. By Lemma 8.3.1 we know that $y \in GS_d(x, M)$. There are now two cases to consider.

Case 1: The arc C is contained in $\overline{GO}_d(x, M)$. It thus follows from part (ii) of this theorem that $C \subset \exp_x(O_{\delta_p}(O_3, T_xM))$. Observe that y has R-coordinate d with respect to x. The desired result now follows from an argument whose outline, is as follows (details are left to the reader): By the compactness of C we see that, except for some arbitrarily small neighborhood of x, the arc C can be broken up into small subarcs each of which is contained in the image of a single geodesic polar coordinate patch D_x; by applying Lemma 8.3.5 to each subarc and adding the resulting inequalities (dropping the absolute value signs), it is seen that the length of C is at least d.

Case 2. The arc C is not contained in $\overline{GO}_d(x, M)$. See Figure A8.1.3. It must be the case that C intersects $GS_d(x, M)$ at some point in the interior of C. Let $z \in C$ be the point in C that intersects $GS_d(x, M)$ closest to x; such a point must exist by the least upper bound property. Observe that the geodesic arc from x to z has length d, and that the subarc of C from x to z is contained in $\overline{GO}_d(x, M)$. By Case 1 we know that the length of the subarc of C from x to z is at least d, and hence so is the length of all of C. □

Exercises

A8.1.1*. Prove that the Jacobian matrix $D\Phi(p_0)$ used in the proof of Proposition 8.2.3 is given by

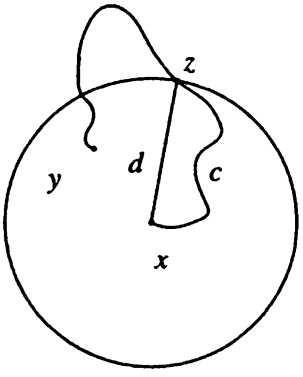


Figure A8.1.3

$$D\Phi(p_0) = \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \end{pmatrix}.$$

A8.1.2*. Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p, q \in M$ be points, let A be a regular arc in M with endpoints p and q, and let δ_p be as in Proposition 8.2.3. If the length of A is d for some $d < \delta_p$, show that $A \subset GO_d(p, M)$.

A8.1.3. Use the ideas in the proof of Theorem A8.1.2 to give an alternative proof of Theorem 7.3.1. More specifically, let $s_0 \in (a, b)$ be any point. Let $q = c(s_0)$, and choose some number $\eta > 0$ small enough so that $c((s_0 - \eta, s_0 + \eta)) \subset GO_{\gamma_q}(q, M)$. Show that the regular arc $c([s_0 - \eta, s_0 + \eta])$ is a geodesic arc, and thus $\frac{Dc'(s)}{ds}|_{s=s_0} = 0$.

Appendix A8.2 Geodesic Triangulations

This section is devoted to a proof of Theorem 8.4.2. We start with two definitions. Let γ_p be as in Theorem A8.1.2.

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. A **convex geodesic ball** in M is any set of the form $GO_r(p, M)$ for some point $p \in M$ and some number r such that $0 < r \leq \gamma_p$. $\Diamond$

Definition. Let $M \subset \mathbb{R}^3$ be a smooth surface. A **geodesic polygonal arc** in M is an arc A in M such that $A = A_1 \cup \cdots \cup A_n$, where the A_i are geodesic arcs, and $A_i \cap A_{i+1}$ is an endpoint of both A_i and A_{i+1} for all i; a **geodesic polygon** in M is a 1-sphere in M that is similarly the union of a finite number of geodesic arcs in M; a **geodesic polygonal disk** in M is a disk in M such that the boundary of the disk is a geodesic polygon. $\Diamond$

Observe that a geodesic triangle is an example of a geodesic polygonal disk. We now have four lemmas, which show that in certain respects geodesics inside geodesically convex balls behave very much like straight lines in the plane.

Lemma A8.2.1. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $U \subset M$ be a convex geodesic ball and let $C \subset U$ be a geodesic polygon. Then there is a geodesic polygonal disk $B \subset U$ such that $\partial B = C$.*

Proof. Exercise A8.2.1. $\square$

Lemma A8.2.2. *Let $M \subset \mathbb{R}^3$ be a smooth surface, let $p \in M$ be a point and let r be a number such that $0 < r \leq \gamma_p$. Then there is a geodesic polygonal disk $B \subset GO_r(p, M)$ such that $p \in \operatorname{int} B$.*

Proof. Let r' be a number such that $0 < r' < r$. Choose a collection of equally spaced points $x_1, \ldots, x_n \in GS_{r'}(p, M)$ for some large positive integer n. Since the points $x_1, \ldots, x_n$ are contained in the convex geodesic ball $GO_r(p, M)$, it follows from Theorem A8.1.2 that for each $i \in \{1, \ldots, n\}$ there is a unique, length-minimizing geodesic arc l_i contained in $GO_r(p, M)$ that joins x_i to x_{i+1} (where addition is mod n). If n is chosen large enough, then each pair of points x_i and x_{i+1} will be so close that the geodesic arc l_i will lie entirely between the two geodesic rays originating at p and containing x_i and x_{i+1} respectively. See Figure A8.2.1. (This assertion follows from the fact that the geodesic arc l_i can intersect each geodesic ray originating at p in at most one point, and if n is large enough then l_i will not contain p; hence l_i must be entirely contained in one of the two regions into which the geodesic rays containing x_i and x_{i+1} divide $GO_r(p, M)$, and if n is large enough l_i will be contained in the smaller region.) The set $l_1 \cup \cdots \cup l_n$ is now seen to be a geodesic polygon contained in $GO_r(p, M)$. The desired result now follows by applying Lemma A8.2.1 to this geodesic polygon. $\square$

Lemma A8.2.3. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let Δxyz be a geodesic triangle contained in a convex geodesic ball. If $t \in \overline{yz}$ is any point, then the*

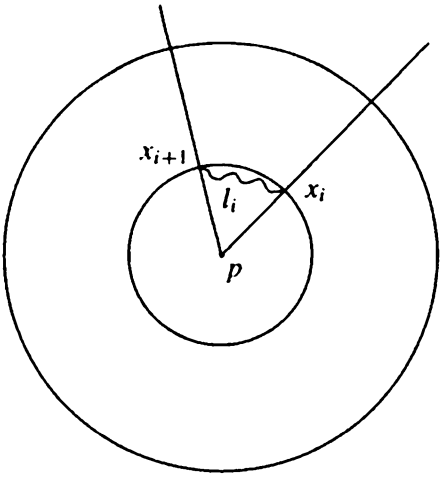


Figure A8.2.1

unique geodesic arc (in the convex geodesic ball) joining x and t is contained in Δxyz.

Proof. We start with the exact same notation and construction as in the second and third paragraphs of the proof of Theorem 8.4.3, which we will not repeat here. The lemma is now proved as follows. Let $t_0 = (D_x)^{-1}(t)$, which is a point on the image of the curve $\widetilde{c}$. Since this curve is the graph of some function of the form $R = f(\theta)$, we see that the vertical line through t_0 intersects the graph of f only at t_0. Tracing through all our definitions it is seen that the vertical line segment from the θ-axis to the point t_0 is contained in $(D_x)^{-1}(\Delta xyz)$ and is mapped by D_x to the geodesic arc in M from x to t, which proves the lemma. □

Lemma A8.2.4. *Let $M \subset \mathbb{R}^3$ be a smooth surface and let $B \subset M$ be a geodesic polygonal disk contained in a convex geodesic ball C. Then B can be written as the union of finitely many geodesic triangles such that if any two of the geodesic triangles intersect, then their intersection is either a common edge or a common vertex.*

Proof. The proof proceeds by induction on the number n of geodesic arcs in the boundary of B. If $n = 3$ then B is a geodesic triangle, and there is nothing to prove. Now suppose that $n > 3$, and that ∂B consists of n geodesic arcs. We assume that the result holds for all geodesic polygonal disks contained in convex balls with fewer than n geodesic arcs in their boundaries. Let y, x and z be

consecutive vertices in ∂B. By definition $C = GO_r(p, M)$ for some appropriate p and r; by Theorem A8.1.2 we know that $C \subset \exp_x(O_{\delta_p}(O_3, T_xM))$. Observe that $(\exp_x)^{-1}(x)$ is the origin; let $Y = (\exp_x)^{-1}(y)$ and $Z = (\exp_x)^{-1}(z)$. Consider the set $J = (\exp_x)^{-1}(\partial B)$. This set is a 1-sphere contained in the open disk $O_{\delta_p}(O_3, T_xM))$ that passes through the origin; the disk bounded by J is just $(\exp_x)^{-1}(B)$. We can think of J as the union of arcs (namely the inverse images of the geodesic arcs that make up ∂B, and any two of these arcs intersect in at most the inverse image of one of the vertices of ∂B. In general, these arcs need not be line segments, though the two arcs containing the origin must be the line segments from the origin to Y and from the origin to Z (since these line segments are the inverse images under $\exp_x$ of the geodesic arcs $\overline{xy}$ and $\overline{xz}$). See Figure A8.2.2. Of the two arcs in J that join Y and Z, let L denote the arc that does not contain the origin.

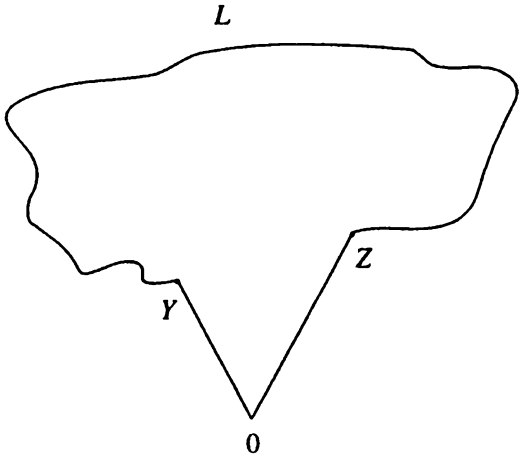


Figure A8.2.2

We now consider all the rays in T_xM that start at the origin and that intersect the interior of $(\exp_x)^{-1}(B)$ near the origin. Each such ray must intersect the arc L at some point; if a ray intersects L in more than one point, consider the point of intersection closest to the origin. See Figure A8.2.3. There are now two possibilities: Either some such ray intersects L in the inverse image of a vertex of ∂B, or not. If we assume the former case, let V denote the inverse image of the appropriate vertex of ∂B (there may be more than one such vertex, so choose one; in any case note that V cannot be either Y or Z). Then the image under

$\exp_x$ of the line segment from the origin to V will be a geodesic arc that cuts B into two geodesic polygonal disks, each of which has fewer than n geodesic arcs in its boundary; we can then finish the proof by using the inductive hypothesis.

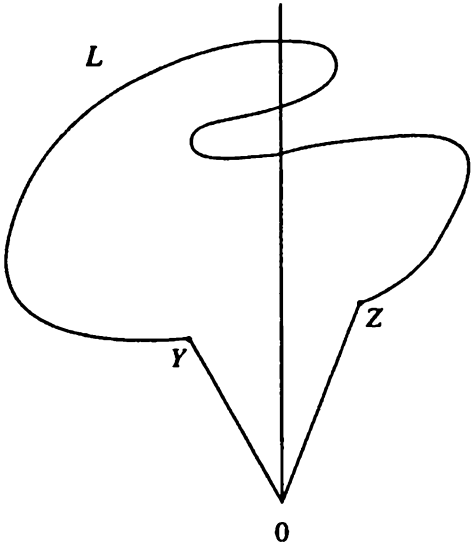


Figure A8.2.3

The other case to be considered is when no ray as discussed in the previous paragraph intersects L in the inverse image of a vertex of ∂B (at the intersection point nearest the origin). In that case it is not hard to see that all such rays must intersect L in the interior of a single arc in L; call this arc α. See Figure A8.2.4. We now have a few subcases. If the arc α does not have either Y or Z as one of its endpoints, then choose the point of intersection of any one of the rays under discussion with α; call this point V; see Figure A8.2.4. Then once again the image under $\exp_x$ of the line segment from the origin to V will be a geodesic arc that cuts B into two geodesic polygonal disks, each of which has fewer than n geodesic arcs in its boundary, and the proof can be finished by using the inductive hypothesis. If α has both Y and Z as endpoints, then B is a geodesic triangle, contradicting our assumption that $n > 3$, so the only remaining case is when α contains precisely one of Y and Z. Assume without loss of generality that α contains Y but not Z; see Figure A8.2.5. Here B can be cut up into pieces by extending the line segment from the origin to Z until it hits α (which it must do, since all the rays converging to this ray intersect α,

and an arc is a closed set); let V denote the point of intersection of the ray from the origin through Z with α. This time we use the image under $\exp_\tau$ of the line segment from Z to V, also a geodesic arc, to cut B into two geodesic polygonal disks, each of which has fewer than n geodesic arcs in its boundary (except in the case when $n = 4$, which we leave to the reader). □

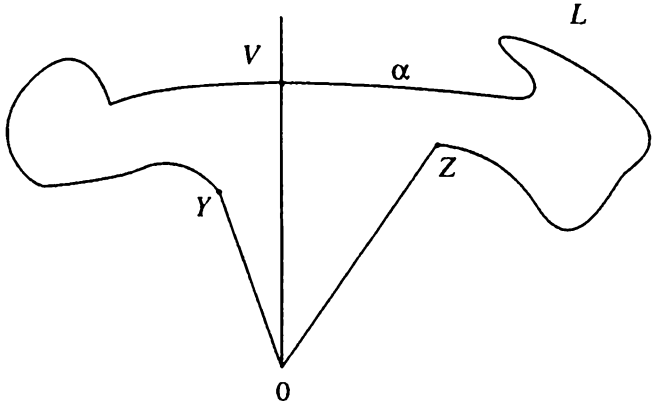


Figure A8.2.4

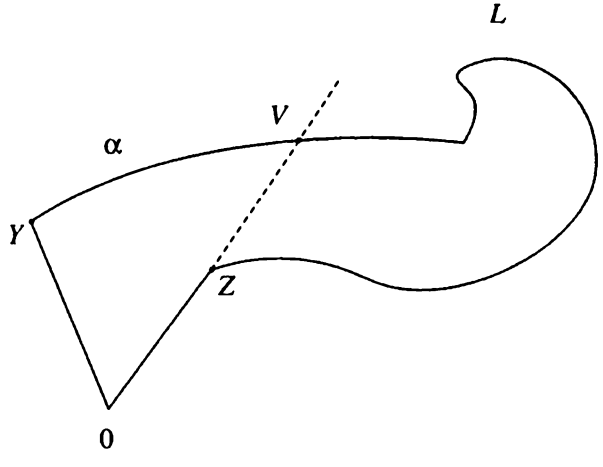


Figure 8A.2.5

We are now ready for the proof of our main result. Our proof is modeled after the construction of triangulations for topological surfaces given in [TH]. Note that any two geodesic arcs in a convex geodesic ball can intersect in at most one point (by the uniqueness of geodesics joining any two points in such a ball).

Proof of Theorem 8.4.2. For each point $p \in M$, let r_p be some number such that $0 < r_p < \min\{\gamma_p, \epsilon\}$. By Lemma A8.2.2 there is a geodesic polygonal disk $B_p \subset GO_{r_p}(p, M)$ such that $p \in \text{int } B_p$. Let $C_p = \partial B_p$. The collection of sets $\{\text{int } B_p\}_{p \in M}$ is an open cover of M, and by the compactness of M there must be a finite collection of sets $\{\text{int } B_{p_1}, \ldots, \text{int } B_{p_k}\}$ that cover M for some positive integer k. It certainly follows that the disks $\{B_{p_1}, \ldots, B_{p_k}\}$ cover M. We may assume without loss of generality that no B_{p_i} is contained entirely in another B_{p_j} (for in that case simply throw out the former set).

First we verify that the union $T = C_{p_1} \cup \cdots \cup C_{p_k}$ partitions M into finitely many geodesic polygonal disks, the boundaries of which are formed out of subarcs of the C_{p_i}. See Figure A8.2.6. Since each C_{p_i} is contained in a geodesically convex set, and since any two geodesic arcs in a geodesically convex set can intersect in at most one point, the set T can be thought of as the union of finitely many geodesic arcs, any two of which intersect in at most a single common endpoint. Let $p \in M - T$ be any point; then p must be contained in $\text{int } B_{p_i}$ for some i. Since any two geodesic arcs in a geodesically convex set can intersect in at most one point, we see that $T \cap B_{p_i}$ can be obtained by taking C_{p_i} and adding to it one at a time finitely many geodesic polygonal arcs $A_1, \ldots, A_q$ made up of subsets of the C_{p_j}. We start with $C_{p_i} \cup A_1$, noting that A_1 intersects C_{p_i} precisely in its two endpoints. Hence $C_{p_i} \cup A_1$ is a theta-curve, as discussed in Exercise 2.2.14. Using that exercise, and the fact that everything is taking place inside a subset of M that is homeomorphic to $\mathbb{R}^2$, it follows that $B_{p_i} - (C_{p_i} \cup A_1)$ consists of two open disks, the boundaries of which are contained in $C_{p_i} \cup A_1$. The point p must be contained in one of these open disks, and we restrict attention to the closure of this disk, denoted D. We now proceed as above, using D and the arc A_2. If A_2 does not intersect D then we proceed to A_3; if it does intersect A_2 then once again $\partial D \cup A_2$ forms a theta-curve, and the point p is now contained in an even smaller open disk, the boundary of which is contained in $\partial D \cup A_2 \subset C_{p_i} \cup A_1 \cup A_2$. Proceeding in this manner until all the A_j are used up, we see that p is contained in some open disk the boundary of which is contained in $C_{p_i} \cup A_1 \cup \cdots \cup A_q = B_{p_i} \cap T$.

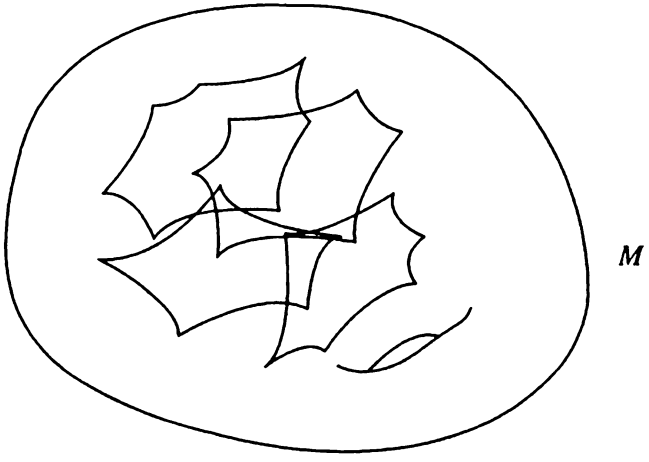


Figure A8.2.6

Our claim now follows.

We can now think of M as broken up into finitely many geodesic polygonal disks, any two of which intersect in at most some collection of geodesic arcs in their boundaries. The next step is to apply Lemma A8.2.4 to each of these geodesic polygonal disks one at a time. The result is that M can be written as the union of finitely many geodesic triangles such that if any two of the geodesic triangles intersect, then their intersection is either a common edge or a common vertex. The desired geodesic triangulation $t: |K| \to M$ can now be constructed using an argument similar to the construction used in the proof of Lemma 3.3.10 (i); details are left to the reader. □

Exercise

A8.2.1*. Prove Lemma A8.2.1.

Endnotes

Notes for Section 8.1

(A) There are higher-dimensional versions of the Gauss–Bonnet Theorem. The original higher-dimensional formulations ([AD], [FN] and [A-W]) were

geometric in nature; the more modern version of this theorem is formulated in ultimate generality using such high-tech tools as bundles, cohomology and characteristic classes (see, for example, [SK3, vol. V, Chapter 13] and [M-S, Appendix C]).

(B) In addition to the Gauss–Bonnet Theorem there are some other very nice results relating the Euler characteristic to certain geometric quantities defined for smooth surfaces (and more generally smooth manifolds). Two such results are the Poincaré–Hopf Theorem concerning vector fields on manifolds (see [MI3, §6] or [DO1, p. 282]) and the Morse inequalities (see [MI2, §5], or [HR, p. 161]).

Notes for Section 8.3

We showed that simplicial curvature and Gaussian curvature are somewhat analogous by showing that the latter can be expressed in terms of a smooth angle defect. Alternately, we might attempt to construct a simplicial analog of the Gauss map, and then prove that the angle defect can be expressed in terms of the simplicial Gauss map. This approach can be found in [BL, §§2, 6, 7]; the sections referred to are elementary and can be read independently of the rest of the paper.

Notes for Section 8.4

Our proof of the Gauss–Bonnet Theorem (the essence of which is the proof of Theorem 8.4.3) is not very geometrically appealing. A nicer approach (which we have not taken to avoid a number of technicalities) would be to make a more explicit use of some version of Stokes' theorem (for example Green's theorem, which relates an integral over a region in the plane to a line integral over the boundary of the region — essentially a two-dimensional version of the Fundamental Theorem of Calculus). We could then apply such a theorem to the integral of Gaussian curvature over a polygon in the surface, where the boundary of the polygon is made up of a number of smooth arcs glued end-to-end. Piecing together the integrals over these polygonal regions would then yield the Gauss–Bonnet Theorem. Such a proof can be found in [DO1] or [SK3, vol. III].

Notes for Section A8.1

Theorem A8.1.2 was first proved in [WH1]; we follow the treatment in [DO1, §4-7].

Notes for Section A8.2

(A) Rather than using geodesic triangulations, some differential geometry texts make do with C^∞ triangulations, which are essentially triangulations in which the restriction of the triangulation map to each simplex is a smooth map (where we think of an i-simplex as sitting in $\mathbb{R}^i$). See [WH2] or [MU1] for more details. It is much easier to find C^∞ triangulations than geodesic triangulations. However, to avoid the use of geodesic triangulations in the proof of the Gauss–Bonnet Theorem (Theorem 8.1.1), other technicalities (which we have not discussed) are needed as compensation.

(B) In those books that do make use of geodesic triangulations, the reader will be hard-pressed to find a proof of the fact that all compact smooth surfaces have geodesic triangulations, or even a reference for a proof. This result is what is known as a "folk-theorem"; everyone knows that it is true, but it is hard to find a detailed proof. The only relevant reference the author could find is [PI], in German, which proves that any topological triangulation of a smooth surface can be approximated by a triangulation with edges that are piecewise geodesics; the result that we need could then be deduced from this theorem.

Appendix

Affine Linear Algebra

We assume that the reader is familiar with the fundamentals of linear algebra (at least insofar as it applies to $\mathbb{R}^n$), including bases, linear maps, matrices, determinants and eigenvalues; see for example [LA1] or [FR] for more details. Because we will be working not only with subspaces of $\mathbb{R}^n$ but with translates of subspaces (such as lines and planes in $\mathbb{R}^3$ that do not contain the origin), we need the variant of linear algebra called affine linear algebra. In linear algebra the fundamental construction is that of a linear combination of vectors; subspaces are closed under linear combinations, and linear maps preserve linear combinations. Because all the coefficients in a linear combination can be zero, the zero vector is contained in any subspace. To obtain translates of subspaces we restrict our attention to linear combinations in which the sum of the coefficients equals 1. The material in this section is analogous to standard results in linear algebra; we leave it to the reader to provide examples and most of the proofs. We start with the analogs of linear independence and span.

Definition. Let $x_0, \ldots, x_k \in \mathbb{R}^n$ be points. An **affine combination** of these points is a linear combination $\sum_{i=0}^k t_i x_i$ where $t_0, \ldots, t_k \in \mathbb{R}$ are numbers such that $\sum_{i=0}^k t_i = 1$. The **affine span** of these points, denoted $\text{aspan}\{x_0, \ldots, x_k\}$, is the set of all affine combinations of the points. The set $\{x_0, \ldots, x_k\}$ is **affinely independent** if the conditions $\sum_{i=0}^k t_i x_i = O_n$ and $\sum_{i=0}^k t_i = 0$ for numbers $t_0, \ldots, t_k \in \mathbb{R}$ imply that $t_i = 0$ for all $i \in \{0, \ldots, k\}$. $\Diamond$

A geometric characterization of affinely independent sets is given in the first part of the following lemma.

Lemma A.1. *Let $x_0, \ldots, x_k \in \mathbb{R}^n$ be points.*

(i) *The set $\{x_0, \ldots, x_k\}$ is affinely independent iff the set $\{x_1 - x_0, \ldots, x_k - x_0\}$ is linearly independent.*

(ii) *If $\{x_0, \ldots, x_k\}$ is affinely independent, then each $x \in \text{aspan}\{x_0, \ldots, x_k\}$ is uniquely expressible as $x = \sum_{i=0}^k t_i x_i$ for some numbers $t_0, \ldots, t_k \in \mathbb{R}$ such that $\sum_{i=0}^k t_i = 1$.*

Proof. (i). Suppose $\{x_0, \ldots, x_k\}$ is not affinely independent, so that there exist numbers $t_0, \ldots, t_k \in \mathbb{R}$ that are not all zero such that $\sum_{i=0}^{k} t_i x_i = O_n$ and $\sum_{i=0}^{k} t_i = 0$. By the last equality it cannot be the case that only one of the t_i is non-zero; hence one of $t_1, \ldots, t_k$ is non-zero. We now have

$$O_n = \sum_{i=0}^{k} t_i x_i = \sum_{i=0}^{k} t_i x_i - \Big(\sum_{i=0}^{k} t_i\Big) x_0 = \sum_{i=1}^{k} t_i (x_i - x_0),$$

and hence the set $\{x_1 - x_0, \ldots, x_k - x_0\}$ is linearly dependent.

Conversely, suppose that $\{x_1 - x_0, \ldots, x_k - x_0\}$ is linearly dependent, so that there exist numbers $t_1, \ldots, t_k \in \mathbb{R}$ that are not all zero such that $\sum_{i=1}^{k} t_i (x_i - x_0) = O_n$. If we let $t_0 = -\sum_{1=1}^{k} t_i$, then $\sum_{i=0}^{k} t_i = 0$ and

$$\sum_{i=0}^{k} t_i x_i = t_0 x_0 + \sum_{i=1}^{k} t_i x_i = -\Big(\sum_{1=1}^{k} t_i\Big) x_0 + \sum_{i=1}^{k} t_i x_i = \sum_{i=1}^{k} t_i (x_i - x_0) = O_n.$$

Hence $\{x_0, \ldots, x_k\}$ is affinely independent.

(ii). The existence of the numbers $t_0, \ldots, t_k$ with the desired properties follows immediately from the definition of affine span, and the uniqueness follows from a standard argument that we leave to the reader. □

It is now straightforward to see that a collection consisting of a single point is always affinely independent; two points are affinely independent if they are distinct; three points are affinely independent iff they are not collinear; four points are affinely independent if they are not coplanar. Observe that a collection of more than $n + 1$ points in $\mathbb{R}^n$ cannot be affinely independent. Next, we have the analogs of subspaces.

Definition. A subset $X \subset \mathbb{R}^n$ is an **affine subspace** if it is closed under affine combinations, that is, if $x_0, \ldots, x_k \in X$ are points and $t_0, \ldots, t_k \in \mathbb{R}$ are numbers such that $\sum_{i=0}^{k} t_i = 1$, then $\sum_{i=0}^{k} t_i x_i \in X$. An **affine basis** for an affine subspace $X \subset \mathbb{R}^n$ is an affinely independent set of points $\{x_0, \ldots, x_k\} \subset X$ such that $\operatorname{aspan}\{x_0, \ldots, x_k\} = X$. ◊

The following lemma is analogous to a standard result in linear algebra.

Lemma A.2. *Every affine subspace of $\mathbb{R}^n$ has an affine basis, and all such affine bases have the same number of elements (which is finite).*

Proof. Let $X \subset \mathbb{R}^n$ be an affine subspace. If $X = \emptyset$ then the result holds vacuously, so assume that $X \neq \emptyset$. Choose any point $p \in X$, and let $Y =$

$\{x - p \mid x \in X\}$. We start by showing that Y is a vector subspace of $\mathbb{R}^n$. Let $v, w \in Y$ and $a, b \in \mathbb{R}$. Then $v + p, w + p \in X$. Further, let $c = 1 - (a + b)$. Hence $a + b + c = 1$, so $a(v + p) + b(w + p) + cp \in X$ by the definition of affine subspaces. Since $a(v + p) + b(w + p) + cp = (av + bw) + p$, it follows that $av + bw \in Y$. Thus Y is a vector subspace.

Any vector subspace of $\mathbb{R}^n$ has a finite basis, so let $\{v_1, \ldots, v_k\} \subset Y$ be a basis for Y. We will show that the set $\{p, v_1 + p, \ldots, v_k + p\}$ is an affine basis for X. First, suppose that $t_0 p + t_1(v_1 + p) + \cdots + t_k(v_k + p) = O_n$ for some numbers $t_0, \ldots, t_k \in \mathbb{R}$ such that $\sum_{i=0}^{k} t_i = 0$. It follows that $\sum_{i=1}^{k} t_i v_i = O_n$, and by linear independence we deduce that $t_i = 0$ for all $i \in \{1, \ldots, k\}$. It then follows that $t_0 = 0$, so $\{p, v_1 + p, \ldots, v_k + p\}$ is affinely independent. Next, let $x \in X$ be a point. Then $x - p \in Y$, so there are numbers $t_1, \ldots, t_k \in \mathbb{R}$ such that $x - p = \sum_{i=1}^{k} t_i v_i$. If we let $t_0 = 1 - \sum_{i=1}^{k} t_i$, then it is seen that $x = t_0 p + t_1(v_1 + p) + \cdots + t_k(v_k + p)$ and $\sum_{i=0}^{k} t_i = 1$. Hence $x \in \operatorname{aspan}\{p, v_1 + p, \ldots, v_k + p\}$.

The fact that all affine bases of X have the same number of elements follows from a similar argument, using the fact that all bases for Y have the same number of elements. □

By the above lemma we can make the following definition.

Definition. An affine subspace of $\mathbb{R}^n$ has **dimension** k if it has an affine basis with $k + 1$ points. A k-dimensional affine subspace of $\mathbb{R}^n$ is called a **k-plane.** ◊

A k-plane in $\mathbb{R}^n$ need not be a vector subspace of $\mathbb{R}^n$, though a k-dimensional vector subspace is a k-plane. The precise relation between vector subspaces and k-planes is given in the following lemma; the proof of this lemma is just like the proof of Lemma A.2.

Lemma A.3. *If $N \subset \mathbb{R}^n$ is a k-dimensional vector subspace of $\mathbb{R}^n$ and $p \in \mathbb{R}^n$ is a point, then the set $\{x + p \mid x \in N\}$ is a k-plane; conversely, any k-plane in $\mathbb{R}^n$ has this form.*

We see, for example, that a 1-plane in $\mathbb{R}^n$ is just a straight line and a 2-plane is just a plane. The only n-plane in $\mathbb{R}^n$ is $\mathbb{R}^n$ itself. The following lemma is once again as expected.

Lemma A.4. *Let $x_0, \ldots, x_k \in \mathbb{R}^n$ be points. Then $\operatorname{aspan}\{x_0, \ldots, x_k\}$ is an affine subspace of $\mathbb{R}^n$. Suppose the set $\{x_0, \ldots, x_k\}$ is affinely independent.*

Then aspan$\{x_0, \ldots, x_k\}$ is a k-plane, and it is the unique k-plane containing these points.

We now turn to the analog of linear maps.

Definition. Let $X \subset \mathbb{R}^n$ be an affine subspace. A map $F: X \to \mathbb{R}^m$ is **affine linear** if it preserves affine combinations, that is, if $x_0, \ldots, x_k \in X$ are points and $t_0, \ldots, t_k \in \mathbb{R}$ are numbers such that $\sum_{i=0}^k t_i = 1$, then $F(\sum_{i=0}^k t_i x_i) = \sum_{i=0}^k t_i F(x_i)$. $\Diamond$

The following lemma shows the relation of affine linear maps $\mathbb{R}^n \to \mathbb{R}^m$ to linear maps.

Lemma A.5. *Let $F: \mathbb{R}^n \to \mathbb{R}^m$ be an affine linear map. Then there is a linear map $L: \mathbb{R}^n \to \mathbb{R}^m$ and a vector $p \in \mathbb{R}^m$ such that $F(x) = L(x) + p$ for all $x \in \mathbb{R}^n$.*

Proof. Define the map $L: \mathbb{R}^n \to \mathbb{R}^m$ by $L(x) = F(x) - F(\mathrm{O}_n)$ for all $x \in \mathbb{R}^n$, and let $p = F(\mathrm{O}_n)$. To prove the lemma it suffices to show that the map L so defined is linear. Let $v, w \in \mathbb{R}^n$ and $a, b \in \mathbb{R}$. If we let $c = 1 - (a + b)$ then $a + b + c = 1$, and by the definition of affine linearity we have

$$\begin{aligned} L(av + bw) &= F(av + bw) - F(\mathrm{O}_n) = F(av + bw + c\mathrm{O}_n) - F(\mathrm{O}_n) \\ &= aF(v) + bF(w) + cF(\mathrm{O}_n) - F(\mathrm{O}_n) \\ &= a(F(v) - F(\mathrm{O}_n)) + b(F(w) - F(\mathrm{O}_n)) \\ &= aL(v) + bL(w). \quad \square \end{aligned}$$

An affine linear map need not be a linear map. If $X \subset \mathbb{R}^n$ is an affine subspace, and if $F: X \to \mathbb{R}^m$ is an affine linear map, then F is the restriction to X of an affine linear map $\mathbb{R}^n \to \mathbb{R}^m$; this claim follows from Lemma A.7 below and the fact that any affinely independent set can be extended to an affine basis. Hence F is also the restriction of a linear map followed by a translation.

The following lemma is analogous to another standard result in linear algebra.

Lemma A.6. *Let $X \subset \mathbb{R}^n$ be an affine subspace, and let $F: X \to \mathbb{R}^m$ be an affine linear map.*

(i) *The image of F is an affine subspace of $\mathbb{R}^m$. If F is injective then the dimension of $F(X)$ equals the dimension of X.*

(ii) *If F is injective, then the inverse map $F^{-1}: F(X) \to X$ is affine linear.*

We see from this lemma that an injective affine linear map takes straight lines to straight lines. Just as a linear map is determined by what it does to an appropriate number of linearly independent vectors, affine linear maps behave similarly to an appropriate number of affinely independent points, as seen in the following lemma.

Lemma A.7. *Let $X \subset \mathbb{R}^n$ be a k-plane, let $\{x_0, \ldots, x_k\} \subset X$ be an affine basis of X, and let $y_0, \ldots, y_k \in \mathbb{R}^m$ be any points (not necessarily affinely independent). Then there is a unique affine linear map $F: X \to \mathbb{R}^m$ such that $F(x_i) = y_i$ for all $i \in \{1, \ldots, k\}$. The map F is injective iff the points $y_0, \ldots, y_k$ are affinely independent.*

Further Study

The material in this book is merely an introduction to a number of branches of mathematics, and here we recommend some books for each topic. This list is certainly not exhaustive.

1. Collateral Reading

A number of books can be used as supplements to the present text. A classic expository work on geometry, which includes some nice illustrations in differential geometry, is [H-CV]. Of more recent vintage is the highly recommended [WE], which gives an intuitive treatment of the geometry and topology of surfaces and 3-manifolds (very much inspired by the work of Weeks' adviser, W. Thurston). The first few chapters are particularly germane to our topic, though for our purposes one need not pay attention to the emphasis on hyperbolic geometry. This is also the only book I know that includes tic-tac-toe on the torus and the Klein bottle.

To do [WE] correctly, read the classic [AB] first. This little volume, written by a Victorian schoolmaster, was as much a satire of Victorian society as a mathematics text, but it has served as an inspiration to many mathematicians and non-mathematicians alike in thinking about higher-dimensional space. A more recent sequel to [AB] is [BU], which contains some nice mathematical topics, but which does not have the satirical style of its predecessor. Another book on higher dimensions for a popular audience is [RU], although some of the more speculative parts of this book should be taken with many grains of salt. A very nice recent book on higher dimensions, of interest to mathematicians and non-mathematicians alike, is [BA4].

On a more standard note, a textbook that might make a nice complement to the present book is [NA]. This book, like the present text, focuses on geometric questions concerning subsets of Euclidean spaces, but it covers topics we do not, such as the fundamental group, simplicial homology, and differential topology. A nice little volume that not only discusses the topology of surfaces but also gives many historical references to the development of the subject is [F-F]. Two very recent books that discuss a number of topics concerning surfaces that we have skipped (such as group actions and covering spaces) are [ST] and [KY].

The recent book [FO] treats a variety of topics, including some differential geometry and geometric topology of surfaces, and supplements rigor with remarkable drawings. Another phenomenal source of drawings of surfaces in various stages of deformation, with accompanying explanations, is [FC]. The text [MCL] nicely puts the differential geometry of surfaces in the context of the development of non-Euclidean geometry, in much more detail than our brief discussion in Section 8.5. Another text that expands upon our treatment of smooth surfaces in $\mathbb{R}^3$ is [MG], which gives a very concrete introduction to Riemannian geometry via surfaces in $\mathbb{R}^n$, discussing general relativity in the process.

2. Point Set Topology (also known as General Topology)

For more advanced work it is necessary to discuss topological properties of sets that do not naturally sit in any Euclidean space. The most general setting for such study is the notion of a topological space; the study of the axiomatic properties of such spaces is called point set topology. Point set topology is both the foundational material for all branches of topology and a good place to practice proof techniques. An excellent text on point set topology is [MU2]; it would suffice to read Chapters 2–4, although the reader familiar with groups should definitely read Chapter 8 about the fundamental group. Another nice text is [AR], which may not quite match [MU2] for pure expository style, but which has the advantage of moving as quickly as possible to geometric topics. Another such text is [JA], although its lack of exercises is a serious drawback. A classic point set textbook often recommended by authors of an earlier generation is [KE]. This book covers important material and is considered to have good problems; on the other hand, it does not have a single figure.

3. Algebraic Topology

Algebraic topology is the study of topological problems using tools from abstract algebra such as groups and rings. The basic idea is to associate with each topological space various algebraic objects that reflect the properties of the original space. The first topics usually studied in algebraic topology are the fundamental group, covering spaces, homology groups and homotopy groups.

Even someone primarily interested in geometric questions will need these tools for advanced work. Algebraic topology has become a subject in its own right as well as a tool for other branches of mathematics. A classic text on the fundamental group and covering spaces (as well as surfaces) is [MS1]. Other introductory texts covering simplicial homology and the fundamental group (among other things) are [NA], [AR] and [CR]. Of the many advanced texts on algebraic topology, two of the more accessible are [MU3], and [MS2]. The text [MU3] is particularly recommended for its geometric approach, including a nice treatment of simplicial cohomology. A slightly older book that covers point set and algebraic topology from a geometric point of view is [H-Y]. The ultimate reference book for algebraic topology is [SP], though as a first exposure to the subject one should proceed at one's own risk.

4. Geometric Topology

Geometric topology focuses primarily on the study of manifolds, which are the higher-dimensional analog of surfaces. The restriction to manifolds, as opposed to general topological spaces, allows for a more geometric flavor (of the "rubber-sheet" variety) than in point set topology. Manifolds come in three varieties: topological; piecewise linear (abbreviated PL), which generalize what we have been calling simplicial surfaces; and differential (also known as smooth). In the two-dimensional case (that is, surfaces) these three categories essentially coincide, in that any surface of one type is homeomorphic to a surface of any of the other two types. In higher dimensions the three categories of manifolds behave quite differently from one another; for example, there are topological manifolds that are not homeomorphic to any differential manifold. The topological properties of differential manifolds are the subject of differential topology and will be discussed in Item 6 below. Geometric topology focusses on topological and PL manifolds. One needs to learn about general topological spaces and the fundamental group (at least) before attempting the books mentioned here on topological and PL manifolds.

A very nice text, and one upon which the current book draws fairly heavily, is [MO], dedicated to surfaces and 3-manifolds. The book contains proofs of the triangulability of 2-manifolds and 3-manifolds, the latter being quite difficult (and which was first proved by Moise). Also discussed are things such as wild spheres and wild arcs. Another book on geometric topology that

maintains a low-dimensional, geometric point of view is [BI]. (Moise and Bing were classmates at the University of Texas, working under R.L. Moore.) Some standard texts on PL topology, all at the graduate level, are [ZE], [R-S], [HU], and [GL]. PL topology often appears quite formal at first encounter, especially [HU]. The text [ZE] is quite nice, and served as an inspiration for later texts on the subject, but is only available as unpublished lecture notes.

One very pretty geometric topic is knot theory, a branch of geometric topology but with a flavor all its own. Not only is this subject geometrically appealing, but there have recently been found some connections between it and such applied fields as quantum mechanics and DNA. Some books on the subject are [RO], [B-Z], and [KA].

Though closer to geometry and combinatorics than geometric topology, the study of polyhedra is both of inherent geometric interest and of use in applications. A nice discussion of the history of the study of polyhedra, which goes back to the ancient world, is given in [S-F, §4]. The combinatorial approach to polyhedra is taken in [GR1], [GR2] and [BD]. Applications to optimization can be found in [Y-K-K].

5. Differential Geometry

Differential geometry is an older subject than topology, having received a major impetus from the work of Gauss and Riemann. For historical comments see the appendix of [M-P]. Classical differential geometry is concerned with curves and surfaces in $\mathbb{R}^3$, as discussed in the present text. Three books taking the classical point of view, which contain material not covered here and upon which the current text has relied, are [M-P], [KL] and [DO1]. The last is particularly recommended; the text [M-P] is the most elementary, though not always elegant; there is much nice material in [KL], but the discussion is often rather terse.

Two main changes occur when moving beyond classical differential geometry: higher-dimensional manifolds are treated, and more advanced technologies (such as moving frames, differential forms, lie groups and vector bundles) are used. Although these more advanced techniques may be applied to surfaces, the advanced techniques are crucial in higher dimensions, where there are complications that do not arise in the case of surfaces in $\mathbb{R}^3$. Three undergraduate texts, slightly more advanced than the three mentioned above, are [ON], [S-T] and [TR]. The first of these two books treats moving frames, and the second

discusses point set and algebraic topology as well as differential geometry, and includes the famous deRham Theorem. Some graduate level differential geometry texts are [DO2], [HI], [BO] and [K-N]. The ultimate introduction to differential geometry is the five volume opus [SK3 vols. I– V]. The coverage in this work is as follows: vol. I — the basics of smooth manifolds, differential forms, etc.; vol. II — an extremely thorough treatment of curvature and connections, in which the same topic is discussed via a sequence of approaches that roughly follows historical development, starting with the work of Gauss and Riemann; vol. III — classical surface theory (although one needs tools from the first two volumes); vol. IV — higher-dimensional manifolds; vol. V — advanced topics, including the generalized Gauss–Bonnet Theorem. The bibliography in vol. V is quite thorough. The five volumes [SK3 vols. I–V] are known for their exploratory and sometimes humorous style.

6. Differential Topology

This area is at the intersection of the various parts of the current text: the study of the topological properties of differential manifolds. Though certainly serving as foundational material for advanced differential geometry, differential topology has become a subject area distinct from either geometric topology or differential geometry and has seen major advances in the past 40 years. To study differential topology, advanced Calculus is definitely needed; see, for example, the classic [SK1] or the recent [MU4]. Some point set topology is also needed, and basic algebraic topology is necessary for the more advanced texts. An excellent place to start is the beautiful little book [MI3]. Two other introductory texts are [WA] and [B-J]. There is also some introductory material on differential topology in [NA]. Other books to look at, all at the graduate level, are [BO], [HR], [MU1], [WR] and [SK3 vol. I].

Finally, two books to which any student interested in the study of smooth manifolds should aspire are [MI2] and [M-S]. Both these books are influential graduate level texts, covering beautiful material and written in a style many mathematicians seek to emulate. Both books are based on notes taken during lectures by J. Milnor, one of the most important topologists of the last 40 years; one of the note-takers for [MI2] was M. Spivak, author of [SK3].

References

[AB] Abbott, E. A., *Flatland*, Dover, New York, 1952.

[AL2] Alexander, J. W., *An example of a simply connected surface bounding a region which is not simply connected*, Proc. Nat. Acad. Sci. U.S.A. **10** (1924), 8–10.

[AL3] ———, *Remarks on a point set constructed by Antoine*, Proc. Nat. Acad. Sci. U.S.A. **10** (1924), 10–12.

[AD] Allendoerfer, C. B., *The Euler number of a Riemann manifold*, Amer. J. Math. **62** (1940), 243–248.

[A-W] Allendoerfer, C. B., and Weil, A., *The Gauss–Bonnet theorem for Riemannian polyhedra*, Trans. Amer. Math. Soc. **53** (1943), 101–129.

[AN1] ———, *Sur la possibilité d'étendre l'homéomorphie de deux figures à leurs voisinages*, C. R. Acad. Sci. Paris **171** (1920), 661–663.

[AN2] Antoine, L., *Sur l'homéomorphie de figures et de leurs voisinages*, J. Math. Pures Appl. **86** (1921), 221–324.

[AR] Armstrong, M. A., *Basic Topology*, Springer-Verlag, New York, 1983.

[BA1] Banchoff, T., *Critical points and curvature for embedded polyhedra*, J. Diff. Geom. **1** (1967), 245–256.

[BA2] ———, *Critical points and curvature for embedded polyhedral surfaces*, Amer. Math. Monthly **77** (1970), 475–485.

[BA3] ———, *Critical points and curvature for embedded polyhedra II*, Progress in Math. **32** (1983), 34–55.

[BA4] ———, *Beyond the Third Dimension*, Scientific American Library, New York, 1990.

[BP] Barr, S., *Experiments in Topology*, Crowell, New York, 1964.

[BT] Bartle, R. G., *The Elements of Real Analysis*, John Wiley & Sons, New York, 1964.

[BE] Berger, M., *Convexity*, Amer. Math. Monthly **97** (1990), 650–678.

[BI] Bing, R H, *The Geometric Topology of 3-Manifolds*, AMS Colloquium Publications, vol. 40, American Mathematical Society, Providence, RI, 1983.

[BL] Bloch, E. D., *A combinatorial Chern–Weil theorem for 2-plane bundles with even Euler characteristic*, Israel J. Math. **67** (1989), 193–216.

[BO] Boothby, W. M., *An Introduction to Differentiable Manifolds and Riemannian Geometry*, Academic Press, New York, 1975.

[BR] Brahana, H. R., *Systems of circuits on two-dimensional manifolds*, Ann. of Math. **23** (1922), 144–68.

[B-J] Bröcker, Th., and Jänich, K., *Introduction to Differential Topology*, Cambridge U. Press, Cambridge, 1982.

[BD] Brøndsted, A., *An Introduction to Convex Polytopes*, Springer-Verlag, New York, 1983.

[BN] Brown, M., *Locally flat imbeddings of topological manifolds*, Ann. of Math. **75** (1962), 331–341.

[B-Z] Burde, G., and Zieschang, H., *Knots*, De Gruyter, New York, 1985.

[BU] Burger, D., *Sphereland*, Perennial Library, Harper & Row, New York, 1965.

[BG] Burgess, C. E., *Classification of surfaces*, Amer. Math. Monthly **92** (1985), 349–454.

[CA1] Cairns, S. S., *Homeomorphisms between topological manifolds and analytic manifolds*, Ann. of Math. **41** (1940), 796–808.

[CA2] ———, *An elementary proof of the Jordan–Schoenflies theorem*, Proc. Amer. Math. Soc. **2** (1951), 860–867.

[CS] Cassels, J. W. S., *Economics for Mathematicians*, London Math. Soc. Lecture Note Series 22, Cambridge U. Press, Cambridge, 1981.

[CE] Cederberg, J. N., *A Course in Modern Geometries*, Springer-Verlag, New York, 1989.

[CH] Chern, S.-S., *What is geometry?*, Amer. Math. Monthly **97** (1990), 679–686.

[C-M-S] Cheeger, J., Muller, W., and Schrader, R., *On the curvature of piecewise flat spaces*, Commun. Math. Phys. **92** (1984), 405–454.

[CR] Croom, F. H., *Basic Concepts of Algebraic Topology*, Springer-Verlag, New York, 1978.

[DE] Debreu, G., *Theory of Value*, Yale U. Press, New Haven, CT, 1959.

[DI] Dierker, E., *Topological Methods in Walrasian Economics*, Lecture Notes in Economics and Mathematical Systems #92, Springer-Verlag, Berlin, 1974.

[DO1] Do Carmo, M., *Differential Geometry of Curves and Surfaces*, Prentice-Hall, Englewood Cliffs, NJ, 1976.

[DO2] Do Carmo, M., *Riemannian Geometry*, Birkhäuser, Boston, 1992.

[D-F-N] Dubrovin, B. A., Fomenko, A. T., and Novikov, S. P., *Modern Geometry — Methods and Applications*, parts I–III, Springer-Verlag, New York, 1984.

[DU] Dugundji, J., *Topology*, Allyn & Bacon, Boston, 1966.

[EI] Eisenhardt, L. P., *A Treatise on the Differential Geometry of Curves and Surfaces*, Ginn, Boston, 1909.

[EU] Euclid, *The Thirteen Books of the Elements*, Dover, New York, 1956.

[FE] Federico, P. J., *Descartes on Polyhedra*, Springer-Verlag, New York, 1982.

[FN] Fenchel, W., *On total curvatures of Riemannian manifolds: I*, J. London Math. Soc. **15** (1940), 15–22.

[FL] Flanders, H., *Differential Forms*, Academic Press, New York, 1963.

[FO] Fomenko, A., *Visual Geometry and Topology*, Springer-Verlag, New York, 1994.

[FC] Francis, G. K., *A Topological Picture Book*, Springer-Verlag, New York, 1987.

[F-F] Fréchet, M., and Fan, K., *Initiation to Combinatorial Topology*, Prindle, Weber & Schmidt, Boston, 1967.

[F-L] Freedman, M. H., and Luo, F., *Selected Applications of Geometry to Low-Dimensional Topology*, American Mathematical Society, Providence, RI, 1989.

[F-Q] Freedman, M. H., and Quinn, F., *Topology of 4-manifolds*, Princeton U. Press, Princeton, 1990.

[FR] Friedberg, S., Insel, A., and Lawrence, E., *Linear Algebra*, 2nd ed., Prentice-Hall, Engelwood Cliffs, NJ, 1989.

[GA] Gauss, K. F., *General Investigations of Curved Surfaces*, Raven Press, New York, 1965.

[GL] Glaser, L. C., *Geometrical Combinatorial Topology*, vols. I–II, Van Nostrand Reinhold, New York, 1970.

[GE] Greenberg, M. J., *Euclidean and Non-Euclidean Geometry*, W. H. Freeman, San Francisco, 1974.

[GR1] Grünbaum, B., *Convex Polytopes*, Wiley, New York, 1967.

[GR2] ———, *Grassman angles of convex polytopes*, Acta. Math. **121** (1968), 293–302.

[HM] Hamilton, A. G., *Numbers, Sets and Axioms*, Cambridge U. Press, Cambridge, 1982.

[HE] Hempel, J., *3-manifolds*, Ann. of Math. Studies, vol. 86, Princeton U. Press, Princeton, 1976.

[HI] Hicks, N., *Notes on Differential Geometry*, Van Nostrand, Princeton, 1965.

[HIL] Hilbert, D., *Über Flächen von konstanten Gausscher Krümmung*, Trans. Amer. Math. Soc. **1** (1901), 87–99.

[H-CV] Hilbert, D., and Cohn-Vossen, S., *Geometry and the Imagination*, Chelsea, New York, 1956.

[HR] Hirsch, M. W., *Differential Topology*, Springer-Verlag, New York, 1976.

[H-Y] Hocking, J. G., and Young, G. S., *Topology*, Addison-Wesley, Reading, MA, 1961.

[HU] Hudson, J. F. P., *Piecewise Linear Topology*, Benjamin, Menlo Park, CA, 1969.

[HZ] Hurewicz, W., *Lectures on Ordinary Differential Equations*, M.I.T. Press, Cambridge, MA, 1966.

[H-W] Hurewicz, W., and Wallman, H., *Dimension Theory*, Princeton U. Press, Princeton, 1941.

[JA] Jänich, K., *Topology*, Springer-Verlag, New York, 1984.

[JU] Juster, N., *The Dot and the Line*, Random House, New York, 1963.

[KA] Kauffman, Louis H., *On Knots*, Ann. of Math. Studies, vol. 115, Princeton U. Press, Princeton, 1987.

[KE] Kelley, John L., *General Topology*, Van Nostrand, Princeton, 1955.

[KN] Kendig, K., *Elementary Algebraic Geometry*, Springer-Verlag, New York, 1977.

[KY] Kinsey, L. C., *Topology of Surfaces*, Springer-Verlag, New York, 1993.

[KI] Kirby, R., *Stable homeomorphisms and the annulus conjecture*, Ann. of Math. **89** (1969), 575–582.

[K-S] Kirby, R., and Siebenmann, L., *Foundational Essays on Topological Manifolds, Smoothings, and Triangulations*, Ann. of Math. Studies, vol. 88, Princeton U. Press, Princeton, 1977.

[KL] Klingenberg, W., *A Course in Differential Geometry*, Springer-Verlag, New York, 1978.

[K-N] Kobayashi, S., and Nomizu, K., *Foundations of Differential Geometry*, I, II, Interscience, New York, 1963, 1969.

[LA1] Lang, S., *Linear Algebra*, Addison-Wesley, Reading, MA, 1966.

[LA2] Lang, S., *Analysis I*, Addison-Wesley, Reading, MA, 1968.

[LY] Lyusternik, L. A., *Convex Figures and Polyhedra*, Dover, New York, 1963.

[MA] Malitz, J., *Introduction to Mathematical Logic*, Springer-Verlag, New York, 1979.

[MK] Markov, A. A., *Unsolvability of the problem of homeomorphism*, Proc. Int. Cong. Math., 1958, pp. 300–306 (Russian).

[MT] Martin, G., *The Foundations of Geometry and the Non-Euclidean Plane*, Springer-Verlag, New York, 1975.

[MS1] Massey, W. S., *Algebraic Topology: An Introduction*, Springer-Verlag, New York, 1967.

[MS2] ———, *Singular Homology Theory*, Springer-Verlag, New York, 1980.

[MCL] McCleary, J., *Geometry from a Differentiable Viewpoint*, Cambridge U. Press, Cambridge, 1994.

[MC] McMullen, P., *Non-linear angle-sum relations for polyhedral cones and polytopes*, Math. Proc. Cambridge Philos. Soc. **78** (1975), 247–261.

[M-P] Millman, R. S., and Parker, G. D., *Elements of Differential Topology*, Prentice-Hall, Englewood Cliffs, NJ, 1977.

[MI1] Milnor, J. W., *Differential Manifolds which are Homotopy Spheres*, unpublished notes.

[MI2] ———, *Morse Theory*, Ann. of Math. Studies, vol. 51, Princeton U. Press, Princeton, 1963.

[MI3] ———, *Topology from the Differentiable Viewpoint*, U. of Virginia Press, Charlottesville, VA, 1965.

[MI4] ———, *Analytic proofs of the "Hairy Ball Theorem" and the Brouwer Fixed Point Theorem*, Amer. Math. Monthly **85** (1978), 521–524.

[M-S] Milnor, J.W., and Stasheff, J., *Characteristic Classes*, Ann. of Math. Studies, vol. 76, Princeton U. Press, Princeton, NJ, 1974.

[MO] Moise, E. E., *Geometric Topology in Dimensions 2 and 3*, Springer-Verlag, New York, 1977.

[MR] Moore, G. H., *Zermelo's Axiom of Choice*, Springer-Verlag, New York, 1982.

[MG] Morgan, F., *Riemannian Geometry: A Beginner's Guide*, Jones and Bartlett, Boston, 1993.

[MU1] Munkres, J. R., *Elementary Differential Topology*, Ann. of Math. Studies, vol. 54, Princeton U. Press, Princeton, 1966.

[MU2] ———, *Topology, A First Course*, Prentice-Hall, Englewood Cliffs, NJ, 1975.

[MU3] ———, *Elements of Algebraic Topology*, Addison-Wesley, Menlo Park, CA, 1984.

[MU4] ———, *Analysis on Manifolds*, Addison-Wesley, Menlo Park, CA, 1991.

[NA] Naber, G., *Topological Methods in Euclidean Spaces*, Cambridge University Press, Cambridge, 1980.

[ON] O'Neill, B., *Elementary Differential Geometry*, Academic Press, New York, 1966.

[OS1] Osserman, R., *A Survey of Minimal Surfaces*, Van Nostrand Reinhold, New York, 1969.

[OS2] ———, *Curvature in the Eighties*, Amer. Math. Monthly **97** (1990), 731–756.

[PI] Pietsch, H., *Geodätische approximation einer topologischen triangulation*, Deutsche Math. **4** (1939), 583–589.

[RI] Richards, I., *On the classification of noncompact surfaces*, Trans. Amer. Math. Soc. **106** (1963), 259–269.

[RO] Rolfsen, D., *Knots and Links*, Publish or Perish, Inc., Berkeley, CA, 1976.

[RT] Rotman, J. J., *Theory of Groups*, 2nd ed., Allyn & Bacon, Boston, 1973.

[R-S] Rourke, C., and Sanderson, B., *Introduction to Piecewise-Linear Topology*, Ergebnesse der Mathematik 69, Springer-Verlag, New York, 1972.

[RU] Rucker, R., *The Fourth Dimension*, Houghton Mifflin, Boston, 1984.

[RD] Rudin, M. E., *An unshellable triangulation of a tetrahedron*, Bull. Amer. Math. Soc. **64** (1958), 90–91.

[SA] Samelson, H., *Orientability of hypersurfaces in* $\mathbb{R}^n$, Proc. Amer. Math. Soc. **22** (1969), 301–302.

[S-F] Senechal, M., and Fleck, G., *Shaping Space*, Birkhäuser, Boston, 1988.

[S-T] Singer, I. M., and Thorpe, J. A., *Lecture Notes on Elementary Topology and Geometry*, Springer-Verlag, New York, 1967.

[SP] Spanier, E., *Algebraic Topology*, McGraw-Hill, New York, 1966.

[SK1] Spivak, M., *Calculus on Manifolds*, Benjamin, New York, 1965.

[SK2] ———, *Calculus*, Benjamin, New York, 1967.

[SK3] ———, *A Comprehensive Introduction to Differential Geometry*, vols. I–V, Publish or Perish, Inc., Boston, 1975.

[ST] Stillwell, J., *Geometry of Surfaces*, Springer-Verlag, New York, 1992.

[SR] Struik, Dirk J., *Lectures on Classical Differential Geometry*, Addison-Wesley, Reading, MA, 1950.

[TH] Thomassen, C., *The Jordan–Schönflies theorem and the classification of surfaces*, Amer. Math. Monthly **99** (1992), 116–130.

[TR] Thorpe, J. A., *Elementary Topics in Differential Geometry*, Springer-Verlag, New York, 1979.

[TU] Trudeau, R. J., *The Non-Euclidean Revolution*, Birkhäuser, Boston, 1987.

[VA] Valentine, F. A., *Convex Sets*, McGraw-Hill, New York, 1964.

[WA] Wallace, A., *Differential Topology*, Benjamin/Cummings, Reading, MA, 1968.

[WR] Warner, F. W., *Foundations of Differentiable Manifolds and Lie Groups*, Springer-Verlag, New York, 1983.

[WE] Weeks, J. R., *The Shape of Space*, Marcel Dekker, New York, 1985.

[WH1] Whitehead, J. H. C., *Convex regions in the geometry of paths*, Quart. J. Math. **3** (1932), 33–42, 226–227.

[WH2] ———, *On* C^1*-complexes*, Ann. of Math. **41** (1940), 809–824.

[WH3] ———, *Manifolds with transverse fields in Euclidean space*, Ann. of Math. **73** (1961), 154–212.

[Y-K-K] Yemelichev, V. A., Kovalev, M. M., and Kravtsov, M. K., *Polytopes, Graphs and Optimisation*, Cambridge U. Press, Cambridge, 1984.

[YU] Yu, Y.-L., *Combinatorial Gauss–Bonnet–Chern formula*, Topology **22** (1983), 153–163.

[ZE] Zeeman, E. C., *Seminar on Combinatorial Topology*, unpublished notes, I.H.E.S., Paris, 1963.

Hints for Selected Exercises

Section 1.2

1.2.2. Take a well-chosen nested family of open intervals in $\mathbb{R}$.

1.2.12. The goal is to show that $U = \mathbb{R}^n - A$ is open in $\mathbb{R}^n$. Let $p \in U$ be a point; show that $O_{D/2}(p, \mathbb{R}^n)$ contains at most one member of A. Now find a number $r > 0$ such that $O_r(p, \mathbb{R}^n) \subset U$.

1.2.17. Let $x = \operatorname{lub} S$, and suppose that $x \notin S$; obtain a contradiction by showing that $\mathbb{R} - S$.

Section 1.3

1.3.7. Look for a function $f\colon (0, 1) \to \mathbb{R}$ with slope that goes to infinity.

1.3.8. Divide $\mathbb{R}$ into two parts, one on which $f(x) \geq g(x)$, and one on which $g(x) \geq f(x)$.

1.3.9. Use Condition (3) of Proposition 1.3.3. Let $p \neq 0$ be a real number; first, find numbers $m, \delta_1 > 0$ such that $m < |xp|$ for all $|x - p| < \delta$; now find the desired δ.

1.3.10. Using Lemmas A.5 and 1.3.8, and writing F out in coordinates with respect to the standard basis of $\mathbb{R}^n$, it suffices to show that any function $f\colon \mathbb{R}^n \to \mathbb{R}$ of the form $f\left(\begin{pmatrix} x_1 \\ \vdots \\ x_n \end{pmatrix}\right) = a_1x_1 + \cdots + a_nx_n + d$ (where $a_1, \ldots, a_n, d \in \mathbb{R}$ are any numbers) is continuous; the proof that f is continuous is similar to Example 1.3.4.

1.3.11. For each possible combination of continuous, open, and closed, an example is given, and sometimes a hint on how to prove that the desired properties are satisfied.

(1) Continuous, open and closed: The identity map $1_{\mathbb{R}}\colon \mathbb{R} \to \mathbb{R}$.

(2) Open, closed and not continuous: The map $f\colon [0, 2] \to [0, 1] \cup (2, 3]$ given by

$$f(x) = \begin{cases} x, & \text{if } x \in [0, 1]; \\ x + 1, & \text{if } x \in (1, 2]. \end{cases}$$

For non-continuity, consider the subset $U = (\frac{1}{2}, 1]$ of the codomain $[0, 1] \cup (2, 3]$. For openness, consider separately the open subset of $[0, 2]$ that contains the point 1 and those that do not. For closedness, use the bijectivity of f and Exercise 1.3.5.

(3) Continuous, closed and not open: Any constant map $f: \mathbb{R} \to \mathbb{R}$.

(4) Continuous, open and not closed: The projection map $\pi_1: \mathbb{R}^2 \to \mathbb{R}$, given by $\pi_1(\binom{x}{y}) = x$. For closedness, consider the subset $C \subset \mathbb{R}^2$ that is the sequence

$$C = \{\binom{1/2}{2}, \binom{1/3}{3}, \binom{1/4}{4}, \dots\}.$$

Use Exercise 1.2.12.

(5) Open, not continuous and not closed: The map $f: \mathbb{R}^2 \to \mathbb{R}$ given by

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

This is similar to part (4).

(6) Closed, not continuous and not open: The map $f: \mathbb{R} \to \mathbb{R}$ given by

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

For non-continuity look at $f^{-1}((\frac{1}{2}, \frac{3}{2}))$.

(7) Continuous, not open and not closed: The inclusion map $i: [0, 1) \to \mathbb{R}$.

(8) Not continous, not open and not closed: The map $f: \mathbb{R} \to \mathbb{R}$ given by

$$f(x) = \begin{cases} 2, & \text{if } x = 0; \\ x, & \text{if } x \neq 0. \end{cases}$$

For non-continuity look at $f^{-1}((1, 3))$. For non-openness look at $f((-1, 1))$. For non-closedness look at $f([0, 1])$.

Section 1.4

1.4.8. Use equivalence relations if you are familiar with the concept; otherwise, show directly that for any $x, y \in X$, if $[x]$ and $[y]$ have non-empty intersection then they are in fact equal sets.

Section 1.5

1.5.3. It suffices to prove that the product of two connected subsets of Euclidean space is connected; the result for products of more than two connected sets would then follow by induction on the number of factors in the product; let $A \subset \mathbb{R}^n$ and $B \subset \mathbb{R}^m$ be connected; choose a point $(a, b) \in A \times B$; for each $x \in A$ let $T_x = (A \times \{b\}) \cup (\{x\} \times B)$; show that each T_x is connected (use Exercise 1.5.2); observe that $A \times B = \bigcup_{x \in A} T_x$, and deduce that $A \times B$ is connected.

1.5.4. The tricky part is showing that components are closed subsets; let C be a component and let $x \in A - C$ be a point; conclude that $C \cup \{x\}$ is not connected, and deduce that there is some number $\epsilon > 0$ such that $O_\epsilon(x, A) \subset A - C$.

1.5.6. Use the Intermediate Value Theorem, though you need to decide what to apply it to.

1.5.8. For the first part, let $r > 0$ be a number such that $O_r(x, \mathbb{R}^2) \subset U$; if $y, z \in U$ are any two points, then by hypothesis there is a path in U from y to z; if the path does not contain x there is nothing to prove; if the path does contain x, then show how to modify the path inside $O_r(x, \mathbb{R}^2)$ so that it misses x.

1.5.11. Let $x, y \in U$ be points for which there is no path from x to y in $U - \{a\}$; by hypothesis there must be a path from x to y in U, and hence this path must contain a; use openness find points x' and y' with the same properties as x and y but contained in V; now suppose that $V - \{a\}$ is path connected, and obtain a contradiction.

1.5.13. Suppose to the contrary that there is a component of A that intersects both B and $A - B$.

Section 1.6

1.6.4. This is a tricky problem; one possibility is to let $A = \mathbb{R} \subset \mathbb{R}^2$, to define an injective continuous map $f: A \to \mathbb{R}^n$ with the desired properties, and to let $B = f(A)$.

1.6.7. Cover the set $[a, b] \times \{0\}$ with open squares (rather than open disks) such that the function f is positive on each square.

1.6.8. To find a maximal element, consider the collection of all sets of the form $(-\infty, a)$ for $a \in A$.

1.6.10. Use the Extreme Value Theorem to find a point $x \in (a, b)$ at which f has a maximal or minimal value and $f(x) \neq f(a)$; use the Intermediate Value Theorem to prove the desired result.

1.6.11. Consider the function $d: A \times B \to \mathbb{R}$ defined by $d(a, b) = \|a - b\|$.

1.6.12. Pick any finite subcover of A by compactness; let U_{i_1} be any element of the finite subcover that contains p; if U_{i_1} contains q we are finished, so assume otherwise; use connectivity to show that there is some set in the finite subcover that intersects U_{i_1}, and call this set U_{i_2}; keep going until one of these sets contains q.

Section 2.2

2.2.5. For part (i), use the Schönflies Theorem to show that $\mathbb{R}^2 - C$ has precisely two components; show that at least one of these components must be unbounded, since $\mathbb{R}^2$ is unbounded and C is compact; show that not both of the components are unbounded, again using the compactness of C; for part (ii) use the hint for Exercise 2.2.6.

2.2.6. First reduce to the case where $B_1 = D^2$; then, by definition, there is some homeomorphism $g: D^2 \to B$; observe that $g(S^1) = \partial B$; using the compactness of D^2 show that $[\mathbb{R}^2 - \partial B] - g(\text{int } D^2)$ is open in $\mathbb{R}^2 - \partial B$; use Invariance of Domain to show that $g(\text{int } D^2)$ is open in $\mathbb{R}^2 - \partial B$; use the connectivity of int D^2 and Exercise 1.5.13 to deduce that $g(\text{int } D^2)$ is precisely one of the two components of $\mathbb{R}^2 - \partial B$; use the compactness of int D^2 to show that $g(\text{int } D^2)$ is the bounded component of $\mathbb{R}^2 - \partial B$; use a similar (though simpler) argument to show that $h(\text{int } D^2)$ is also the bounded component of $\mathbb{R}^2 - \partial B$; deduce the desired result.

2.2.8. The set B cannot be just the origin, so pick some point $z \in B$ other than the origin; show that the intersection of B with the line containing O_2 and z is a compact set; use Exercise 1.5.8 to find a point x in this intersection with maximal distance from the origin; use Invariance of Domain to show that $x \in \partial B$; show that x is as desired; to show that there are at least two such points x, use Invariance of Domain to show that B is not contained in a single line.

2.2.9. Reduce to the situation where $B_1 = D^2$; use Exercise 2.2.8 to find two points $x, y \in \partial B_2$ satisfying the conclusion of that exercise; break up $D^2 - \text{int } B_2$

into two disks by forming two 1-spheres using the points x and y and then using Corollary 2.2.5; use these disks to construct a homeomorphism from A onto $D^2 - \text{int } B_2$.

2.2.10. First suppose that $n > m$, and derive a contradiction to Invariance of Domain by thinking of $\mathbb{R}^m$ as a subset of $\mathbb{R}^n$ as usual; next assume $n < m$, and that A is open in $\mathbb{R}^m$; let $x \in A$ be a point, so that there exists some number $\epsilon > 0$ such that $O_\epsilon(x, \mathbb{R}^m) \subset A$; hence $\mathbb{R}^n$ contains a subset homeomorphic to $O_\epsilon(x, \mathbb{R}^m)$; think of $\mathbb{R}^n$ as a subset of $\mathbb{R}^m$ as usual; derive a contradiction using Invariance of Domain.

2.2.12. First show that for each point in ∂J there is a point arbitrarily close to it in int J; then show that if a point is in int B then there is a minimal positive distance from it to all points in ∂B.

2.2.13. Let $A \subset S^1$ be a proper subset homeomorphic to S^1; show that S^1 with a point removed is homeomorphic to $\mathbb{R}$, and hence a homeomorphic copy of A sits in $\mathbb{R}$; using the compactness and connectivity of A show that A must be a closed interval; show that a closed interval cannot be homeomorphic to S^1, yielding a contradiction.

Section 2.3

2.3.3. Let $V \subset Q$ be an open set containing p that is homeomorphic to int $D2$, and let h: int $D2 \to V$ be a homeomorphism; consider the set $h^{-1}(O_\epsilon(p, Q))$, and find appropriate disks there.

2.3.4. Let $p \in U$ be a point, and let $V \subset Q$ be an open subset containing Q that is homeomorphic to int $D2$; consider $U \cap V$, and use Invariance of Domain.

2.3.9. Let H: $Q_1 - \text{int } B_1 \to Q_2 - \text{int } B_2$ be a homeomorphism; show that $H(\partial B_1) = \partial B_2$; extend H over B_1 so that $H(B_1) = B_2$.

Section 2.4

2.4.2. The "obvious" thing you might try, namely cutting the rectangle in Figure 2.4.9 (i) in two with a horizontal line half-way up, does not work; a more judicious cut is needed.

Section 2.5

2.5.3. In the case of planes, let $p \in Q$ be a point, and let Π be a plane in $\mathbb{R}^n$ which contains an open neighborhood of p in Q; consider the set $T = \Pi \cap Q$; use the definition of relative closedness to show that T is closed in Q; use the hypothesis on Q to show that T is open in Q, the crucial observation being that a non-empty open subset of a plane in $\mathbb{R}^n$ cannot simultaneously be an open subset of a different plane in $\mathbb{R}^n$; conclude that $T = Q$.

Section 2.6

2.6.1. Use Lemma 2.4.5.

Appendix A2.2

A2.2.1. It suffices to show that the surfaces Q and Q_r in the proof of parts (i) and (ii) of the proposition are homeomorphic under the hypothesis that at least one of Q_1 and Q_2 is disk-reversible; assume without loss of generality that Q_2 is disk-reversible; the goal is to apply Exercise 1.4.9 to the maps f' and $r' \circ f'$; let $H\colon Q_2 \to Q_2$ be a homeomorphism such that $H(T_2') = T_2'$ and $H|\partial T_2'$ is an orientation reversing homeomorphism, which exists by Exercise 2.5.3; consider the map

$$d = r' \circ f' \circ (f')^{-1} \circ (H|\partial T_2')^{-1}\colon \partial T_2' \to \partial T_2',$$

and proceed as in the proof of parts (i) and (ii) of the proposition.

A2.2.3. Transfer everything to D^2 using Corollary 2.2.6 and Exercise 2.2.6, and use the method of Lemma A2.2.2.

A2.2.5. Find a homeomorphism $g\colon S^1 \to S^1$ such that $f_2 = f_1 \circ g$; then use Exercise A2.2.4.

A2.2.6. Pick any pair of antipodal points $a, b \in S^1$; define the function $f\colon \overrightarrow{ab} \to \mathbb{R}$ by letting $f(z)$ be the length of the arc from $f(-z)$ to $-f(z)$, where the length is positive if the arc is counterclockwise and negative if the arc is clockwise (note that by injectivity $f(-z) \neq f(z)$, so that there is never ambiguity in this definition); if $f(a)$ and $f(b)$ are antipodal there is nothing to prove, so assume otherwise; show that one of $f(a)$ and $f(b)$ is positive and the

other is negative; since $\overrightarrow{ab}$ is an arc, the Intermediate Value Theorem (Theorem 1.5.4) can be applied to f; deduce the result.

A2.2.7. There are many ways to construct the map F, the simplest being as follows: The idea is to map each horizontal slice of $[-1, 1] \times [0, 1]$ homeomorphically to itself, squeezing the homeomorphism f to a point as one moves up $[0, 1]$, fixing the endpoints and increasing amounts of $[-1, 1]$; graphically this map F is suggested in Figure H.1; find an explicit formula.

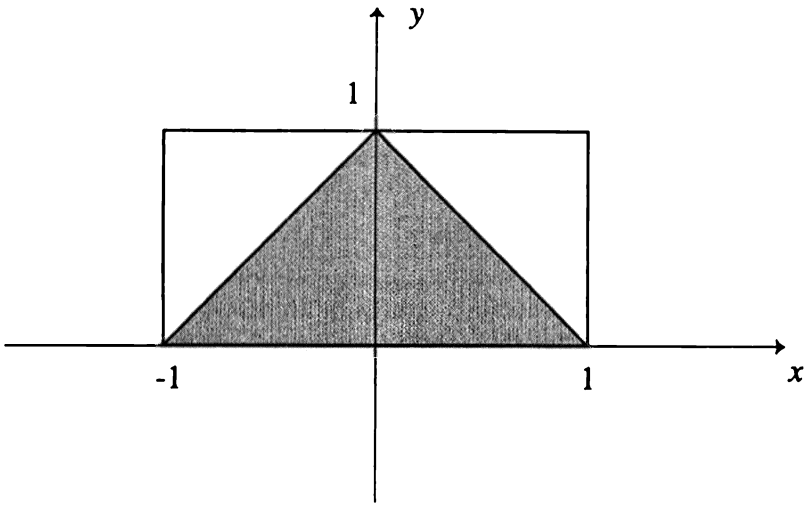


Figure H.1

A2.2.10. Use Proposition A2.2.6.

A2.2.11. Use Proposition A2.2.6.

Section 3.2

3.2.5. Linear independence.

3.2.6. Suppose $\eta = \langle a_0, \ldots, a_k \rangle$; translate the whole situation by a_0, and consider the issue of k-dimensional subspaces of $\mathbb{R}^n$ containing k given vectors.

3.2.8. Use Lemmas A.6 and A.7 and Exercise 1.3.10.

Section 3.3

3.3.6. The tricky part is (3) $\Rightarrow$ (2); suppose (2) holds but (3) is false, so that there exist simplices σ and τ of K for which there does not exist a connecting chain of simplices as in condition (3); let $C \subset |K|$ be the union of all simplices of K that can be connected to σ by a chain of simplices, and let D be the union of the rest of the simplices of K; show that $C \cup D = |K|$, that $C \cap D = \emptyset$ and that both C and D are non-empty and closed in $|K|$.

3.3.9. Although the idea is intuitively very straightforward — one simply chops up the polygonal disk into triangles — the proof is a bit trickier (though not longer) than might be expected; see [HO] for a proof and discussion of some false proofs that have been published.

Section 3.4

3.4.3. No need to repeat the proof of Theorem 3.4.1; use Exercise 3.3.10, although you need to check what happens with the links after subdivision.

Section 3.5

3.5.2. Use the method of Example 3.5.3 (2).

3.5.4. First show that no two of S^2, T^2, $T^2 \# T^2$, $T^2 \# T^2 \# T^2, \ldots$ are homeomorphic by using the Euler characteristic; then show that no two of P^2, $P^2 \# P^2$, $P^2 \# P^2 \# P^2, \ldots$ are homeomorphic; use Proposition 2.6.6 to show that no surface from one of these lists is homeomorphic to a surface in the other list.

Section 3.7

3.7.2. Break up $f_2(K)$ into parts.

3.7.5. Equation 3.7.1.

3.7.6. Let K be a simplicial surface with $f_0(K) = V_\zeta$; create new simplicial surfaces with $f_0(K)$ any integer greater than V_ζ by subdividing K, adding one 0-simplex at a time. Observe that if ζ and $f_0(K)$ are known, then $f_1(K)$ and $f_2(K)$ are determined.

Section 3.8

3.8.1. (i). Let m be the dimension of K, let τ be an m-simplex of K and let $x \in \operatorname{Int} \tau$ is a point; show that x can be chosen to be in $\operatorname{int}|K|$; show that $m = 2$, as in the proof of Theorem 3.4.1.

(ii) & (iii). Use the fact that $\mathbb{R}^2 \not\approx \mathbb{H}^2$ (Exercise 2.2.11), and that small enough open balls in $|K|$ about any two points in the interior of the same simplex of K are homeomorphic to deduce that the interior of each simplex of K is entirely contained either in $\operatorname{int}|K|$ or $\partial|K|$, as in the proof of Theorem 3.4.1; show that every 1-simplex of K, the interior of which is contained in $\operatorname{int}|K|$, is a face of two 2-simplices, and the underlying space of the link of every 0-simplex of K contained in $\operatorname{int}|K|$ is a 1-sphere.

Let η be a 1-simplex of K such that $\operatorname{Int}\eta \subset \partial|K|$; using an argument similar to that used in the proof of Theorem 3.4.1, show that η is the face of at least one 2-simplex of K; suppose η is contained in 2-simplices $\sigma_1, \ldots, \sigma_p$, where $p \geq 2$, let $x \in \operatorname{Int}\eta$ be a point and let $U \subset |K|$ be an open set containing x that is homeomorphic to $\mathbb{H}^2$; deduce that any point in $U \cap \partial|K|$ has an open neighborhood in $|K|$ that is homeomorphic to $\mathbb{H}^2$; by an argument to Exercise 2.3.3, choose U small enough so that it is entirely contained in $\operatorname{Int}\eta \cup \operatorname{Int}\sigma_1 \cup \cdots \cup \operatorname{Int}\sigma_p$; show that $U \cap \partial|K| \subset \operatorname{Int}\eta$; use Exercise 1.5.12 to show that $U \cap \operatorname{int}|K|$ is entirely contained in one of the $\operatorname{Int}\sigma_i$; deduce that U is entirely contained in σ_i, and derive a contradiction using Exercise 1.2.18 (ii) and an argument similar to that found in the proof of Theorem 3.4.1; deduce that η is contained in precisely one 2-simplex.

If w is a 0-simplex of K contained in $\partial|K|$, show that $|\operatorname{link}(w, K)|$ is an arc, as in the proof of Theorem 3.4.1; deduce that parts (ii) and (iii) of the theorem will both follow, except for the fact that $\operatorname{Bd} K$ is a subcomplex; deduce this remaining claim from Exercise 2.2.12.

3.8.2. Use induction on p.

3.8.5. Combine Corollary 2.2.5 (ii), Corollary A2.2.5 and Exercise 3.8.4.

Section 4.2

4.2.1. The proof is very similar to the proof of Theorem 4.2.2.

4.2.2. The proof is similar to the proofs of Proposition 5.2.5 (ii) and Lemma 7.2.2 (iii), using Exercise 4.2.1 instead of Theorem 4.2.2.

4.2.3. Use the chain rule.

4.2.5. One first needs to show that the image of c intersects each vertical line in $\mathbb{R}^2$ at most once; let c_1 and c_2 denote the x and y coordinate functions of c; show that c_1 is bijective and that its image is an open interval; then use Exercise 4.2.3 to conclude that c_1 is a diffeomorphism; then consider the function $c_2 \circ (c_1)^{-1}$.

Section 4.3

4.3.3. First show that if $h(d, e)(a, b)$ is any diffeomorphism (independent of c) then $h(t) \neq 0$ for all $t \in (d, e)$; hence either $h'(t) > 0$ for all $t \in (d, e)$ or $h'(t) < 0$ for all $t \in (d, e)$; now apply Proposition 4.3.4 (i) to c; if the derivative of h has the wrong sign, modify h.

4.3.10. First find $\epsilon_p > 0$ such that $c|(p - \epsilon_p, q]$ is injective as follows; use Exercise 4.2.5 to find a number $\epsilon_1 > 0$ such that $c|(p - \epsilon_1, p + \epsilon_1)$ is injective; by compactness find the minimal distance $D > 0$ from $c([p + \epsilon_1, q])$ to $c(p)$; find a number $\delta > 0$ be such that $c((p - \delta, p + \delta)) \subset O_{D/2}(c(p), \mathbb{R}^3)$; show that $\epsilon_p = \min\{\epsilon_1, \delta\}$ has the desired property; now use a similar argument to find a number $\epsilon_q > 0$ such that $c|[p - \epsilon_p/2, q + \epsilon_q)$ is injective; let $\epsilon = \frac{1}{2}\min\{\epsilon_p/2, \epsilon_q/2\}$; now use Proposition 1.6.14 (iii) and the analog for arc of Exercise 2.2.4 applied to $[p - \epsilon, q + \epsilon]$.

Section 4.4

4.4.2. First consider the case where the curve lies in the x-y plane, and then reduce the general case to the first case using rotations and translations of $\mathbb{R}^3$.

4.4.4. Use Exercise 4.4.2.

Section 4.5

4.5.2. Rotation matrices preserve inner product and cross product.

4.5.4. Observe that $c(t) = \int_p^t T(s)\, ds + c(p)$ for any fixed $p \in (a, b)$.

4.5.5. For the "if" part, for each $t \in (a, b)$ we have $c(t) - x_0 = \lambda(t)\, T(t)$ for some function $\lambda\colon (a, b) \to \mathbb{R}$; what can you say about $\kappa(t)$ in this case?

Section 4.6

4.6.1. Use a circle and a right circular helix or appropriate radii.

4.6.2. For step (1), use Equations 4.6.4 and 4.6.5; for step (3) use Theorem 4.2.6.

Section 4.7

4.7.4. Use Exercise 4.7.3.

4.7.5. Define $f\colon (a, b) \to \mathbb{R}$ to be $f(t) = \langle c(t), c(t)\rangle$; what can you say about $f'(q)$ and $f''(q)$? Use this information to show that $\bar{N}(q) = \pm c(q)/R$, and that $|\kappa(q)| \geq \frac{1}{R}$.

Section 5.2

5.2.6. Use the chain rule for partial derivatives.

5.2.8. Use Proposition 5.2.5 (ii).

Section 5.3

5.3.5. The c curve is in the x-z plane; it might help to sketch the surface.

5.3.6. Use the x-y plane as the surface, though the equation $z = 0$ for this surface does not do what we want.

Section 5.4

5.4.6. If $c\colon (-\epsilon, \epsilon) \to M$ is a curve in M, use the chain rule to show that $DF(c(t))$ is perpendicular to $c'(t)$ for all $t \in (-\epsilon, \epsilon)$.

5.4.7. For the case $k = 0$ compute the partial derivatives of $\langle x, n\rangle$ with respect to s and t; for the case $k \neq 0$ compute $\|x - (-\frac{1}{k}w\|$.

Section 5.5

5.5.9. For step (2), show that

$$\frac{\partial g_{jk}}{\partial u_i} = \langle x_{ij}, x_k \rangle + \langle x_{ik}, x_j \rangle$$
$$\frac{\partial g_{ik}}{\partial u_j} = \langle x_{ij}, x_k \rangle + \langle x_{jk}, x_i \rangle$$

by permuting the subscripts i, j and k in the equation found in Step (1) and using the equality of mixed partial derivatives. Now solve these two equations together with the equation in Step (1).

Section 5.6

5.6.4. In both cases, choose a curve $c\colon (-\epsilon, \epsilon) \to M$ such that $c(0) = p$ and $c'(0) = v$; then use the chain rule on $f \circ c$.

Section 5.7

5.7.1. For part (i), at each point $\bar{q} \in U$ one can find numbers $Z^1(\bar{q})$ and $Z^2(\bar{q})$ such that $Z \circ x(\bar{q}) = Z^1(\bar{q})x_1(\bar{q}) + Z^2(\bar{q})x_2(\bar{q})$ by using linear algebra; to show that the resulting functions Z^i are smooth, use Cramer's rule.

5.7.2. Assume without loss of generality that $\bar{p} = O_2$; if $\{e_1, e_2\}$ are the standard basis vectors for $\mathbb{R}^2$, use the curve $c\colon (-\epsilon, \epsilon) \to M$ given by $c(t) = x(te_i)$ to compute $\widetilde{\nabla}_{x_i(\bar{p})} Z$.

5.7.8. Choose a coordinate patch, and do everything in coordinates; start off by stating and proving an analog of Exercise 5.2.6 for functions of two variables.

Section 5.9

5.9.2. One scheme is to show (1) $\Leftrightarrow$ (2) and (3) $\Rightarrow$ (4) $\Rightarrow$ (5) $\Rightarrow$ (2); for (2) $\Rightarrow$ (1) show that for each point $p \in M$ there is an open subset $V \subset M$ containing p such that $f(V)$ is open in N and $f|V\colon V \to f(V)$ is a diffeomorphism; use Equation 5.9.1 to show that df_p is non-singular, and then use Proposition

5.9.3; for (4) $\Rightarrow$ (5) let $x: U \to M$ be a coordinate patch as in part (4); use Lemma 5.8.1 to show that letting $A = U$ works; for (5) $\Rightarrow$ (2) let $v \in T_pM$ be a vector; suppose that $c: (-\epsilon, \epsilon) \to M$ be a smooth curve such that $c(0) = p$ and $c'(0) = v$; for each $t \in (-\epsilon/2, \epsilon/2)$ show that

$$\int_{-\epsilon/2}^{t} \|c'(s)\|\, ds = \int_{-\epsilon/2}^{t} \|(f \circ c)'(s)\|\, ds;$$

deduce that $\|c'(t)\| = \|(f \circ c)'(t)\|$ for all $t \in (-\epsilon/2, \epsilon/2)$; conclude that $\|df_p(v)\| = \|v\|$; now use the fact that a linear map that preserves lengths of vectors also preserves inner products (see [LA1, chapter VIII §5]).

5.9.5. First show that Ψ is bijective; use the Inverse Function Theorem and Exercise 1.4.4 to show that Ψ is actually a homeomorphism.

5.9.9. Let $y: U \to M$ be a coordinate patch such that $p \in y(U)$; let $\bar{v} = (dy_p)^{-1}(v)$ and similarly for $\bar{w}$; let $F: \mathbb{R}^2 \to \mathbb{R}^2$ be the linear map that sends $\binom{1}{0}$ to $\bar{v}$ and $\binom{0}{1}$ to $\bar{w}$; consider the map $x = y \circ F|F^{-1}(U)$; use Exercises 5.9.7 and 5.9.8.

Section 6.1

6.1.1. For part (2), start with any coordinate patch whose image contains p, and construct the monge patch from it.

Section 6.2

6.2.1. By symmetry it suffices to describe the Weingarten map at any one point in the cylinder; examine the effect of the Weingarten map on a well-chosen basis for the tangent plane at the point you choose.

7. Use the normal vector field $n = \frac{Z}{\|Z\|}$ to compute the Weingarten map.

Section 6.3

6.3.1. By symmetry, both kinds of curvature are constant.

6.3.2. One need not completely know the Weingarten map to compute Gaussian curvature in this case. What can be said about the normal vectors along each ruling?

6.3.5. For part (1) use Exercises 5.4.6 and 6.2.7.

Section 6.4

6.4.5. Use the Weingarten Equations.

6.4.6. First show that the L_{ij} are all zero; then show that n is constant, and apply Exercise 5.4.7.

6.4.7. To show that the Weingarten map is zero at all points use $K = H = 0$ to compute the principle curvatures at all points; the proof of Exercise 6.4.6 can now be used to show that if $x: U \to M$ is a coordinate patch with connected domain, then $x(U)$ is contained in a plane; now use Exercise 2.5.3.

6.4.8. If the statements of each of (i)–(iii) can be proved for the image of each coordinate patch $x: U \to M$ for which U is connected, then the result for all of M can be pieced together using Exercise 2.5.3; using Exercise 6.3.4 (ii) there must be a function $d: U \to \mathbb{R}$ such that $n_i(\bar{p}) = d(\bar{p})x_i(\bar{p})$ for all $\bar{p} \in U$ and $i = 1, 2$; show that the function d is smooth; show that $d_1(\bar{p})x_2(\bar{p}) = d_2(\bar{p})x_1(\bar{p})$ for all $\bar{p}$, where d_1 and d_2 denote the partial derivatives of d; deduce that d_1 and d_2 are constantly zero, and it follows that d is constant; now show that x satisfies the hypotheses, and hence the conclusion, of Exercise 5.4.7.

6.4.9. First find a general criterion for a vector being an eigenvector for a 2×2 matrix.

Section 6.5

6.5.1. For step (1), use Equation 6.5.6 and Exercise 5.7.4. For step (2), use step (1) together with Equation 6.4.1 and some manipulating of the expression we are trying to derive.

Section 7.2

7.2.4. Use Exercises 5.5.6 and 7.2.3.

7.2.6. For step (3), there must be numbers $p, q \in (x, y]$ such that $h^{-1} \circ c(p) = 0$ and $h^{-1} \circ c(q) = 1$; show that there must be some number u between p and q such that $c(u) = c(x)$, a contradiction to injectivity.

7.2.7. For part (iii) use Exercise 7.2.5.

Section 8.2

8.2.2. Write $\bar{p} = \begin{pmatrix} p_1 \\ p_2 \end{pmatrix}$ and $\bar{q} = \begin{pmatrix} q_1 \\ q_2 \end{pmatrix}$, and consider the matrix Df; use Exercises 4.2.2 and 4.2.3.

8.2.5. Some parts of the corollary follow from the compactness of M and the statement of Proposition 8.2.3, whereas other parts require looking at the proof of Proposition 8.2.3.

8.2.6. Use the Lebesgue Covering Lemma (Theorem 1.6.9).

Section 8.3

8.3.1. Use Exercise 6.5.2.

8.3.2. Define $\sqrt{G(\begin{pmatrix} 0 \\ \theta \end{pmatrix})}, \ldots, \frac{\partial^3 \sqrt{G}}{\partial R^3}(\begin{pmatrix} 0 \\ \theta \end{pmatrix})$ using the limiting values given in Lemma 8.3.3, and use the Taylor polynomial with remainder.

8.3.4. For the second part of the exercise suppose that both cases are false; then there is some closed interval $I \subset (a, b)$, where I may be a single point, such that I is a maximal set upon which the curve $(D_p)^{-1} \circ c$ has horizontal tangent vectors at all points in I; see Figure H.2; consider the images under D_p of lines of the form $\theta = k$ and the curve $(D_p)^{-1} \circ c$, and obtain a contradiction to Theorem 7.2.6 by looking at one of the endpoints of I.

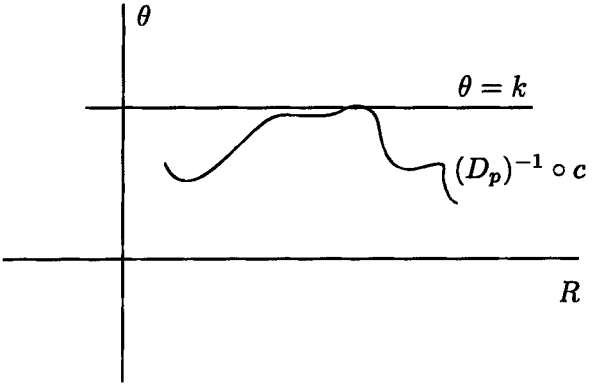


Figure H.2

8.3.5. The strategy is similar to the analogous part of the proof of Proposition 8.3.6; use Equation 8.3.6 to extend the function $\sqrt{G(\binom{R}{0})}$ smoothly over $(-\delta_p, \delta_p) \times (-\pi, 3\pi)$; take the derivative and restrict to an appropriate compact set.

Appendix A8.1

A8.1.1. If $\bar{q} = \binom{q_1}{q_2} \in O_{\epsilon_1}(\bar{p}, U)$ and $\bar{v} = \binom{v_1}{v_2} \in O_{\epsilon\epsilon_2/2}(O_2, \mathbb{R}^2)$ are any two points, then show that

$$D\Phi(p_0) = \begin{pmatrix} \frac{\partial q_1}{\partial q_1} & \frac{\partial q_1}{\partial q_2} & \frac{\partial q_1}{\partial v_1} & \frac{\partial q_1}{\partial v_1} \\ \frac{\partial q_2}{\partial q_1} & \frac{\partial q_2}{\partial q_2} & \frac{\partial q_2}{\partial v_1} & \frac{\partial q_2}{\partial v_1} \\ \frac{\partial c_1(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial q_1} & \frac{\partial c_1(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial q_2} & \frac{\partial c_1(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial v_1} & \frac{\partial c_1(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial v_1} \\ \frac{\partial c_2(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial q_1} & \frac{\partial c_2(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial q_2} & \frac{\partial c_2(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial v_1} & \frac{\partial c_2(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial v_1} \end{pmatrix},$$

where everything is evaluated at $p_0 = \begin{pmatrix} p_1 \\ p_2 \\ 0 \\ 0 \end{pmatrix}$. The first two rows of this matrix are straightforward to compute. We discuss two of the other entries in the matrix; the remaining ones are similar to these two.

To compute $\frac{\partial c_1(\epsilon/2, \bar{q}, 2\bar{v}/\epsilon)}{\partial q_1}|_{p_0}$, hold all the variables other than q_1 constant (setting $q_2 = p_2$ and $v_1 = v_2 = 0$) and then taking the derivative; observe that $c_1(\frac{\epsilon}{2}, \binom{q_1}{p_2}, O_2) = q_1$, since any geodesic of the form $c_{\bar{q},O_2}(t)$ is the constant map $c_{\bar{q},O_2}(t) = x(q)$ for all t by the uniqueness of geodesics.

To compute $\frac{\partial c_1(\epsilon/2, \bar{q}, 2\bar{v}/\epsilon)}{\partial v_1}|_{p_0}$, fix $q_1 = p_1$ and $q_2 = p_2$ and $v_2 = 0$, varying v_1; for notational ease let $t = v_1$, and take the derivative with respect to t at $t = 0$. Show that

$$\frac{\partial c_1(\frac{\epsilon}{2}, \bar{q}, \frac{2\bar{v}}{\epsilon})}{\partial v_1}|_{p_0} = \frac{d}{dt} c_1(\frac{\epsilon}{2}, p, t\frac{2}{\epsilon}\begin{pmatrix} 1 \\ 0 \end{pmatrix})|_{t=0}.$$

Use Exercise 7.2.5 to show that $c_{\bar{q},\bar{v}}(ta) = c_{\bar{q},t\bar{v}}(a)$ for any number a and all sufficiently small values of t, and use Exercise 5.9.7 to show that $c'_{\bar{p},\bar{v}}(0) = dx_{\bar{p}}(\bar{v}) = v_1 x_1(\bar{p}) + v_2 x_2(\bar{p})$, and hence $\frac{dc_i(s,\bar{p},\bar{v})}{ds}|_{s=0} = v_i$ for $i = 1, 2$. Put these observations together to obtain the desired value.

A8.1.2. Suppose the result is false; it must then be the case that A intersects $GS_d(p, M)$; use the ideas of the proof of the length minimization part of Theorem A8.1.2 to show that, in fact, the length of A must be greater than d, a contradiction.

A8.1.3. First, show that $c([s_0 - \eta, s_0 + \eta])$ must have minimal length of all regular arcs in M with its endpoints; assume the contrary, and deduce that $c([a, b])$ could not be the regular arc of minimal length with endpoints p and q. Use the uniqueness in Theorem A8.1.2 (i) to deduce that $c([s_0 - \eta, s_0 + \eta])$ is a geodesic arc.

Appendix A8.2

A8.2.1. Use the Schönflies Theorem.

Index

$\text{Area}(S)$	area of a region in a surface, 5.8
df_p	differential of f at p, 5.9
$Tx(U)$	tangent space over $x(U)$, 5.9
$\hat{n}$	Gauss map, 6.1
$\text{Area}_o(\hat{n}(T))$	oriented area of the image of the Gauss map, 6.1
L	Weingarten map, 6.2
II_p	second fundamental form at p, 6.2
II	second fundamental form, 6.2
(L_{ij})	matrix for Weingarten map, 6.2
(l_{ij})	matrix for second fundamental form, 6.2
n_i	partial derivative of n, 6.2
$K(p)$	Gaussian curvature at p, 6.3
$H(p)$	mean curvature at p, 6.3
k_1, k_2	principle curvatures, 6.3
Ω_v	oriented plane generated by v, 6.3
v_Ω	unit vector generated by Ω, 6.3
E, F, G	metric coefficients, 6.4
A, B, C	matrix entries for the second fundamental form, 6.4
ρ_v	bound for the domain of a geodesic, 7.2
$\Lambda(t)$	length of variation of a curve, 7.3
E_p	domain of exponential map, 8.2
$\exp_p$	exponential map, 8.2
δ_p	radius of ball in the domain for the exponential map, 8.2
δ_M	radius of ball in the domain for the exponential map, 8.2
ϵ_M	radius of ball in the image of the exponential map, 8.2
Υ_p	orthogonal map, 8.2
Exp_p	exponential coordinate patch, 8.2
$GS_r(p, M)$	geodesic circle, 8.2
$GO_r(p, M)$	open geodesic ball, 8.2
$\overline{GO}_r(p, M)$	closed geodesic ball, 8.2
rect	polar to rectangular map, 8.3
D_p	geodesic coordinate patch, 8.3
$(D_p)_i$	partial derivative of the geodesic coordinate patch, 8.3
$\overline{E}, \overline{F}, \overline{G}$	metric coefficients of Exp_p, 8.3
$\gamma_\theta(s)$	geodesic ray, 8.3
$\alpha_R(t)$	geodesic circle, 8.3
L_r	length of geodesic circle of radius r, 8.3
γ_p	radius of convex geodesic ball, A8.2

Index

Ethan D. Bloch

A First Course in Geometric Topology and Differential Geometry

The uniqueness of this text in combining geometric topology and differential geometry lies in its unifying thread: the notion of a surface. With numerous illustrations, exercises and examples, the student comes to understand the relationship of the modern abstract approach to geometric intuition. The text is kept at a concrete level, avoiding unnecessary abstractions, yet never sacrificing mathematical rigor. The book includes topics not usually found in a single book at this level.

A number of intuitively appealing definitions and theorems concerning surfaces in the topological, polyhedral and smooth cases are presented from the geometric view. Point set topology is restricted to subsets of Euclidean spaces. The treatment of differential geometry is classical, dealing with surfaces in $\mathbf{R}^3$. Included are the classification of compact surfaces, the Gauss-Bonnet Theorem and the geodesic nature of length minimizing curves on surfaces.

The material here should be accessible to math majors at the junior/senior level in an American university or college, the minimal prerequisites being standard Calculus sequence (including multivariable Calculus and an acquaintance with differential equations), linear algebra (including inner products), and familiarity with proofs and the basics of sets and functions.

ISBN 0-8176-3840-7