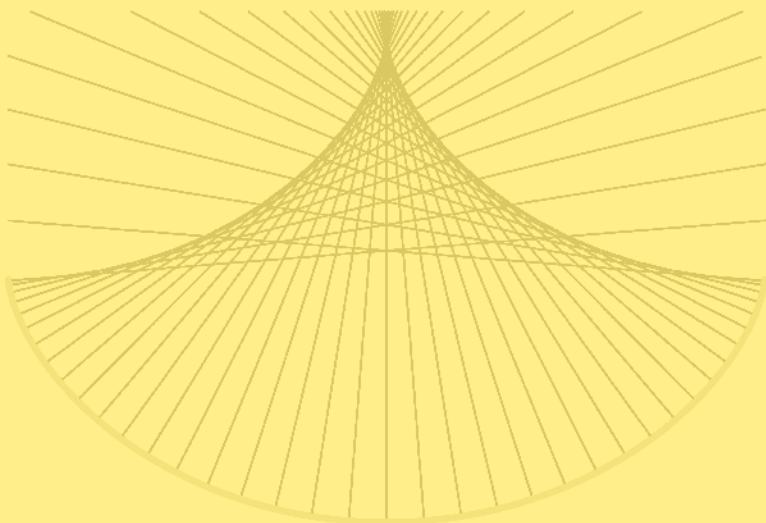


Undergraduate Texts in Mathematics

UTM

Kristopher Tapp

Differential Geometry of Curves and Surfaces



 Springer

Undergraduate Texts in Mathematics

Undergraduate Texts in Mathematics

Series Editors:

Sheldon Axler

San Francisco State University, San Francisco, CA, USA

Kenneth Ribet

University of California, Berkeley, CA, USA

Advisory Board:

Colin Adams, *Williams College*

David A. Cox, *Amherst College*

Pamela Gorkin, *Bucknell University*

Roger E. Howe, *Yale University*

Michael E. Orrison, *Harvey Mudd College*

Lisette G. de Pillis, *Harvey Mudd College*

Jill Pipher, *Brown University*

Fadil Santosa, *University of Minnesota*

Undergraduate Texts in Mathematics are generally aimed at third- and fourth-year undergraduate mathematics students at North American universities. These texts strive to provide students and teachers with new perspectives and novel approaches. The books include motivation that guides the reader to an appreciation of interrelations among different aspects of the subject. They feature examples that illustrate key concepts as well as exercises that strengthen understanding.

More information about this series at <http://www.springer.com/series/666>

Kristopher Tapp

Differential Geometry of Curves and Surfaces



Springer

Kristopher Tapp
Department of Mathematics
Saint Joseph's University
Philadelphia, PA, USA

ISSN 0172-6056 ISSN 2197-5604 (electronic)
Undergraduate Texts in Mathematics
ISBN 978-3-319-39798-6 ISBN 978-3-319-39799-3 (eBook)
DOI 10.1007/978-3-319-39799-3

Library of Congress Control Number: 2016942561

© Springer International Publishing Switzerland 2016

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

Contents

| | |
|---|---------|
| Introduction | VII |
| About Differential Geometry | VII |
| About This Book | VII |
| Prerequisites | VIII |
| Chapter 1. Curves | 1 |
| 1. Parametrized Curves | 2 |
| 2. The Inner Product (<i>Linear Algebra Background</i>) | 9 |
| 3. Acceleration | 16 |
| 4. Reparametrization | 18 |
| 5. Curvature | 24 |
| 6. Plane Curves | 32 |
| 7. Space Curves | 41 |
| 8. Rigid Motions | 48 |
| 9. Overview of Curvature Formulas | 60 |
| Chapter 2. Additional Topics in Curves | 61 |
| 1. Theorems of Hopf and Jordan | 61 |
| 2. Convexity and the Four Vertex Theorem (<i>Optional</i>) | 72 |
| 3. Fenchel's Theorem (<i>Optional</i>) | 78 |
| 4. Green's Theorem (<i>Calculus Background</i>) | 81 |
| 5. The Isoperimetric Inequality (<i>Optional</i>) | 97 |
| 6. Huygens's Tautochrone Clock (<i>Optional</i>) | 101 |
| Chapter 3. Surfaces | 113 |
| 1. The Derivative of a Function from \mathbb{R}^m to \mathbb{R}^n | 113 |
| 2. Regular Surfaces | 125 |
| 3. Tangent Planes | 141 |
| 4. Area Distortion and Orientation (<i>Linear Algebra Background</i>) | 147 |
| 5. Orientable Surfaces | 151 |

| | |
|---|---------|
| 6. Surface Area | 160 |
| 7. Isometries and the First Fundamental Form | 165 |
| 8. Equiareal and Conformal Maps (<i>Optional</i>) | 170 |
| 9. The First Fundamental Form in Local Coordinates | 182 |
| 10. An Alternative Characterization of Regular Surfaces <i>(Optional)</i> | 188 |
| Chapter 4. The Curvature of a Surface | 193 |
| 1. The Gauss Map | 195 |
| 2. Self-Adjoint Linear Transformations <i>(Linear Algebra Background)</i> | 201 |
| 3. Normal Curvature | 206 |
| 4. Geometric Characterizations of Gaussian Curvature | 213 |
| 5. The Second Fundamental Form in Local Coordinates | 217 |
| 6. Minimal Surfaces (<i>Optional</i>) | 227 |
| 7. The Fary–Milnor Theorem (<i>Optional</i>) | 237 |
| Chapter 5. Geodesics | 247 |
| 1. Definition and Examples of Geodesics | 247 |
| 2. The Exponential Map | 257 |
| 3. Gauss’s Remarkable Theorem | 268 |
| 4. Complete Surfaces | 275 |
| 5. Parallel Transport and the Covariant Derivative | 280 |
| 6. Geodesics in Local Coordinates | 289 |
| 7. Gaussian Curvature Measures Infinitesimal Holonomy | 298 |
| 8. Arc-Length Variation: Tire Tracks on a Curved Surface <i>(Optional)</i> | 303 |
| 9. Jacobi Fields (<i>Optional</i>) | 309 |
| Chapter 6. The Gauss–Bonnet Theorem | 319 |
| 1. The Local Gauss–Bonnet Theorem | 320 |
| 2. The Global Gauss–Bonnet Theorem | 326 |
| 3. Compact Surfaces | 334 |
| 4. A Sampling of Other Global Theorems | 342 |
| Appendix A. The Topology of Subsets of \mathbb{R}^n | 345 |
| 1. Open and Closed Sets and Limit Points | 345 |
| 2. Continuity | 350 |
| 3. Connected and Path-Connected Sets | 352 |
| 4. Compact Sets | 353 |
| Recommended Excursions | 357 |
| Image Credits | 359 |
| Index | 363 |

Introduction

About Differential Geometry

Differential geometry uses calculus to study curved shapes, such as the trajectory of a missile, the shape of an airplane wing, and the curvature of Einstein's spacetime. The story begins with characters familiar from multivariable calculus: parametrized curves and surfaces. Along the way, techniques from calculus, linear algebra, and real analysis are blended together, often within a single proof. All of the material is visually grounded with substantial motivation. In the end, the various prerequisite topics will not seem so different from each other or from geometry. Looking forward, differential geometry prepares the reader for a sizable portion of modern physics and mathematics.

About This Book

The differential geometry of curves and surfaces is a well-developed topic about which volumes have been written. My goal is to engage the modern reader with clear and colorful explanations of the essential concepts, culminating in the famous Gauss–Bonnet Theorem.

I wrote this book to serve a variety of readers. For readers seeking an elementary text, I kept the prerequisites minimal, included plenty of examples and intermediate steps within proofs, and clearly identified as optional the more excursive applications and the more advanced topics. For readers bound for graduate school in math or physics, this is a clear, concise, rigorous development of the topic including the deep global theorems. For the benefit of *all* readers, I employed every trick I know to render the difficult abstract ideas **herein** more understandable and engaging.

Over 300 color illustrations bring the mathematics to life, instantly clarifying concepts in ways that grayscale could not. Green-boxed definitions and purple-boxed theorems help to visually organize the mathematical content.

Applications abound! The study of conformal and equiareal functions is grounded in its application to cartography. Evolutes, involutes, and cycloids are introduced through Christiaan Huygens's fascinating story. He attempted to solve the famous longitude problem with a mathematically improved pendulum clock. Along the way, he invented mathematics that would later be applied to optics and gears. Clairaut's theorem is presented as a conservation law for angular momentum. Green's theorem makes possible a drafting tool called a planimeter. Foucault's pendulum helps one visualize a parallel vector field along a latitude of the Earth. Even better, a south-pointing chariot helps one visualize a parallel vector field along any curve in any surface.

In truth, the most profound application of differential geometry is to modern physics, which is beyond the scope of this book. The GPS in my car wouldn't work without general relativity, formalized through the language of differential geometry. The above-mentioned applications don't purport to match the significance of modern physics, but instead they serve a crucial pedagogical role within this book: to ground each abstract idea in something concrete. Search YouTube for "south-pointing chariot" and you will learn about a fascinating toy, but in this book it's more than that—it's a concrete device that buttresses the rigorous definition of a parallel vector field and motivates the variational formulas for arc length. Throughout this book, applications, metaphors, and visualizations are tools that motivate and clarify the rigorous mathematical content, but never replace it.

To emphasize geometric concepts, I have systematically put local coordinate formulas in their proper place: near the end of each chapter. These formulas empower the reader to compute various curvature measurements in particular examples, but they do not define these measurements. Local coordinate formulas are necessary for certain proofs, but it's amazing how much can be done without them. To me, differential geometry represents a beautiful interwoven collection of visual insights, made rigorous with a bit of calculus and linear algebra. My goal is make it as easy as possible for readers to internalize the proofs and the intuitions that support them.

Prerequisites

This text requires only a minimal one-semester course in each of the following three prerequisite topics:

- (1) Multivariable calculus (not necessarily including Green's theorem)
- (2) Linear algebra
- (3) Real analysis (not necessarily including multivariable content)

Along the way, I have included brief overviews of some of this prerequisite content, plus an appendix covering topology (continuity, connectedness, and compactness) at a pace that assumes some previous exposure.



*Spiderweb segments dangle in the shape of **catenary** curves, exemplifying aspects of the general theory of curves presented in this chapter.*

Curves

In this chapter, we develop the mathematical tools needed to model and study a moving object. The object might be moving in the **plane**:

$$\mathbb{R}^2 = \{(x, y) | x, y \in \mathbb{R}\} \quad (\text{all ordered pairs of real numbers}),$$

like a car driving on a flat parking lot. Or it might be moving in **space**:

$$\mathbb{R}^3 = \{(x, y, z) | x, y, z \in \mathbb{R}\} \quad (\text{all ordered triples of real numbers}),$$

like a **honeybee** flying about. Many results in this chapter will apply generally to an object moving in n -dimensional **Euclidean space**:

$$\mathbb{R}^n = \{(x_1, x_2, \dots, x_n) | \text{ all } x_i \in \mathbb{R}\} \quad (\text{all ordered } n\text{-tuples of real numbers}),$$

but we are primarily concerned with motion in \mathbb{R}^2 (the plane) and \mathbb{R}^3 (space).

1. Parametrized Curves

Suppose that $x(t)$, $y(t)$, and $z(t)$ are the coordinates at time t of an object moving in space. The object's **position vector**¹ at any particular time t is the vector

$$\gamma(t) = (x(t), y(t), z(t)).$$

Let $I \subset \mathbb{R}$ denote the interval of time during which we will study the object's motion, so that $\gamma : I \rightarrow \mathbb{R}^3$. It is reasonable to assume that γ is **smooth**, which means that each of the three **component functions**, $x(t)$, $y(t)$, and $z(t)$, separately is smooth in the sense that it can be differentiated any number of times. In summary, a moving object in space will be modeled by the $n = 3$ case of a parametrized curve:

DEFINITION 1.1.

A **parametrized curve** in \mathbb{R}^n is a smooth function $\gamma : I \rightarrow \mathbb{R}^n$, where $I \subset \mathbb{R}$ is an interval.

Recall that an **interval** means a nonempty connected subset of \mathbb{R} . Every interval has one of the following forms (by Exercise 1.1):

$$(a, b), [a, b], (a, b], [a, b), (-\infty, b), (-\infty, b], (a, \infty), [a, \infty), (-\infty, \infty).$$

For types of intervals that include boundary points, Exercise 1.2 discusses the minor technical issue of how to correctly interpret the hypothesis that the component functions are smooth at these boundary points.

For brevity, we will **henceforth** use the term **curve** as an abbreviation for "parametrized curve."

EXAMPLE 1.2 (A circle). The function $\gamma(t) = (\cos t, \sin t)$, $t \in (-\infty, \infty)$, is a curve in \mathbb{R}^2 ("a plane curve") that models an object traveling counter-clockwise around the unit circle.

EXAMPLE 1.3 (A helix). The function $\gamma(t) = (\cos t, \sin t, t)$, $t \in (-\infty, \infty)$, is a curve in \mathbb{R}^3 ("a space curve"). It models an object whose shadow in the xy -plane traverses the circle from Example 1.2, while simultaneously its z -coordinate steadily increases with time. This path looks like a helix.

EXAMPLE 1.4 (A graph). The function $\gamma(t) = (t, t^2)$, $t \in (-\infty, \infty)$, is a plane curve that models an object traveling along the parabola $y = x^2$ in such a way that its x -coordinate always equals the time parameter.

More generally, if I is an interval and $f : I \rightarrow \mathbb{R}$ is a smooth function, then $\gamma(t) = (t, f(t))$ models an object traversing the graph of $y = f(x)$.

¹Many calculus books enclose vectors in pointed brackets, like $\langle x(t), y(t), z(t) \rangle$, but we will always use round parentheses. The term "vector" is a synonym for "element of Euclidean space." In some situations, it is best visualized as a point, and in others as an arrow with its tail at some particular position. Names of vectors (and vector-valued functions) will be typeset in boldface throughout Chaps. 1 and 2. For handwritten math, we recommend over-arrows rather than bold, like $\vec{\gamma}$.

EXAMPLE 1.5 (A line). The plane curve $\gamma(t) = (2 + 3t, 4 - t)$, $t \in (-\infty, \infty)$, has linear component functions, so you might guess that it models an object moving along a straight line. To confirm this guess, rewrite the formula as

$$\gamma(t) = (2, 4) + t(3, -1),$$

and picture it as in Fig. 1.1.

In the above examples, the component functions are smooth because the familiar elementary functions from calculus are known to be smooth. Specifically, the following classes of functions are smooth on their domains: polynomial, rational, exponential, logarithmic, trigonometric, and inverse-trigonometric. Power functions are also smooth on their domains, except possibly at zero (for example, $f(x) = x^{1/3}$ is not differentiable at $x = 0$).

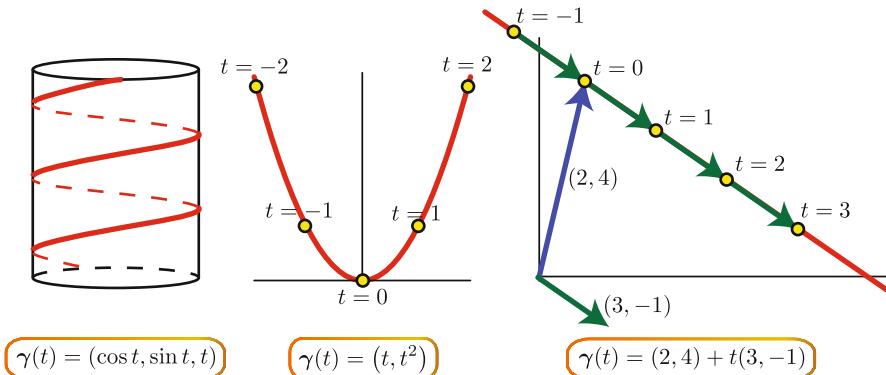


FIGURE 1.1. The curves from Examples 1.3, 1.4, and 1.5

Furthermore, the sum, difference, product, quotient, or composition of smooth functions is a smooth function on its domain.

Derivatives of curves are computed componentwise:

DEFINITION 1.6.

If $\gamma : I \rightarrow \mathbb{R}^n$ is a curve with components $\gamma(t) = (x_1(t), x_2(t), \dots, x_n(t))$, then its **derivative**, $\gamma' : I \rightarrow \mathbb{R}^n$, is the curve defined as $\gamma'(t) = (x'_1(t), x'_2(t), \dots, x'_n(t))$. Higher-order derivatives are defined analogously.

For example, the space curve $\gamma(t) = (x(t), y(t), z(t))$ has first derivative $\gamma'(t) = (x'(t), y'(t), z'(t))$, second derivative $\gamma''(t) = (x''(t), y''(t), z''(t))$, and so on. The following alternative definition is easier to interpret visually:

PROPOSITION 1.7.

The derivative of a curve $\gamma : I \rightarrow \mathbb{R}^n$ at time $t \in I$ is given by the formula

$$\gamma'(t) = \lim_{h \rightarrow 0} \frac{\gamma(t+h) - \gamma(t)}{h}.$$

PROOF. Exercise 1.5. □

If t is a boundary point of I , then this limit must be interpreted as a left- or right-hand limit, as explained in Exercise 1.2.

This looks like the familiar formula from calculus, but notice that the numerator involves a subtraction of two vectors in \mathbb{R}^n . Visually, $\gamma(t+h) - \gamma(t)$ is the vector that when drawn with its tail at $\gamma(t)$, will have its tip at $\gamma(t+h)$. After this vector is divided by h , its length approximates the object's speed, and its direction approximately the direction of motion. Thus, if $\gamma'(t)$ is drawn with its tail at $\gamma(t)$, then it will be tangent to the path of motion, and its length will equal the object's speed (see Fig. 1.2). It is therefore reasonable to define *speed* and *arc length* as follows:

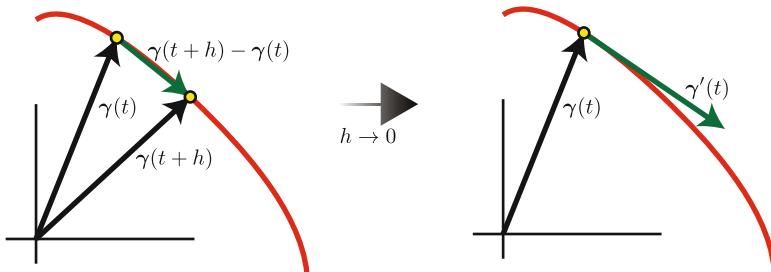


FIGURE 1.2. The visual interpretation of $\gamma'(t) = \lim_{h \rightarrow 0} \frac{\gamma(t+h) - \gamma(t)}{h}$

DEFINITION 1.8.

The **speed** at time $t \in I$ of a curve $\gamma : I \rightarrow \mathbb{R}^n$ is $|\gamma'(t)|$. The **arc length** between times t_1 and t_2 is $\int_{t_1}^{t_2} |\gamma'(t)| dt$.

The vertical bars denote the length (also called the **norm**) of a vector, defined as

$$|(x_1, x_2, \dots, x_n)| = \sqrt{x_1^2 + x_2^2 + \dots + x_n^2}.$$

Think of arc length as “total distance traveled,” which is found by the familiar method of integrating speed.

DEFINITION 1.9.

Let $\gamma : I \rightarrow \mathbb{R}^n$ be a curve. It is called **regular** if its speed is always nonzero ($|\gamma'(t)| \neq 0$ for all $t \in I$). It is called **unit-speed** or **parametrized by arc length** if its speed is always equal to 1 ($|\gamma'(t)| = 1$ for all $t \in I$).

The term “parametrized by arc length” is appropriate because for such a curve, the arc length between t_1 and t_2 equals $\int_{t_1}^{t_2} |\gamma'(t)| dt = \int_{t_1}^{t_2} 1 dt = t_2 - t_1$.

EXAMPLE 1.10 (An Irregular Plane Curve). The function $\gamma(t) = (t^3, t^2)$, $t \in (-\infty, \infty)$, is a plane curve. Its component functions $x(t) = t^3$ and $y(t) = t^2$ satisfy the equation $y(t) = x(t)^{2/3}$ for all t . Therefore, γ models an object that moves on the graph of the equation $y = x^{2/3}$, and in fact visits the entire graph of this equation (see Fig. 1.3). The component functions are smooth, so it might seem surprising that the graph contains a sharp point at the origin. This is possible because $|\gamma'(0)| = 0$, so γ is not regular.

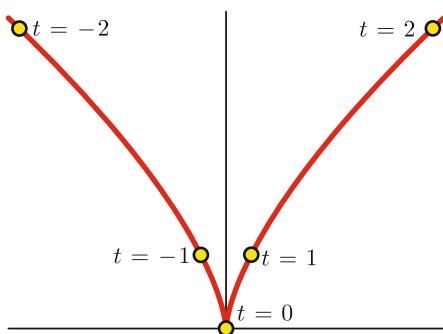


FIGURE 1.3. The graph of $\gamma(t) = (t^3, t^2)$

Since forces don't spontaneously disappear, this is just an approximation (good when h is small):

$$(1.1) \quad \gamma(t_0 + h) \approx \gamma(t_0) + h\gamma'(t_0).$$

Physicists call $\gamma'(t_0)$ the **velocity vector** at time t_0 because it encodes the direction and speed of motion. They call $\gamma''(t_0)$ the **acceleration vector** at time t_0 . In the next few sections, we'll discuss why the term “acceleration” is appropriate.

It is really only for a **regular** curve that Fig. 1.2 provides an appropriate visual interpretation of $\gamma'(t)$. In this case, there are no sharp points; after we have zoomed in enough near $\gamma(t)$, the object's path looks like a straight line, and $\gamma'(t)$ gives the direction of that straight line. If all forces were to disappear at time t_0 , then the object would begin following the straight line that passes through $\gamma(t_0)$ in the direction of $\gamma'(t_0)$, so that h seconds later, its position would be $\gamma(t_0) + h\gamma'(t_0)$.

EXERCISES

EXERCISE 1.1. An interval means a nonempty connected subset of \mathbb{R} . Prove that every interval has one of the following forms:

$$(a, b), [a, b], (a, b], [a, b), (-\infty, b), (-\infty, b], (a, \infty), [a, \infty), (-\infty, \infty).$$

EXERCISE 1.2 (Definition of Smoothness at a Boundary Point).

If I is an interval containing at least one boundary point, then a function $f : I \rightarrow \mathbb{R}$ is called **smooth** if it extends to a smooth function on an *open* interval containing I . In this case, the k th-order **derivative** of f , denoted by $f^{(k)} : I \rightarrow \mathbb{R}$, means the restriction to I of the k th-order derivative of such an extension.

For example, $f : [a, b] \rightarrow \mathbb{R}$ is called smooth if there exist $\epsilon > 0$ and a smooth function $\hat{f} : (a - \epsilon, b + \epsilon) \rightarrow \mathbb{R}$ such that $f(t) = \hat{f}(t)$ for all $t \in [a, b]$. In this case, $f^{(k)} : [a, b] \rightarrow \mathbb{R}$ is defined as the restriction to $[a, b]$ of $\hat{f}^{(k)}$.

If $f : [a, b] \rightarrow \mathbb{R}$ is smooth, show that $f^{(k)} : [a, b] \rightarrow \mathbb{R}$ can equivalently be defined without reference to any extension (so the above definition is independent of the choice of extension). For example, $f'(a)$ can be defined as the *right-hand limit* $f'(a) = \lim_{h \rightarrow 0^+} \frac{f(a+h) - f(a)}{h}$, and $f'(b)$ can be similarly defined as a *left-hand limit*.

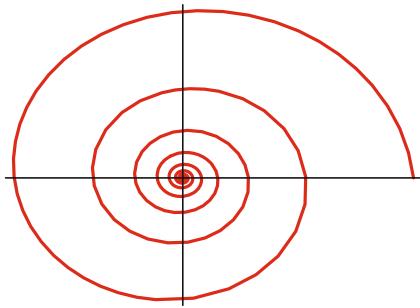


FIGURE 1.4. A portion of a logarithmic spiral

COMMENT: It can be shown that $f : [a, b] \rightarrow \mathbb{R}$ is smooth if and only if its k th-order derivative (defined as above using right- and left-hand limits) exists for every positive integer k . In this way, smoothness can be characterized without reference to any extension.

EXERCISE 1.3. A **logarithmic spiral** means a plane curve of the form

$$\gamma(t) = c(e^{\lambda t} \cos(t), e^{\lambda t} \sin(t)), \quad t \in \mathbb{R},$$

where $c, \lambda \in \mathbb{R}$ with $c \neq 0$. Figure 1.4 shows the restriction to $[0, \infty)$ of a logarithmic spiral with $\lambda < 0$. Use an improper integral to prove that such a restriction has finite arc length even though it makes infinitely many loops around the origin.

EXERCISE 1.4. The curve $\gamma(t) = (\sin(t), \cos(t) + \ln(\tan(t/2))), t \in (\pi/2, \pi)$, is shown in Fig. 1.5. Demonstrate that for every point \mathbf{p} of its image, the segment of the tangent line at \mathbf{p} between \mathbf{p} and the y -axis has length 1. This curve is called a **tractrix** (from the Latin “to drag”), because it represents the path of an object **tethered** by a length-1 rope to a tractor moving upward

along the positive y -axis. *COMMENT:* See the Wikipedia “tractrix” page for a nice animation of this.

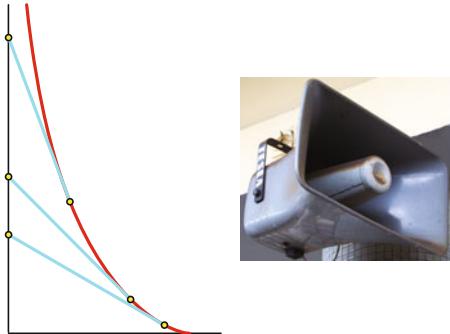


FIGURE 1.5. The tractrix from Exercise 1.4. This curve has acoustic advantages in the design of **horn** speakers. The term “tractrix horn” features prominently in the marketing material of **Klipsch** speakers

EXERCISE 1.5. Prove Proposition 1.7.

EXERCISE 1.6. If all three component functions of a space curve γ are quadratic functions, prove that the image of γ is contained in a plane.

EXERCISE 1.7. Find a plane curve parametrized by arc length that traverses the unit circle clockwise starting at $(0, -1)$.

EXERCISE 1.8. Compute the arc length of $\gamma(t) = (2t, 3t^2)$, $t \in [0, 1]$.

EXERCISE 1.9. Let $a, b > 0$. Find the maximum and minimum speed of the ellipse $\gamma(t) = (a \cos t, b \sin t)$.

EXERCISE 1.10. Prove that the arc length, L , of the graph of the polar coordinate function $r(\theta)$, $\theta \in [a, b]$, is

$$L = \int_a^b \sqrt{r(\theta)^2 + r'(\theta)^2} d\theta.$$

EXERCISE 1.11. Figure 1.6 shows a **polygonal approximation** of the regular curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$. This polygonal approximation is determined by a partition, $a = t_0 < t_1 < t_2 < \dots < t_k = b$. The sum of the lengths of the line segments equals $L = \sum_{i=0}^{k-1} |\gamma(t_{i+1}) - \gamma(t_i)|$. The *mesh* of the partition is defined as $\delta = \max\{t_{i+1} - t_i\}$. Prove that L converges to the arc length of γ for every sequence of partitions for which $\delta \rightarrow 0$.

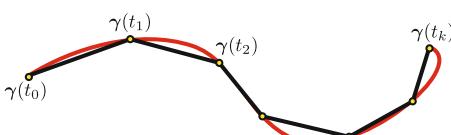


FIGURE 1.6. A polygonal approximation of γ

HINT: For notational simplicity, suppose first that $n = 2$, and denote the component functions by $\gamma(t) = (x(t), y(t))$. For each $i \in \{0, 1, \dots, k-1\}$, the mean value theorem guarantees a sample point $t_i^* \in [t_i, t_{i+1}]$ such that $x'(t_i^*) =$

$\frac{x(t_{i+1}) - x(t_i)}{\Delta t_i}$, where $\Delta t_i = t_{i+1} - t_i$. Similarly, there exists a sample point $t_i^{**} \in [t_i, t_{i+1}]$ such that $y'(t_i^{**}) = \frac{y(t_{i+1}) - y(t_i)}{\Delta t_i}$. Notice that $L = \sum_{i=0}^{k-1} \sqrt{x'(t_i^*)^2 + y'(t_i^{**})^2} \Delta t_i$. If we knew that $t_i^* = t_i^{**}$ for all i , then this expression would be a Riemann sum approximation of $\int_a^b |\gamma'(t)| dt$. But since we can't guarantee that $t_i^* = t_i^{**}$, we must instead use the uniform continuity of x' and y' to show that this expression is close to such a Riemann sum.

EXERCISE 1.12. Use a computer graphing application to plot the following plane curves (all with domain $[0, 2\pi]$):

- (1) The **lemniscate of Bernoulli**

$$\gamma(t) = \left(\frac{\cos t}{1 + \sin^2 t}, \frac{\sin t \cos t}{1 + \sin^2 t} \right).$$

- (2) The **deltoid curve**

$$\gamma(t) = (2n(\cos t)(1 + \cos t) - n, 2n(\sin t)(1 - \cos t))$$

for several choices of the integer $n \geq 1$.

- (3) The **astroid curve**

$$\gamma(t) = c (\cos^3 t, \sin^3 t)$$

for several choices of the constant $c > 0$.

- (4) The **epitrochoid**

$$\gamma(t) = (\cos t, \sin t) - c(\cos(nt), \sin(nt))$$

for several choices of the integer $n > 1$ and the real number $c \in (0, 1)$.

EXERCISE 1.13. Use a computer graphing application to plot the following space curves and view the plots from a variety of angles:

- (1) The **toroidal spiral**

$$\gamma(t) = ((4 + \sin nt) \cos t, (4 + \sin nt) \sin t, \cos nt), \quad t \in [0, 2\pi],$$

for several choices of the positive integer n .

- (2) The **trefoil knot**

$$\gamma(t) = ((2 + \cos 1.5t) \cos t, (2 + \cos 1.5t) \sin t, \sin 1.5t), \quad t \in [0, 4\pi].$$

- (3) The **twisted cubic**

$$\gamma(t) = (t, t^2, t^3), \quad t \in [-1, 1].$$

EXERCISE 1.14. Research the mathematical definition of a **catenary** curve, the physics explanation of why a heavy chain fixed at both ends will dangle in the shape of a catenary, and the history of the study and use of the catenary. Discuss the relationship to the square-wheel tricycle exhibit at the Museum of Mathematics in New York.

2. The Inner Product (*Linear Algebra Background*)

This section reviews some basic facts about the inner product, projections, and orthonormal bases. This material will be necessary for interpreting acceleration in the next section, and will be used repeatedly throughout the book.

DEFINITION 1.11.

The **inner product** of a pair of vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ (with components denoted by $\mathbf{x} = (x_1, x_2, \dots, x_n)$ and $\mathbf{y} = (y_1, y_2, \dots, y_n)$) is

$$\langle \mathbf{x}, \mathbf{y} \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \in \mathbb{R}.$$

The inner product² has the following algebraic properties:

LEMMA 1.12 (Algebraic Properties of the Inner Product).

If $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$ and $\lambda, \mu \in \mathbb{R}$, then:

- (1) $\langle \mathbf{x}, \mathbf{y} \rangle = \langle \mathbf{y}, \mathbf{x} \rangle$ (symmetric),
- (2) $\langle \lambda \mathbf{x} + \mu \mathbf{y}, \mathbf{z} \rangle = \lambda \langle \mathbf{x}, \mathbf{z} \rangle + \mu \langle \mathbf{y}, \mathbf{z} \rangle$,
- $\langle \mathbf{x}, \lambda \mathbf{y} + \mu \mathbf{z} \rangle = \lambda \langle \mathbf{x}, \mathbf{y} \rangle + \mu \langle \mathbf{x}, \mathbf{z} \rangle$ (bilinear),
- (3) $\langle \mathbf{x}, \mathbf{x} \rangle = |\mathbf{x}|^2$, which equals zero only if $\mathbf{x} = \mathbf{0}$.
- (4) $\langle \mathbf{x}, \mathbf{y} \rangle \leq |\langle \mathbf{x}, \mathbf{y} \rangle| \leq |\mathbf{x}||\mathbf{y}|$ (the **Schwarz inequality**).

The visual meaning of $\langle \mathbf{x}, \mathbf{y} \rangle$ involves the angle, $\theta \in [0, \pi]$, between \mathbf{x} and \mathbf{y} :

$$(1.2) \quad \langle \mathbf{x}, \mathbf{y} \rangle = |\mathbf{x}||\mathbf{y}| \cos(\theta).$$

Most calculus texts prove the $n = 2$ case of this formula as a consequence of the law of cosines. For general n , one could regard this formula as the *definition* of angle; that is, the angle between nonzero vectors \mathbf{x} and \mathbf{y} is defined as:

$$\angle(\mathbf{x}, \mathbf{y}) = \cos^{-1} \left(\frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{x}||\mathbf{y}|} \right) \in [0, \pi].$$

Recall that \mathbf{x} and \mathbf{y} are called **orthogonal** if $\langle \mathbf{x}, \mathbf{y} \rangle = 0$. They are called **parallel** if one of them is a scalar multiple of the other. Assuming that \mathbf{x} and \mathbf{y} are nonzero vectors, orthogonal means that $\theta = \pi/2$, while parallel means that $\theta = 0$ or $\theta = \pi$. On the other hand, the **zero vector is both orthogonal and parallel to every other vector**.

The inner product is computed with only multiplication and addition, and yet can be used to calculate certain geometric quantities that one might have expected to require **trigonometry**. In particular, the inner product is useful for computing *components* and *projections*, defined as follows:

²Many calculus books denote this by $x \cdot y$ and call it the *dot product*.

PROPOSITION AND DEFINITION 1.13.

If $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ with $|\mathbf{y}| \neq 0$, then there is a unique way to write \mathbf{x} as a sum of two vectors:

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp},$$

the first of which is parallel to \mathbf{y} and the second of which is orthogonal to \mathbf{y} . The vector \mathbf{x}^{\parallel} is called the **projection** of \mathbf{x} in the direction of \mathbf{y} . The signed length of \mathbf{x}^{\parallel} (that is, the scalar $\lambda \in \mathbb{R}$ such that $\mathbf{x}^{\parallel} = \lambda \frac{\mathbf{y}}{|\mathbf{y}|}$) is called the **component** of \mathbf{x} in the direction of \mathbf{y} .

As shown in Fig. 1.7, the sign of the component signifies whether the angle, $\theta = \angle(\mathbf{x}, \mathbf{y})$, is acute (positive component) or obtuse (negative component), while the absolute value of the component equals the norm of the projection.

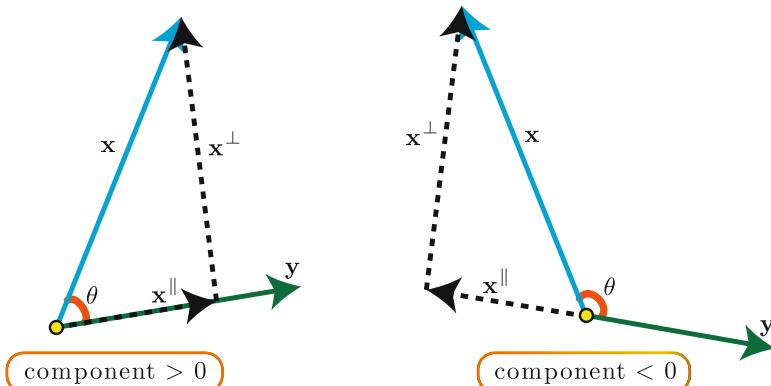


FIGURE 1.7. The vector \mathbf{x}^{\parallel} is called the projection of \mathbf{x} in the direction of \mathbf{y} , while its “signed length” is called the component of \mathbf{x} in the direction of \mathbf{y}

PROOF. We must show that there is a unique value $\lambda \in \mathbb{R}$ such that if we define

$$\mathbf{x}^{\parallel} = \lambda \frac{\mathbf{y}}{|\mathbf{y}|}, \quad \mathbf{x}^{\perp} = \mathbf{x} - \mathbf{x}^{\parallel},$$

then \mathbf{x}^{\perp} is orthogonal to \mathbf{y} . That is, we must solve the following for λ :

$$0 = \langle \mathbf{x}^{\perp}, \mathbf{y} \rangle = \left\langle \mathbf{x} - \frac{\lambda}{|\mathbf{y}|} \mathbf{y}, \mathbf{y} \right\rangle = \langle \mathbf{x}, \mathbf{y} \rangle - \frac{\lambda}{|\mathbf{y}|} \langle \mathbf{y}, \mathbf{y} \rangle = \langle \mathbf{x}, \mathbf{y} \rangle - \lambda |\mathbf{y}|.$$

The unique solution is $\lambda = \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{y}|}$. □

The following formulas can be gleaned from the previous proof:

$$(1.3) \quad \begin{aligned} \text{component} &= \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{y}|} = \underbrace{\langle \mathbf{x}, \mathbf{y} \rangle}_{\text{if } |\mathbf{y}|=1}, & \text{projection} &= \frac{\langle \mathbf{x}, \mathbf{y} \rangle}{|\mathbf{y}|^2} \mathbf{y} = \underbrace{\langle \mathbf{x}, \mathbf{y} \rangle}_{\text{if } |\mathbf{y}|=1} \mathbf{y}. \end{aligned}$$

It is often necessary to project a vector \mathbf{x} in the direction of each member of a collection of vectors. For this activity to be worthwhile, the collection must be orthonormal:

A set $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_k\} \subset \mathbb{R}^n$ is called **orthonormal** if each vector in Y is unit-length and each pair of distinct vectors in Y is orthogonal; in other words,

$$(1.4) \quad \text{for all } i, j \in \{1, \dots, k\}, \quad \langle \mathbf{y}_i, \mathbf{y}_j \rangle = \begin{cases} 1, & \text{if } i = j, \\ 0, & \text{if } i \neq j. \end{cases}$$

Every orthonormal set is linearly independent (Exercise 1.19). So if $k = n$ in the above definition, then Y must be a basis, and therefore deserves to be called an **orthonormal basis** of \mathbb{R}^n .

Even if $k < n$, Y is an orthonormal basis of its **span**. Thus, every $\mathbf{x} \in \text{span}(Y)$ can be written in a unique way as a linear combination: $\mathbf{x} = a_1 \mathbf{y}_1 + \dots + a_k \mathbf{y}_k$, where each $a_i \in \mathbb{R}$. For each i , the component of \mathbf{x} in the direction of \mathbf{y}_i equals the coefficient a_i , because

$$\text{component} = \langle \mathbf{x}, \mathbf{y}_i \rangle = \underbrace{\langle a_1 \mathbf{y}_1 + \dots + a_k \mathbf{y}_k, \mathbf{y}_i \rangle}_{\text{Equation 1.4 and bilinearity}} = a_i.$$

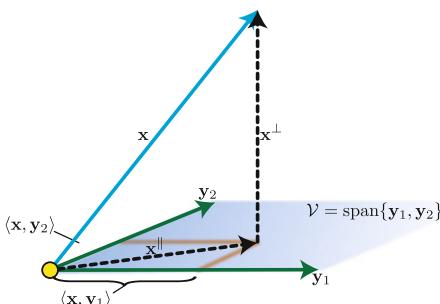


FIGURE 1.8. If $Y = \{\mathbf{y}_1, \mathbf{y}_2\}$ is an orthonormal basis of \mathcal{V} , then $\mathbf{x}^{\parallel} = \langle \mathbf{x}, \mathbf{y}_1 \rangle \mathbf{y}_1 + \langle \mathbf{x}, \mathbf{y}_2 \rangle \mathbf{y}_2$ is the projection of \mathbf{x} onto \mathcal{V}

This is exactly what's advantageous about an *orthonormal* basis—the coefficients used to express a vector \mathbf{x} as a linear combination of the basis elements are just the components of \mathbf{x} in the directions of those basis elements, computed with the inner product. This conclusion is summarized and slightly expanded upon in the following generalization of Definition 1.13, which is illustrated in Fig. 1.8:

PROPOSITION AND DEFINITION 1.14.

If $\mathcal{V} \subset \mathbb{R}^n$ is a subspace and $\mathbf{x} \in \mathbb{R}^n$, then there is a unique way to write \mathbf{x} as a sum of two vectors:

$$\mathbf{x} = \mathbf{x}^{\parallel} + \mathbf{x}^{\perp},$$

the first of which lies in \mathcal{V} and the second of which is orthogonal to every member of \mathcal{V} . The first vector, \mathbf{x}^{\parallel} , is called the **projection** of \mathbf{x} onto \mathcal{V} . If $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ is an orthonormal basis of \mathcal{V} , then \mathbf{x}^{\parallel} is computed as

$$\mathbf{x}^{\parallel} = \langle \mathbf{x}, \mathbf{y}_1 \rangle \mathbf{y}_1 + \dots + \langle \mathbf{x}, \mathbf{y}_k \rangle \mathbf{y}_k.$$

Notice that the coefficients of this linear combination are the components of \mathbf{x} in the directions of the members of Y .

PROOF. Let $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_k\}$ be an orthonormal basis of \mathcal{V} (which exists by Exercise 1.20). We must show that there is a unique choice of scalars $a_1, \dots, a_k \in \mathbb{R}$ such that if we define

$$\mathbf{x}^{\parallel} = a_1 \mathbf{y}_1 + \dots + a_k \mathbf{y}_k, \quad \mathbf{x}^{\perp} = \mathbf{x} - \mathbf{x}^{\parallel},$$

then \mathbf{x}^{\perp} is orthogonal to every vector in \mathcal{V} . Because the inner product is bilinear, this is the same as saying that \mathbf{x}^{\perp} is orthogonal to every element of the basis Y of \mathcal{V} . So we must choose the scalars such that the following is true for each i :

$$0 = \langle \mathbf{x}^{\perp}, \mathbf{y}_i \rangle = \underbrace{\langle \mathbf{x} - a_1 \mathbf{y}_1 - \dots - a_k \mathbf{y}_k, \mathbf{y}_i \rangle}_{\text{Equation 1.4 and bilinearity}} = \langle \mathbf{x}, \mathbf{y}_i \rangle - a_i.$$

The unique solution is $a_i = \langle \mathbf{x}, \mathbf{y}_i \rangle$ for all i , as claimed. \square

Our two uses of the word *projection* are related as follows: the projection of \mathbf{x} in the direction of a vector \mathbf{y} (as in Definition 1.13) equals the projection of \mathbf{x} onto the one-dimensional subspace spanned by \mathbf{y} (as in Definition 1.14).

EXAMPLE 1.15. The **standard orthonormal basis** of \mathbb{R}^n is

$$\{\mathbf{e}_1 = (1, 0, \dots, 0), \mathbf{e}_2 = (0, 1, 0, \dots, 0), \dots, \mathbf{e}_n = (0, \dots, 0, 1)\}.$$

For example, the standard orthonormal basis of \mathbb{R}^3 is

$$\{\mathbf{e}_1 = (1, 0, 0), \mathbf{e}_2 = (0, 1, 0), \mathbf{e}_3 = (0, 0, 1)\}.$$

The components of the vector $\mathbf{v} = (5, 19, -3) \in \mathbb{R}^3$ in the directions of \mathbf{e}_1 , \mathbf{e}_2 , and \mathbf{e}_3 are $\langle \mathbf{v}, \mathbf{e}_1 \rangle = 5$, $\langle \mathbf{v}, \mathbf{e}_2 \rangle = 19$, and $\langle \mathbf{v}, \mathbf{e}_3 \rangle = -3$ respectively.

This shows how our previous informal use of the term “component” agrees with Definition 1.13. Further, the projection of \mathbf{v} onto $\text{span}\{\mathbf{e}_1, \mathbf{e}_2\}$ equals $\langle \mathbf{v}, \mathbf{e}_1 \rangle \mathbf{e}_1 + \langle \mathbf{v}, \mathbf{e}_2 \rangle \mathbf{e}_2 = (5, 19, 0)$.

To return our focus to curves, we now review the product rules for the inner product and scalar multiplication (found in any multivariable calculus text):

LEMMA 1.16.

If $\gamma, \beta : I \rightarrow \mathbb{R}^n$ is a pair of curves, and $c : I \rightarrow \mathbb{R}$ is a smooth function, then:

- (1) $\frac{d}{dt} \langle \gamma(t), \beta(t) \rangle = \langle \gamma'(t), \beta(t) \rangle + \langle \gamma(t), \beta'(t) \rangle$.
- (2) $\frac{d}{dt} (c(t)\gamma(t)) = c'(t)\gamma(t) + c(t)\gamma'(t)$.

Rule (1) is useful for understanding derivatives of pairs of curves that remain always orthonormal:

PROPOSITION 1.17.

Let $\gamma, \beta : I \rightarrow \mathbb{R}^n$ be a pair of curves.

- (1) If γ has constant norm (that is, $|\gamma(t)| = C$ for all $t \in I$), then $\gamma'(t)$ is orthogonal to $\gamma(t)$ for all $t \in I$.
- (2) If $\gamma(t)$ is orthogonal to $\beta(t)$ for all $t \in I$, then

$$\langle \gamma'(t), \beta(t) \rangle = -\langle \gamma(t), \beta'(t) \rangle \text{ for all } t \in I.$$

Notice that the hypotheses of (1) and (2) are both true if $\{\gamma(t), \beta(t)\}$ is orthonormal for all $t \in I$.

PROOF. For part (1), since the expression $\langle \gamma(t), \gamma(t) \rangle = |\gamma(t)|^2 = C^2$ is constant, the derivative of this expression must be zero:

$$0 = \frac{d}{dt} \langle \gamma(t), \gamma(t) \rangle = \langle \gamma'(t), \gamma(t) \rangle + \langle \gamma(t), \gamma'(t) \rangle = 2 \langle \gamma(t), \gamma'(t) \rangle.$$

Thus, $\langle \gamma(t), \gamma'(t) \rangle = 0$, which means that these vectors are orthogonal.

For part (2), since the expression $\langle \gamma(t), \beta(t) \rangle = 0$ is constant, the derivative of this expression must be zero:

$$0 = \frac{d}{dt} \langle \gamma(t), \beta(t) \rangle = \langle \gamma'(t), \beta(t) \rangle + \langle \gamma(t), \beta'(t) \rangle.$$

□

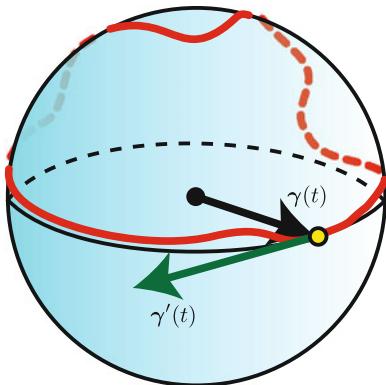


FIGURE 1.9. For an object moving on a sphere, the position vector is always orthogonal to the velocity vector

To visualize part (1) of Proposition 1.17 when $n = 3$, imagine an object moving on a sphere of radius C centered at the origin of \mathbb{R}^3 , as in Fig. 1.9. Can you see why its velocity vector must remain always orthogonal to its position vector? The following natural application of Proposition 1.17 will help us in the next section to interpret the second derivative of a curve geometrically:

PROPOSITION 1.18.

If $\gamma : I \rightarrow \mathbb{R}^n$ is a curve with constant speed, then $\gamma'(t)$ is orthogonal to $\gamma''(t)$ for all $t \in I$.

PROOF. Since speed $= |\gamma'(t)|$ is constant, part (1) of Proposition 1.17 implies that $\gamma'(t)$ is everywhere orthogonal to its derivative, which is $\gamma''(t)$. \square

Although the standard orthonormal basis seems natural, it is often advantageous to “tilt your head,” that is, to choose vectors adapted to the problem at hand. The next example shows how the results from this section let you do exactly that:

EXAMPLE 1.19 (The Shortest Path Between Two Points). Let $\gamma : [a, b] \rightarrow \mathbb{R}^n$ be a curve. Define $\mathbf{p} = \gamma(a)$ and $\mathbf{q} = \gamma(b)$, and let L be the arc length of γ . Since the straight line between these points has arc length $|\mathbf{q} - \mathbf{p}|$, our goal is to prove that $L \geq |\mathbf{q} - \mathbf{p}|$.

To warm up, we will first consider the special case in which $n = 2$, $\mathbf{p} = (0, 0)$, and $\mathbf{q} = (5, 0)$. Denoting the components by $\gamma(t) = (x(t), y(t))$, we have

$$\begin{aligned} L &= \int_a^b |\gamma'(t)| dt = \int_a^b \sqrt{x'(t)^2 + y'(t)^2} dt \geq \int_a^b \sqrt{x'(t)^2 + 0} dt \\ &= \int_a^b |x'(t)| dt \geq \int_a^b x'(t) dt = x(b) - x(a) = 5 - 0 = 5. \end{aligned}$$

Visually, the above equations say that the total distance traveled is greater than or equal to the horizontal distance traveled; see Fig. 1.10 (left).

For the general case, “horizontal” should be replaced with the direction of the unit vector $\mathbf{n} = \frac{\mathbf{q} - \mathbf{p}}{|\mathbf{q} - \mathbf{p}|}$. This head-tilting is computationally achieved as follows:

$$\begin{aligned}
 L &= \int_a^b |\gamma'(t)| dt \geq \underbrace{\int_a^b \langle \gamma'(t), \mathbf{n} \rangle dt}_{\text{Schwarz inequality}} = \int_a^b \frac{d}{dt} \langle \gamma(t), \mathbf{n} \rangle dt \\
 &= \langle \gamma(b), \mathbf{n} \rangle - \langle \gamma(a), \mathbf{n} \rangle = \langle \gamma(b) - \gamma(a), \mathbf{n} \rangle = \left\langle \mathbf{q} - \mathbf{p}, \frac{\mathbf{q} - \mathbf{p}}{|\mathbf{q} - \mathbf{p}|} \right\rangle = |\mathbf{q} - \mathbf{p}|.
 \end{aligned}$$

In the previous special case, we have $\mathbf{n} = e_1 = (1, 0)$, and the picture and logic of our general proof reduce to that of the special case; see Fig. 1.10.

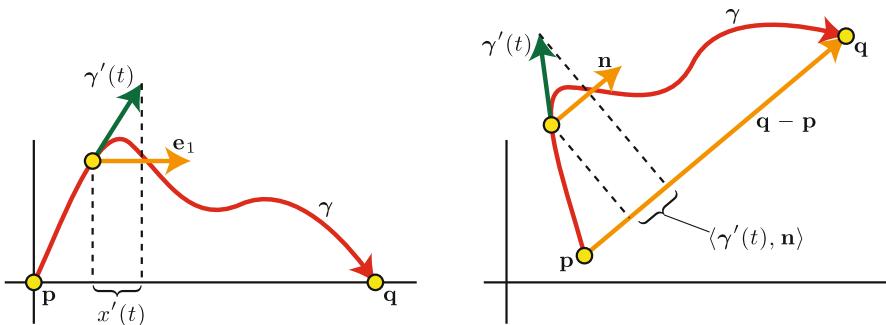


FIGURE 1.10. The arc length of γ is greater than or equal to the integral of the component of velocity in the direction from \mathbf{p} to \mathbf{q}

The above is a small example of a generally useful technique: using a unit vector (or an orthonormal set of vectors) adapted to the problem at hand.

EXERCISES

EXERCISE 1.15. Prove the Schwarz inequality from Lemma 1.12.

COMMENT: Since the Schwarz inequality is needed to ensure that Eq. 1.2 provides a well-defined definition of angle, you may not use Eq. 1.2 to prove the Schwarz inequality.

EXERCISE 1.16. Let γ be a **logarithmic spiral**, as defined in Exercise 1.3 on page 6. Prove that the angle between $\gamma(t)$ and $\gamma'(t)$ is a constant function of t .

EXERCISE 1.17. Let $\mathbf{x} = (0, 2), \mathbf{y} = (3, 4) \in \mathbb{R}^2$. Find the component and the projection of \mathbf{x} in the direction of \mathbf{y} . Write \mathbf{x} as a sum of two vectors, one parallel to \mathbf{y} and the other orthogonal to \mathbf{y} .

EXERCISE 1.18. Let $\mathbf{x} = (1, 2, 3) \in \mathbb{R}^3$ and let $\mathcal{V} = \text{span}\{(1, 0, 1), (1, 1, 0)\}$. Write \mathbf{x} as a sum of two vectors, one in \mathcal{V} and the other orthogonal to every member of \mathcal{V} .

EXERCISE 1.19. Prove that every orthonormal set in \mathbb{R}^n must be linearly independent.

EXERCISE 1.20. Prove that every subspace $\mathcal{V} \subset \mathbb{R}^n$ has an orthonormal basis. *HINT: Begin with an arbitrary basis. Do the following one basis member at a time: subtract from it its projection onto the span of the previous basis members, and then scale it to make it of unit length. This is called the Gram–Schmidt process.*

EXERCISE 1.21. Prove Lemma 1.16

EXERCISE 1.22. Is the converse of part (1) of Proposition 1.17 true?

EXERCISE 1.23. Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve, and let $P \subset \mathbb{R}^3$ be a plane that does not intersect the image of γ . If γ comes closest to P at time t_0 , prove that $\gamma'(t_0)$ is parallel to P .

EXERCISE 1.24. If $t_0 \in I$ is the time at which the curve $\gamma : I \rightarrow \mathbb{R}^n$ is farthest from the origin, prove that $\gamma(t_0)$ is orthogonal to $\gamma'(t_0)$.

EXERCISE 1.25. If $\gamma : I \rightarrow \mathbb{R}^n$ is a regular curve, prove that

$$\frac{d}{dt} |\gamma(t)| = \left\langle \gamma'(t), \frac{\gamma(t)}{|\gamma(t)|} \right\rangle.$$



3. Acceleration

In this section, we interpret the second derivative of a regular curve as acceleration.

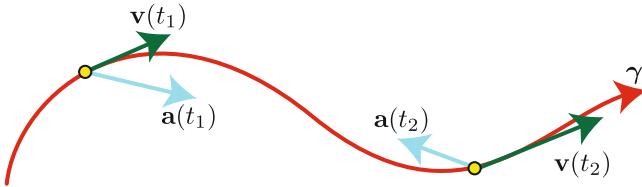
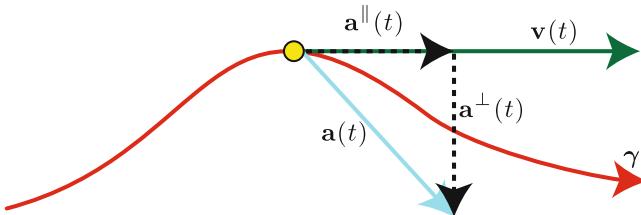
Suppose that $\gamma : I \rightarrow \mathbb{R}^n$ is a regular curve. We will henceforth adopt the following notational convention from physics:

$$\begin{aligned}\mathbf{v}(t) &= \gamma'(t) \quad (\text{the velocity function}) \\ \mathbf{a}(t) &= \gamma''(t) \quad (\text{the acceleration function})\end{aligned}$$

We already know that the term “velocity” is appropriate, because $\mathbf{v}(t)$ encodes the direction and speed of motion. Why is the term “acceleration” appropriate, and how do we visualize $\mathbf{a}(t)$? The physics interpretation of $\mathbf{a}(t)$ comes from the vector version of Newton’s law:

$$\mathbf{F}(t) = m\mathbf{a}(t),$$

where m is the object’s mass, and $\mathbf{F}(t)$ is the vector-valued force acting on the object at time t . Assume for simplicity that $m = 1$. If we draw $\mathbf{a}(t)$ with its tail at $\gamma(t)$, then we can visualize $\mathbf{a}(t)$ as the force vector that is tugging the object so as to make it follow the path that it follows. For example, in Fig. 1.11, the object seems to be speeding up at time t_1 , because $\mathbf{a}(t_1)$ makes an acute angle with $\mathbf{v}(t_1)$, so the force is tugging roughly in the direction of motion. The object seems to be slowing down at time t_2 , because $\mathbf{a}(t_2)$ makes an obtuse angle with $\mathbf{v}(t_2)$, so we have a drag force tugging roughly against the object’s motion. At both times, $\mathbf{a}(t)$ points in the direction in which the path is bending, as required for the force vector to tug the object along its curved path.

FIGURE 1.11. Speeding up at t_1 and slowing down at t_2 FIGURE 1.12. $\mathbf{a}(t) = \mathbf{a}^{\parallel}(t) + \mathbf{a}^{\perp}(t)$

Newton's law does not need to be proven or experimentally verified, because it is the *definition* of force. Nevertheless, we should add some rigor to our claim that $\mathbf{a}(t)$ points in the direction that a force must act in order to make the object follow the path that it follows. The above discussion suggests that we write

$$\mathbf{a}(t) = \mathbf{a}^{\parallel}(t) + \mathbf{a}^{\perp}(t)$$

as a sum of two vectors respectively parallel and orthogonal to $\mathbf{v}(t)$, as in Definition 1.13, and separately interpret the geometric meaning of the two parts (Fig. 1.12).

As discussed above, $\mathbf{a}^{\parallel}(t)$ is the part of the force vector tugging with or against the direction of motion, which should change the object's speed. In fact, the signed length of $\mathbf{a}^{\parallel}(t)$ equals the rate of change of speed:

PROPOSITION 1.20.

$$\frac{d}{dt} |\mathbf{v}(t)| = \frac{\langle \mathbf{a}(t), \mathbf{v}(t) \rangle}{|\mathbf{v}(t)|} = \text{the component of } \mathbf{a}(t) \text{ in the direction of } \mathbf{v}(t).$$

PROOF.

$$\begin{aligned} \frac{d}{dt} |\mathbf{v}(t)| &= \frac{d}{dt} \langle \mathbf{v}(t), \mathbf{v}(t) \rangle^{\frac{1}{2}} \\ &= \frac{1}{2} \langle \mathbf{v}(t), \mathbf{v}(t) \rangle^{-\frac{1}{2}} (\langle \mathbf{v}'(t), \mathbf{v}(t) \rangle + \langle \mathbf{v}(t), \mathbf{v}'(t) \rangle) \\ &= \frac{2 \langle \mathbf{v}'(t), \mathbf{v}(t) \rangle}{2 \langle \mathbf{v}(t), \mathbf{v}(t) \rangle^{\frac{1}{2}}} = \frac{\langle \mathbf{a}(t), \mathbf{v}(t) \rangle}{|\mathbf{v}(t)|}. \end{aligned}$$

□

Notice that Proposition 1.18 follows from Proposition 1.20. It remains only to interpret geometrically $\mathbf{a}^\perp(t)$, which is roughly the part of the force vector that alters the direction in which the object moves. Intuitively, $|\mathbf{a}^\perp(t)|$ should be larger when:

- (1) the object's path is curving more sharply, or
- (2) the object is moving faster.

When you drive in the rain, you avoid turns that are sharp or fast, because in either case, the force of your tire's friction against the wet road might be insufficient to keep you following the road.

To finish this story, in the next two sections we will separate these two effects to obtain a speed-invariant measurement of how sharply the object's path is curving.

EXERCISES

EXERCISE 1.26. What can be said about a space curve with constant acceleration? Compare to Exercise 1.6 on page 7.

EXERCISE 1.27. If γ is a curve in \mathbb{R}^n with $|\gamma(t)| = C$ (a constant), prove that $\langle \mathbf{a}(t), -\gamma(t) \rangle = |\mathbf{v}(t)|^2$. Rewrite this as $\left\langle \mathbf{a}(t), -\frac{\gamma(t)}{|\gamma(t)|} \right\rangle = \frac{|\mathbf{v}(t)|^2}{C}$, and notice that the left side is the component of $\mathbf{a}(t)$ in the direction of the center-pointing vector. Interpret this physically in terms of centripetal force.

EXERCISE 1.28. Find a space curve $\gamma : \mathbb{R} \rightarrow \mathbb{R}^3$ with acceleration function $\mathbf{a}(t) = (t^2 - 1, t^3, t^2 + 1)$. How unique is the solution?



4. Reparametrization

When a regular curve $\gamma : I \rightarrow \mathbb{R}^n$ represents the position function of a moving object, its image $\gamma(I) = \{\gamma(t) \mid t \in I\}$ represents the path that the object follows. This image is called the **trace** of γ . If the object is an airplane, then the trace is the smoke trail it leaves across the sky. Notice that γ is a function, while its trace is a subset of \mathbb{R}^n . The trace contains no time information.

In many situations, we only care about the trace. For example, our goal in the next section will be essentially to assign a *curvature* number to each point of the trace of γ . This measurement should only depend on how sharply the trace bends at that point, not on the time information. In other words, we will want our curvature formula to be unchanged by a *reparametrization*, which roughly means a different association of the time parameter with the points of the trace (literally: “change in parameter”).

EXAMPLE 1.21. Consider the following two regular plane curves:

$$\begin{aligned}\gamma(t) &= (t, t^2), \quad t \in [-2, 2], \\ \tilde{\gamma}(t) &= (2t, (2t)^2), \quad t \in [-1, 1].\end{aligned}$$

They have the same trace, namely the portion of the parabola shown in Fig. 1.13. Define $\phi : [-1, 1] \rightarrow [-2, 2]$ as $\phi(t) = 2t$, and notice that

$$\tilde{\gamma}(t) = \gamma(\phi(t)) = (\gamma \circ \phi)(t).$$

We will therefore call $\tilde{\gamma}$ a reparametrization of γ , and we will refer to γ and $\tilde{\gamma}$ as two parametrizations of the same path.

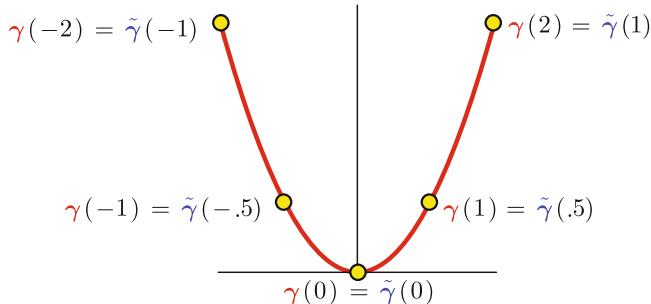


FIGURE 1.13. Two parametrizations of the same path

DEFINITION 1.22.

Suppose that $\gamma : I \rightarrow \mathbb{R}^n$ is a regular curve. A **reparametrization** of γ is a function of the form $\tilde{\gamma} = \gamma \circ \phi : \tilde{I} \rightarrow \mathbb{R}^n$, where \tilde{I} is an interval and $\phi : \tilde{I} \rightarrow I$ is a smooth bijection with nowhere-vanishing derivative ($\phi'(t) \neq 0$ for all $t \in \tilde{I}$); see Fig. 1.14.

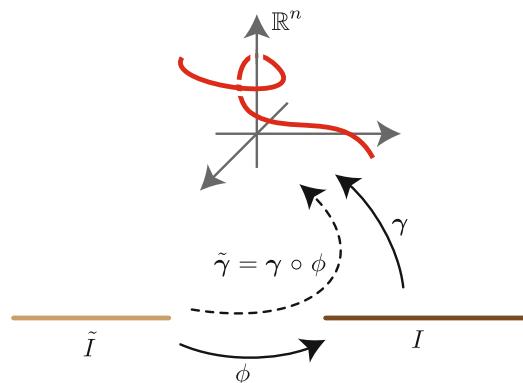


FIGURE 1.14. A reparametrization changes the association of the time parameter with the points of the trace

The requirement that ϕ have a nowhere-vanishing derivative ensures that $\tilde{\gamma}$ is itself a *regular* curve, since the chain rule gives

$$(1.5) \quad \tilde{\gamma}'(t) = \phi'(t)\gamma'(\phi(t)) \neq \mathbf{0} \text{ (the zero vector).}$$

Pause to confirm that this form of the chain rule is valid. It is proved by applying to each component function the familiar chain rule for real-valued functions.

The hypothesis that ϕ has nowhere-vanishing derivative also implies that ϕ must be either monotonically increasing ($\phi' > 0$) or monotonically decreasing ($\phi' < 0$) on all of \tilde{I} , which enables us to make the following definition:

DEFINITION 1.23.

In Definition 1.22, $\tilde{\gamma}$ is called **orientation-preserving** if $\phi' > 0$, and **orientation-reversing** if $\phi' < 0$.

An *orientation* roughly means a choice of direction in which to traverse the trace of a curve (a more precise definition is found in Exercise 1.35). For example, an orientation for the Appalachian Trail is a decision whether to hike it northbound or southbound. Definition 1.23 has to do with whether this choice is preserved or reversed.

EXAMPLE 1.24. The function $\tilde{\gamma}(t) = (-t, (-t)^2)$, $t \in [-2, 2]$, is an orientation-reversing reparametrization of the curve γ from Example 1.21. It traverses the parabola in the opposite direction (right to left).

More generally, if $\gamma : I \rightarrow \mathbb{R}^n$ is any regular curve, then $\tilde{\gamma}(t) = \gamma(-t)$, $t \in \{-s \mid s \in I\}$, is an orientation-reversing reparametrization of γ .

Curves that are parametrized by arc length are particularly easy to compute with. In principle, it is possible to traverse any curve at unit speed.

PROPOSITION 1.25.

A regular curve $\gamma : I \rightarrow \mathbb{R}^n$ can be reparametrized by arc length. That is, there exists a unit-speed reparametrization of γ .

PROOF. Choose any $t_0 \in I$ and consider the arc-length function $s : I \rightarrow \mathbb{R}$ defined as

$$s(t) = \int_{t_0}^t |\gamma'(u)| du.$$

Let $\tilde{I} \subset \mathbb{R}$ denote the image of s . By the fundamental theorem of calculus, $s'(t) = |\gamma'(t)| \neq 0$, from which it can be seen that s is a smooth bijection onto \tilde{I} with nowhere-vanishing derivative. Therefore, s has an inverse function, $\phi : \tilde{I} \rightarrow I$, which is also a smooth bijection with nowhere-vanishing derivative.

Notice that $\tilde{\gamma} = \gamma \circ \phi : \tilde{I} \rightarrow \mathbb{R}^n$ is contrived to be parametrized by arc length. Since s inputs time and outputs arc length, its inverse ϕ must input arc length and output time. Thus, $\gamma(\phi(t))$ is the object's position at the time

when it achieves the arc length t , so it is parametrized by arc length. Or if you prefer computation, the verification looks like this:

$$|\tilde{\gamma}'(t)| = |\phi'(t)\gamma'(\phi(t))| = \phi'(t)|\gamma'(\phi(t))| = \frac{1}{s'(\phi(t))}|\gamma'(\phi(t))| = 1.$$

□

The above proof provides an explicit method to reparametrize any regular curve by arc length; however, in practice this method is usually not computationally reasonable to implement (see Exercise 1.32). Therefore, the value of Proposition 1.25 is more theoretical than computational. We will rarely use it for computation, but our proofs of abstract theorems might include phrases like this: “assume without loss of generality that the regular curve is parametrized by arc length.”

We must slightly modify our definition of “reparametrization” in order for it to correctly apply to a very important class of regular curves, namely those that “close up” by ending at their starting point to smoothly form a loop:

DEFINITION 1.26.

A **closed curve** means a regular curve of the form $\gamma : [a, b] \rightarrow \mathbb{R}^n$ such that $\gamma(a) = \gamma(b)$ and all derivatives match:

$$\gamma'(a) = \gamma'(b), \quad \gamma''(a) = \gamma''(b), \quad \text{etc.}$$

If additionally γ is one-to-one on the domain $[a, b]$, then it is called a **simple closed curve** (Fig. 1.15).

Recall that the derivatives at a and b mentioned in Definition 1.26 are interpreted as left- and right-hand limits, as in Exercise 1.2 on page 6.

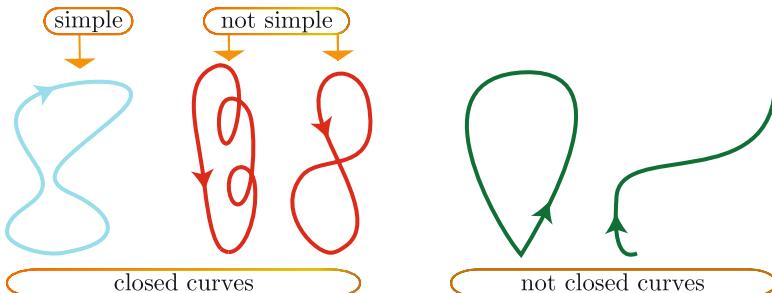


FIGURE 1.15. A simple closed curve closes up smoothly and does not have any self-intersections

The most natural example of a simple closed curve is the circle $\gamma(t) = (\cos t, \sin t)$, $t \in [0, 2\pi]$. This is just the restriction to $[0, 2\pi]$ of the regular curve from Example 1.2, which had the larger domain \mathbb{R} but was described by the same formula; it had period 2π and went around the circle infinitely

many times. More generally, every closed curve can be extended in such a manner:

PROPOSITION 1.27.

A regular curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ is a closed curve if and only if there exists a periodic regular curve $\hat{\gamma} : \mathbb{R} \rightarrow \mathbb{R}^n$ with period $b - a$ such that $\hat{\gamma}(t) = \gamma(t)$ for all $t \in [a, b]$.

PROOF. Exercise 1.33. □

There are two reasons why the definition of *reparametrization* must be modified to apply correctly to the class of closed curves. First, we want every reparametrization of a closed curve to be a closed curve. According to Eq. 1.5 (on page 20), this will require the derivatives of ϕ to match at its domain's two boundary points. Second, it is natural to allow a reparametrization to have a different point at which it begins and ends. The simplest way to change the beginning and ending point of a closed curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ is as follows. Let $\hat{\gamma} : \mathbb{R} \rightarrow \mathbb{R}^n$ denote the periodic extension of γ guaranteed by Proposition 1.27. For $\lambda \in \mathbb{R}$, let γ_λ denote the restriction of $\hat{\gamma}$ to the domain $[a + \lambda, b + \lambda]$. Notice that γ_λ is a closed curve with the same trace as γ , but with a different beginning and ending point. We wish to modify the definition so that γ_λ (or a reparametrization thereof) will qualify as reparametrization of γ . The following will henceforth replace Definition 1.22 for the class of *closed* curves:

DEFINITION 1.28.

*Let $\gamma : [a, b] \rightarrow \mathbb{R}^n$ be a closed curve. A **reparametrization** of γ is a function of the form $\tilde{\gamma} = \gamma_\lambda \circ \phi : [c, d] \rightarrow \mathbb{R}^n$, where $\lambda \in \mathbb{R}$ and $\phi : [c, d] \rightarrow [a + \lambda, b + \lambda]$ is a smooth bijection with nowhere-vanishing derivative, whose derivatives all match at c and d ; that is, $\phi'(c) = \phi'(d)$, $\phi''(c) = \phi''(d)$, etc.*

As mentioned above, the requirement that the derivatives of ϕ match at c and d ensures that every reparametrization of a closed curve is a closed curve. The modified definition of reparametrization for the class of closed curves also ensures the following:

PROPOSITION 1.29.

Two simple closed curves have the same trace if and only if each is a reparametrization of the other.

This conclusion is false for nonsimple closed curves. For example, a trip one time around the unit circle is not a reparametrization of a trip twice around, even though both options have the same trace.

Although we will continue to use the term “curve” as an abbreviation for “parametrized curve,” we mention that some other books define a “curve” to mean an equivalence class of parametrized curves, with two considered equivalent if each is a reparametrization of the other. Exercise 1.35 invites you to explore this viewpoint further and to prove Proposition 1.29. With this latter definition, the information contained in a “curve” might be more than just a trace, but also a decision about how many times to go around the trace.

EXERCISES

EXERCISE 1.29. Consider the following pair of plane curves:

$$\begin{aligned}\gamma(s) &= (\cos s, \sin s), \quad s \in (-\pi, \pi), \\ \tilde{\gamma}(t) &= \left(\frac{1-t^2}{1+t^2}, \frac{2t}{1+t^2} \right), \quad t \in \mathbb{R}.\end{aligned}$$

Verify that $\tilde{\gamma}$ is a reparametrization of γ . *HINT:* $t = \tan(s/2)$.

EXERCISE 1.30. Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve, and let $\tilde{\gamma} = \gamma \circ \phi : \tilde{I} \rightarrow \mathbb{R}^n$ be a reparametrization of γ , as in Definition 1.22.

- (1) Is it possible that I is unbounded while \tilde{I} is bounded?
- (2) Is it possible that I is noncompact while \tilde{I} is compact?
- (3) Is it possible that $I = (a, \infty)$ while $\tilde{I} = [b, \infty)$?

EXERCISE 1.31. Precisely state and prove the assertion that the definition of arc length is independent of parametrization for regular curves.

EXERCISE 1.32. Let $\gamma(t) = (t, t^2)$, $t \in \mathbb{R}$, be the parabolic curve from Example 1.4. Find a unit-speed reparametrization, $\tilde{\gamma}(t) = (\tilde{x}(t), \tilde{y}(t))$, with $\tilde{\gamma}(0) = (0, 0)$. For this, use a computer algebra system to implement the method of the proof of Proposition 1.25. Separately plot the components \tilde{x} and \tilde{y} , which are defined by integrals. Is the computer able to simplify these components into closed-form expressions without integrals?

EXERCISE 1.33. Prove Proposition 1.27.

EXERCISE 1.34. Some texts define a *closed curve* to mean a periodic regular curve $\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$. This viewpoint is equivalent to ours because of Proposition 1.27. Assuming that this alternative definition of a closed curve has been adopted, complete the following necessary modification to Definition 1.28: A *reparametrization* of a closed curve $\gamma : \mathbb{R} \rightarrow \mathbb{R}^n$ is a function of the form $\tilde{\gamma} = \gamma \circ \phi : \mathbb{R} \rightarrow \mathbb{R}^n$, where $\phi : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth bijection with nowhere-vanishing derivative satisfying the following condition: _____.

(Fill in the blank so that a reparametrization of a closed curve is a closed curve.)

EXERCISE 1.35 (“Curve” vs. “Parametrized Curve”). Although this book uses the term “curve” as an abbreviation for “parametrized curve,” some other texts define the term “curve” in an alternative manner that is explored in this exercise.

- (1) Consider two regular parametrized curves in \mathbb{R}^n to be equivalent if one is a reparametrization of the other (as in Definition 1.22). Prove that this is an equivalence relation on the set of all regular parametrized curves in \mathbb{R}^n .
- (2) Consider two parametrized *closed* curves in \mathbb{R}^n to be equivalent if one is a reparametrization of the other (as in Definition 1.28). Prove that this is an equivalence relation on the set of all parametrized closed curves in \mathbb{R}^n .
- (3) Repeat (1) and (2) with “reparametrization” replaced by “orientation-preserving reparametrization.”
- (4) Show that two parametrized simple closed curves have the same trace if and only if they are equivalent (that is, one is a reparametrization of the other).
- (5) If two regular parametrized curves in \mathbb{R}^n have the same trace, and each becomes one-to-one after removing finitely many points from its domain, must they be equivalent?

COMMENTS: The following definition of “curve” is used in some texts. It is really four definitions in one, since each color, red and blue, is separately optional: *Consider the equivalence relation on the set of all regular parametrized (*closed*) curves in \mathbb{R}^n , with two considered equivalent if one is an [orientation-preserving] reparametrization of the other. An equivalence class is called an [oriented] (*closed*) curve in \mathbb{R}^n .* This formalism helps to define more precisely an *orientation* of a curve, C , which should roughly mean a choice of direction in which to traverse it. For this, regard C as an equivalence class, that is, a set of parametrized curves that are all reparametrizations of each other. This set partitions into two oriented curves; that is, two subsets of parametrized curves that are all orientation-preserving reparametrization of each other. An orientation of C means a choice of one of these two subsets. More concisely, an **orientation** of a curve means an oriented curve that it contains.

5. Curvature

Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve. In this section, we will define and study its *curvature function*, $\kappa : I \rightarrow [0, \infty)$. For $t \in I$, the value $\kappa(t)$ will measure how sharply the trace of γ bends as it passes the position $\gamma(t)$. It will

be large if the path bends sharply, and will equal zero if the path looks like a straight line. To calibrate our measurement of curvature, we will compare to circles in \mathbb{R}^2 , with bigger circles declared to have smaller curvature. More precisely, if the trace at $\gamma(t)$ is bending as sharply as a circle of radius r , then $\kappa(t) = \frac{1}{r}$, as in Fig. 1.16.

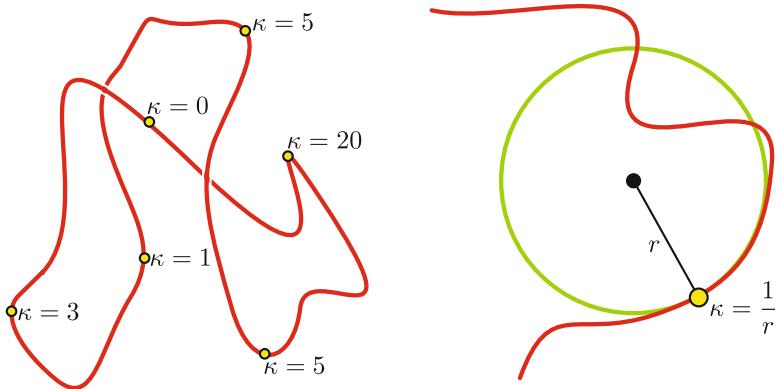


FIGURE 1.16. The curvature, κ , should measure how sharply the trace bends, compared to circles in \mathbb{R}^2

To ensure that our curvature measurement depends only on the path traversed (and not on the speed at which it is traversed), the function κ must be defined by a formula that is *independent of parametrization*; this means that if $\tilde{\gamma} = \gamma \circ \phi$ is a reparametrization of γ (as in Definition 1.22) and $\tilde{\kappa}$ is its curvature function, then we must have

$$\tilde{\kappa} = \kappa \circ \phi.$$

This says that the two parametrizations assign the same curvature number to each point of the trace. For example, $\kappa(1) = \tilde{\kappa}(.5)$ in Fig. 1.13.

Define $\mathbf{v}(t) = \gamma'(t)$ and $\mathbf{a}(t) = \gamma''(t)$ as before. It does *not* quite work to define $\kappa(t)$ to equal $|\mathbf{a}^\perp(t)|$, because this definition would depend on the parametrization. We encountered this conceptually at the end of Sect. 3, when we mentioned that $|\mathbf{a}^\perp(t)|$ depends not only on how sharply the path bends, but also on the object's speed, so it would be increased by a faster parametrization. We can now quantify this dependence. Let $\tilde{\gamma} = \gamma \circ \phi$ be a reparametrization of γ , whose velocity and acceleration functions will be denoted by $\tilde{\mathbf{v}}(t)$ and $\tilde{\mathbf{a}}(t)$ respectively. The chain rule (Eq. 1.5 on page 20) and the product rule (Lemma 1.16(2) on page 13) together give

$$\begin{aligned}\tilde{\mathbf{v}}(t) &= \phi'(t)\mathbf{v}(\phi(t)), \\ \tilde{\mathbf{a}}(t) &= \phi''(t)\mathbf{v}(\phi(t)) + \phi'(t)^2\mathbf{a}(\phi(t)), \\ \tilde{\mathbf{a}}^\perp(t) &= 0 \quad + \phi'(t)^2\mathbf{a}^\perp(\phi(t)).\end{aligned}$$

Suppressing the input parameters, we can summarize this as

$$\tilde{\mathbf{v}} = (\phi')\mathbf{v} \quad \text{and} \quad \tilde{\mathbf{a}}^\perp = (\phi')^2 \mathbf{a}^\perp.$$

So the reparametrization scales \mathbf{v} by a factor of (ϕ') and scales \mathbf{a}^\perp by a factor of $(\phi')^2$. The quantity $\frac{|\mathbf{a}^\perp|}{|\mathbf{v}|^2}$ is therefore unaffected by the reparametrization; in other words,

$$(1.6) \quad \frac{|\tilde{\mathbf{a}}^\perp(t)|}{|\tilde{\mathbf{v}}(t)|^2} = \frac{|\mathbf{a}^\perp(\phi(t))|}{|\mathbf{v}(\phi(t))|^2}.$$

We therefore define the curvature function as follows:

DEFINITION 1.30.

Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve. Its **curvature function**, $\kappa : I \rightarrow [0, \infty)$, is defined as

$$\kappa(t) = \frac{|\mathbf{a}^\perp(t)|}{|\mathbf{v}(t)|^2}.$$

Equation 1.6 confirms that $\tilde{\kappa} = \kappa \circ \phi$, so curvature is independent of parametrization, as desired. Thus, $\kappa(t)$ depends only on the trace of the restriction of γ to a small neighborhood of t in I .

Solving for $|\mathbf{a}^\perp(t)|$ gives

$$|\mathbf{a}^\perp(t)| = \kappa(t) \cdot |\mathbf{v}(t)|^2,$$

which quantifies the intuition discussed in Sect. 3 that $|\mathbf{a}^\perp(t)|$ should increase with speed and with the curvature of the path.

Since curvature is parametrization-independent, it is often useful to choose a unit-speed reparametrization in order to simplify the formula:

PROPOSITION 1.31.

If γ is parametrized by arc length, then $\kappa(t) = |\mathbf{a}(t)|$.

PROOF. By Proposition 1.18 on page 14, $\mathbf{a}(t)$ is orthogonal to $\mathbf{v}(t)$, which means that $\mathbf{a}^\perp(t) = \mathbf{a}(t)$, and we have $\kappa(t) = \frac{|\mathbf{a}^\perp(t)|}{|\mathbf{v}(t)|^2} = \frac{|\mathbf{a}(t)|}{1}$. \square

Proposition 1.31 suggests a nice visual interpretation of the curvature of a unit-speed curve $\gamma : I \rightarrow \mathbb{R}^n$. If each velocity vector $\mathbf{v}(t)$ is drawn with its tail at the origin (rather than at the object's position), then $t \mapsto \mathbf{v}(t)$ is visualized as a path on the **($n - 1$)-dimensional sphere**,

$$S^{n-1} = \{\mathbf{p} \in \mathbb{R}^n \mid |\mathbf{p}| = 1\} = \text{the set of all unit vectors in } \mathbb{R}^n.$$

Figure 1.17 illustrates $n = 3$, where you can think of the sphere $S^2 \subset \mathbb{R}^3$ as a physical globe—a “direction-meter” device that allows you at the origin to monitor a faraway moving object remotely, because the device’s internal arrow, $\mathbf{v}(t)$, always points in the direction in which the object is heading.

The arrow tip's derivative, $\mathbf{a}(t)$, records the change in the object's direction. Thus $\kappa(t) = |\mathbf{a}(t)|$ records the rate at which the object's direction changes.

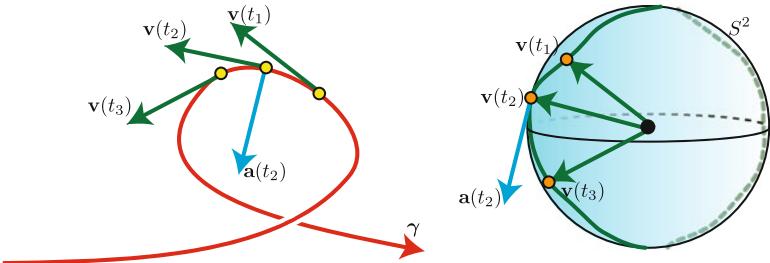


FIGURE 1.17. For a unit-speed curve, curvature measures the rate at which the object's direction changes

EXAMPLE 1.32. The plane curve $\gamma(t) = (r \cos(t), r \sin(t))$, $t \in [0, 2\pi]$, parametrizes the circle with radius r . The velocity and acceleration functions are

$$\begin{aligned}\mathbf{v}(t) &= (-r \sin(t), r \cos(t)), \\ \mathbf{a}(t) &= (-r \cos(t), -r \sin(t)).\end{aligned}$$

Notice that $\mathbf{a}(t) = -\gamma(t)$, reflecting the physical intuition that a center-pointing force is needed to make an object travel in a circle. Since $\langle \mathbf{a}(t), \mathbf{v}(t) \rangle = 0$, we have $\mathbf{a}^\perp(t) = \mathbf{a}(t)$, so

$$\kappa(t) = \frac{|\mathbf{a}^\perp(t)|}{|\mathbf{v}(t)|^2} = \frac{|\mathbf{a}(t)|}{|\mathbf{v}(t)|^2} = \frac{r}{r^2} = \frac{1}{r}.$$

This verifies that our definition agrees with the calibration of curvature indicated by Fig. 1.16: a circle of radius r has constant curvature $\kappa = \frac{1}{r}$.

Since it is convenient to work with orthonormal collections of vectors, we make the following definition:

DEFINITION 1.33.

Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve. Define the **unit tangent** and the **unit normal** vectors at $t \in I$ as

$$\mathbf{t}(t) = \frac{\mathbf{v}(t)}{|\mathbf{v}(t)|}, \quad \mathbf{n}(t) = \underbrace{\frac{\mathbf{a}^\perp(t)}{|\mathbf{a}^\perp(t)|}}_{\text{defined only if } \kappa(t) \neq 0}.$$

Whenever it does not cause confusion, we will suppress the input variable and simply write \mathbf{t}, \mathbf{n} . By construction, $\{\mathbf{t}, \mathbf{n}\}$ is orthonormal.

To generalize the idea of Fig. 1.17 to curves with arbitrary speed, notice that the path $t \mapsto \mathbf{t}(t)$ can always be visualized as a path on S^{n-1} that records the direction in which the object is moving. The derivative of this

path (the rate of change of the object's direction) is related to curvature as follows:

PROPOSITION 1.34.

If $\gamma : I \rightarrow \mathbb{R}^n$ is regular (not necessarily of unit speed), then for all $t \in I$,

$$\kappa(t) = \frac{|\mathbf{t}'(t)|}{|\mathbf{v}(t)|}.$$

PROOF. By Proposition 1.17 (on page 13), $\mathbf{t}' \perp \mathbf{t}$, so

$$\mathbf{a} = \mathbf{v}' = (|\mathbf{v}| \mathbf{t})' = \underbrace{|\mathbf{v}|' \mathbf{t}}_{\mathbf{a}^\parallel} + \underbrace{|\mathbf{v}| \mathbf{t}'}_{\mathbf{a}^\perp}.$$

Therefore, $\kappa = \frac{|\mathbf{a}^\perp|}{|\mathbf{v}|^2} = \frac{||\mathbf{v}| \mathbf{t}'||}{|\mathbf{v}|^2} = \frac{|\mathbf{t}'|}{|\mathbf{v}|}$. \square

Thus, curvature equals the rate at which the object's direction changes divided by its speed (which makes the measurement independent of parametrization). The above proof shows that $\mathbf{a}^\parallel = |\mathbf{v}|' \mathbf{t}$, which provides a simpler verification of Proposition 1.20. It also shows that $\mathbf{a}^\perp = |\mathbf{v}| \mathbf{t}'$. In particular, \mathbf{a}^\perp and \mathbf{t}' point in the same direction, so the following two characterizations of \mathbf{n} are equivalent, and both possibilities capture the idea that \mathbf{n} is the direction in which the curve is turning:

$$(1.7) \quad \mathbf{n} = \frac{\mathbf{a}^\perp}{|\mathbf{a}^\perp|} = \frac{\mathbf{t}'}{|\mathbf{t}'|}.$$

Equation 1.7 and Proposition 1.34 give that $\mathbf{t}' = |\mathbf{t}'| \mathbf{n} = \kappa |\mathbf{v}| \mathbf{n}$. In summary:

PROPOSITION 1.35.

If $\gamma : I \rightarrow \mathbb{R}^n$ is regular (not necessarily of unit speed), then at every time when $\kappa \neq 0$, we have $\boxed{\mathbf{t}' = \kappa |\mathbf{v}| \mathbf{n}}$. Consequently,

$$\underbrace{-\langle \mathbf{n}', \mathbf{t} \rangle}_{\text{by Proposition 1.17}} = \langle \mathbf{t}', \mathbf{n} \rangle = \kappa |\mathbf{v}|.$$

EXAMPLE 1.36 (The Curvature of a Graph at a Critical Point). Let $f : I \rightarrow \mathbb{R}$ be a smooth function with a critical point at $t_0 \in I$, that is, $f'(t_0) = 0$. As mentioned in Example 1.4, the natural parametrization of the graph of f is $\gamma(t) = (t, f(t))$, $t \in I$. Notice that $\mathbf{v}(t) = (1, f'(t))$ and $\mathbf{a}(t) = (0, f''(t))$. In particular, $\mathbf{v}(t_0) = (1, 0)$ and $\mathbf{a}(t_0) = (0, f''(t_0))$ are orthogonal, so $\mathbf{a}^\perp(t_0) = \mathbf{a}(t_0)$. Thus,

$$\kappa(t_0) = \frac{|\mathbf{a}(t_0)|}{|\mathbf{v}(t_0)|^2} = |f''(t_0)|.$$

In summary, curvature = |concavity| at a critical point of a graph.

The conclusion of the above example applies only at a critical point of the graph, but an arbitrary point would look like a critical point if you tilted your head so that \mathbf{t} appeared horizontal. Even better, we will now generalize the conclusion of the above example to every point of every curve in \mathbb{R}^n .

To explain this generalization, assume for the remainder of this section that $\gamma : I \rightarrow \mathbb{R}^n$ is a *unit-speed* curve and that $t_0 \in I$ with $\kappa(t_0) \neq 0$. The (componentwise) second-order Taylor polynomial of γ at time t_0 is

$$\begin{aligned}\gamma(t_0 + h) &\approx \gamma(t_0) + h\gamma'(t_0) + \frac{h^2}{2}\gamma''(t_0) \\ &= \gamma(t_0) + \textcolor{teal}{h}\mathbf{t} + \frac{\kappa h^2}{2}\mathbf{n},\end{aligned}$$

where $\mathbf{t} = \mathbf{t}(t_0)$ and $\mathbf{n} = \mathbf{n}(t_0)$. Therefore, the second-order Taylor polynomials for the components of the displacement vector $\mathbf{D}(h) = \gamma(t_0 + h) - \gamma(t_0)$ in the directions of \mathbf{t} and \mathbf{n} are

$$\begin{aligned}x(h) &= \langle \mathbf{D}(h), \mathbf{t} \rangle \approx \textcolor{teal}{h}, \\ y(h) &= \langle \mathbf{D}(h), \mathbf{n} \rangle \approx \frac{\kappa h^2}{2}.\end{aligned}$$

Furthermore, if $\mathbf{b} \in \mathbb{R}^n$ is any unit vector orthogonal to both \mathbf{t} and \mathbf{n} , then the second-order Taylor polynomial for the component of \mathbf{D} in the direction of \mathbf{b} is

$$\langle \mathbf{D}(h), \mathbf{b} \rangle \approx 0.$$

Every equation above containing an approximation symbol “ \approx ” is a *second-order Taylor polynomial*, so the left and right sides of each such equation differ by an error term, $E(h)$, for which $\lim_{h \rightarrow 0} \frac{|E(h)|}{h^2} = 0$.

The notation $x(h)$ and $y(h)$ is appropriate if you imagine repositioning and tilting your head so that it appears as if $\text{span}\{\mathbf{t}, \mathbf{n}\}$ is the xy -plane with the origin at $\gamma(t_0)$; see Fig. 1.18. From this vantage point, the trace of the second-order Taylor approximation of γ at t_0 is the parabola $y = \frac{\kappa}{2}x^2$. The concavity of this parabola is $y''(0) = \kappa$, which provides a general interpretation of curvature: *curvature equals the concavity of the parabola that approximates the trace of the curve at the point*.

The plane spanned by \mathbf{t} and \mathbf{n} is called the *osculating plane* at time t_0 :

$$\text{osculating plane} = \text{span}\{\mathbf{t}, \mathbf{n}\} = \text{span}\{\mathbf{v}, \mathbf{a}\}.$$

It contains the direction in which the curve is heading and the direction in which the curve is turning, so it is not surprising that the trace of the curve is well approximated by a parabola in the translation of this plane to the object's position.

In the remainder of this section, we discuss how the trace of the curve is also well approximated by a circle in the (translated) osculating plane. Specif-

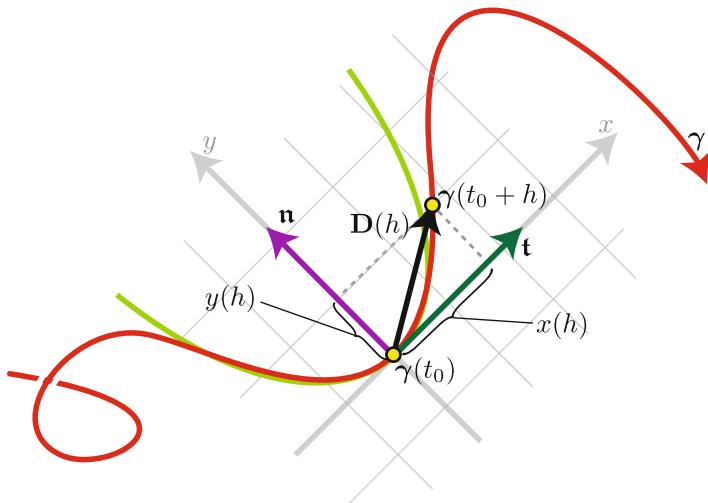


FIGURE 1.18. The trace of γ is approximated near $\gamma(t_0)$ by the parabola with concavity $\kappa(t_0)$ in the $t\mathbf{n}$ -plane (translated to $\gamma(t_0)$)

ically, define the **osculating circle** to be the circle in the osculating plane centered at the origin with radius $\frac{1}{\kappa(t_0)}$. This circle could be parametrized as

$$\mathbf{c}(s) = \frac{1}{\kappa(t_0)} (\cos(s)\mathbf{t} + \sin(s)\mathbf{n}), s \in [0, 2\pi].$$

The definitions are contrived to make the osculating plane become a *subspace* of \mathbb{R}^n , with the osculating circle centered at its origin, which will be advantageous later. But Fig. 1.19 shows the translation of the osculating circle to the position, denoted by $\epsilon(t_0)$, where you naturally imagine it centered. This position is found by starting at $\gamma(t_0)$ and moving the distance $\frac{1}{\kappa(t_0)}$ in the direction of \mathbf{n} :

$$(1.8) \quad \boxed{\epsilon(t_0) = \gamma(t_0) + \frac{1}{\kappa(t_0)}\mathbf{n}.}$$

The term “osculating” comes from the Latin root “to kiss,” because when translated to this position, the osculating circle kisses the trace of the curve at $\gamma(t_0)$. More precisely, for sufficiently small $h > 0$, there is a unique circle containing the three points $\gamma(t_0)$, $\gamma(t_0 + h)$, and $\gamma(t_0 - h)$, and as $h \rightarrow 0$, these unique circles converge to the osculating circle translated to $\epsilon(t_0)$ (Exercise 1.42).

Notice that $t \mapsto \epsilon(t)$ is itself a parametrized curve (not necessarily regular) on any neighborhood of t_0 along which $\kappa \neq 0$. This curve is called the **evolute** of γ , and its geometric significance will be discussed in Sect. 6 of Chap. 2.

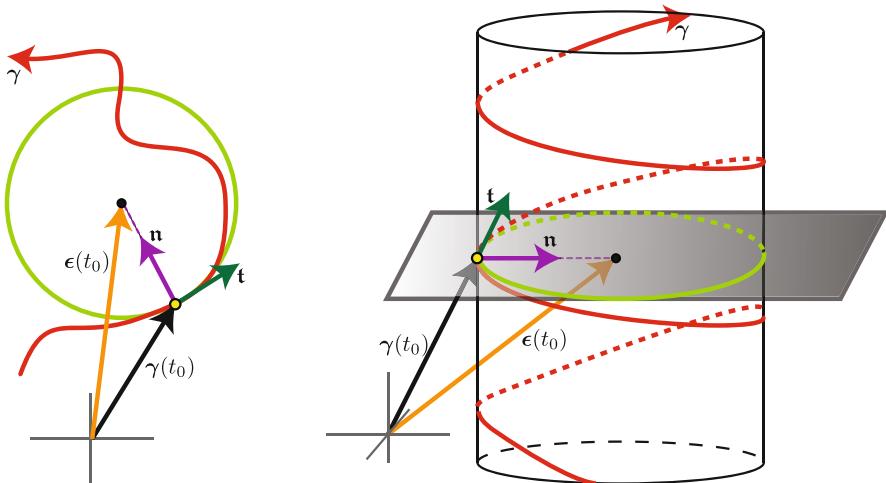


FIGURE 1.19. The osculating circle for a plane curve (*left*) and a space curve (*right*), translated to $\epsilon(t_0)$

EXERCISES

EXERCISE 1.36. *Prove or disprove:* For a regular parametrized curve γ in \mathbb{R}^n , the measurement $f(t) = \frac{|\gamma''(t)|}{|\gamma'(t)|^2}$ is independent of parametrization.

EXERCISE 1.37. In Example 1.32, the curvature could alternatively be computed by reparametrizing the circle by arc length and then applying Proposition 1.31. Verify that this strategy yields the same answer.

EXERCISE 1.38. For the helix in Example 1.3, compute the curvature function:

- (1) directly from Definition 1.30,
- (2) by reparametrizing by arc length and using Proposition 1.31.

EXERCISE 1.39. For constants $a, b, c > 0$, consider the “generalized helix” defined as $\gamma(t) = (a \cos t, b \sin t, ct)$, $t \in \mathbb{R}$. Where is the curvature maximal and minimal?

EXERCISE 1.40. Prove directly that the definition $\kappa(t) = \frac{|\mathbf{t}'(t)|}{|\mathbf{v}(t)|}$ from Proposition 1.34 is independent of parametrization.

EXERCISE 1.41. Let $\gamma : I \rightarrow \mathbb{R}^3$ be a unit-speed space curve with component functions denoted by $\gamma(t) = (x(t), y(t), z(t))$. The plane curve $\bar{\gamma}(t) = (x(t), y(t))$ represents the projection of γ onto the xy -plane. Assume that γ' is nowhere parallel to $(0, 0, 1)$, so that $\bar{\gamma}$ is regular. Let κ and $\bar{\kappa}$ denote the curvature functions of γ and $\bar{\gamma}$ respectively. Let $\mathbf{v}, \bar{\mathbf{v}}$ denote the velocity functions of γ and $\bar{\gamma}$ respectively.

- (1) Prove that $\kappa \geq \bar{\kappa}|\bar{\mathbf{v}}|^2$. In particular, at a time $t \in I$ for which $\mathbf{v}(t)$ lies in the xy -plane, we have $\kappa(t) \geq \bar{\kappa}(t)$.

- (2) Suppose the trace of γ lies on the cylinder $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$. At a time $t \in I$ for which $\mathbf{v}(t)$ lies in the xy -plane (so that γ is tangent to the “waist” of the cylinder), conclude that $\kappa(t) \geq 1$. Is there any *upper* bound for $\kappa(t)$ under these conditions?
- (3) Find an optimal lower bound for $\kappa(t)$ at a time $t \in I$ when $\mathbf{v}(t)$ makes the angle θ with the xy -plane.

EXERCISE 1.42. Let $\gamma : I \rightarrow \mathbb{R}^n$ be a unit-speed curve. Let $t_0 \in I$ and assume $\kappa(t_0) \neq 0$. For sufficiently small $h > 0$, prove that the three points $\gamma(t_0 - h)$, $\gamma(t_0)$, and $\gamma(t_0 + h)$ are not collinear, so there is a unique plane P_h containing them and a unique circle C_h containing them. Precisely formulate and prove the following:

- (1) As $h \rightarrow 0$, P_h converges to the osculating plane (translated to $\gamma(t_0)$).
- (2) As $h \rightarrow 0$, C_h converges to the osculating circle (translated to $\epsilon(t_0)$).

HINT: For (1), use the Taylor approximation formulas from this section. For (2), for fixed h , let $\mathbf{p}(h)$ denote the center of C_h , and define

$$f(s) = |\gamma(t_0 + s) - \mathbf{p}(h)|^2.$$

Since $f(-h) = f(0) = f(h)$, the mean value theorem says that there exist $\delta_1 \in (-h, 0)$ and $\delta_2 \in (0, h)$ with $f'(\delta_1) = f'(\delta_2) = 0$, and then that there exists $\epsilon \in (\delta_1, \delta_2)$ with $f''(\epsilon) = 0$, which becomes

$$\begin{aligned} 0 &= f'(\delta_i) = 2 \langle \gamma'(t_0 + \delta_i), \gamma(t_0 + \delta_i) - \mathbf{p}(h) \rangle \quad (\text{for } i \in \{1, 2\}), \\ 0 &= f''(\epsilon) = 2 \langle \gamma''(t_0 + \epsilon), \gamma(t_0 + \epsilon) - \mathbf{p}(h) \rangle + 2|\gamma'(t_0 + \epsilon)|^2. \end{aligned}$$

Now consider the limit as $h \rightarrow 0$.

EXERCISE 1.43. Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve. Assume that the function $t \mapsto |\gamma(t)|$ has a local maximum value of r occurring at time t_0 . Prove that

$$\kappa(t_0) \geq \frac{1}{r}.$$

Is there any *upper* bound for $\kappa(t_0)$ under these conditions?



6. Plane Curves

Our results so far about curves in \mathbb{R}^n have been valid for all n . But every dimension has its own unique characteristics. In this section, we explore specialized properties of regular *plane curves* (regular curves in the plane \mathbb{R}^2).

What is special about $n = 2$? It is the only dimension in which the terms “clockwise” and “counterclockwise” make sense for describing how a regular curve is turning. More rigorously, consider the linear isomorphism $R_{90} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined as

$R_{90}(x, y) = (-y, x),$

whose effect is to rotate the vector (x, y) by 90 degrees counterclockwise. If you prefer to think of \mathbb{R}^2 as the complex plane, then R_{90} is the “multiplication by i ” function. Notice that $R_{90}(R_{90}(\mathbf{v})) = -\mathbf{v}$ for all $\mathbf{v} \in \mathbb{R}^2$.

Now suppose that $\gamma : I \rightarrow \mathbb{R}^2$ is a *unit-speed* plane curve. At any time $t \in I$, notice that $\mathbf{a}(t)$ and $R_{90}(\mathbf{v}(t))$ are both orthogonal to $\mathbf{v}(t)$, so they must be parallel to each other, which means we can write

$$(1.9) \quad \boxed{\mathbf{a}(t) = \kappa_s(t) \cdot R_{90}(\mathbf{v}(t))}$$

for some scalar $\kappa_s(t) \in \mathbb{R}$. We call $\kappa_s : I \rightarrow \mathbb{R}$ the **signed curvature function**. It is negative if the curve is turning clockwise at t , and positive if counterclockwise, as shown in Fig. 1.20.

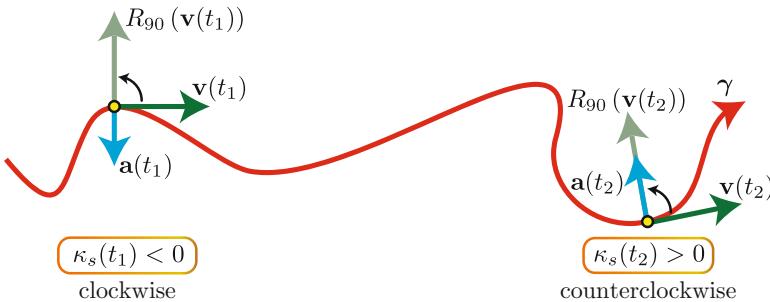


FIGURE 1.20. Signed curvature is positive when the path is turning counterclockwise

This measurement is called “signed curvature” because its sign is interpreted as above, while its absolute value equals the curvature:

LEMMA 1.37.

With the above notation and assumptions, $|\kappa_s(t)| = \kappa(t)$.

PROOF. Notice that $|R_{90}(\mathbf{v}(t))| = |\mathbf{v}(t)| = 1$, so by Proposition 1.31, we have

$$\kappa(t) = |\mathbf{a}(t)| = |\kappa_s(t)R_{90}(\mathbf{v}(t))| = |\kappa_s(t)|.$$

□

Equation 1.9 can be expressed as

$$(1.10) \quad \kappa_s(t) = \langle \mathbf{a}(t), R_{90}(\mathbf{v}(t)) \rangle.$$

Up to this point, we have assumed that γ is of unit speed. But even when it is not, its signed curvature can be computed by a generalization of Eq. 1.10:

DEFINITION 1.38.

If $\gamma : I \rightarrow \mathbb{R}^2$ is a regular plane curve (not necessarily parametrized by arc length), then for all $t \in I$,

$$\kappa_s(t) = \frac{\left\langle \mathbf{a}(t), R_{90} \left(\frac{\mathbf{v}(t)}{|\mathbf{v}(t)|} \right) \right\rangle}{|\mathbf{v}(t)|^2} = \frac{\langle \mathbf{a}(t), R_{90}(\mathbf{v}(t)) \rangle}{|\mathbf{v}(t)|^3}.$$

This definition of $\kappa_s(t)$ agrees with our previous definition for the special case in which γ is of unit speed. Furthermore, the formula is unchanged by orientation-preserving reparametrizations (Exercise 1.50). It follows that $|\kappa_s(t)| = \kappa(t)$ even for non-unit-speed curves. This could also be verified by comparing the definition $\kappa(t) = \frac{|\mathbf{a}^\perp|}{|\mathbf{v}|^2}$ to the above formula for $\kappa_s(t)$, noticing that

$$\left\langle \mathbf{a}(t), R_{90} \left(\frac{\mathbf{v}(t)}{|\mathbf{v}(t)|} \right) \right\rangle = \pm |\mathbf{a}^\perp|.$$

In the previous section, Fig. 1.17 (on page 27) illustrated for a unit-speed *space* curve how $\kappa(t)$ measures the rate at which the object's direction changes. For a unit-speed *plane* curve, the analogous picture is even simpler. If each velocity vector $\mathbf{v}(t)$ is drawn with its tail at the origin (rather than at the object's position), then $t \mapsto \mathbf{v}(t)$ is visualized as a path on the unit circle,

$$S^1 = \{ \mathbf{p} \in \mathbb{R}^2 \mid |\mathbf{p}| = 1 \}.$$

Think of S^1 as a “compass” at the origin whose needle always points in the direction in which the object is heading. Notice that $\kappa(t) = |\mathbf{v}'(t)|$ represents the compass needle's speed, while $\kappa_s(t)$ represents the compass needle's “counterclockwise speed” (it is negative when the needle moves clockwise); see Fig. 1.21. In summary, for an object moving at unit speed in the plane, κ_s represents the rate at which its direction is turning counterclockwise.

Since the position of a compass needle is determined by an angle, this viewpoint suggests that we define an *angle function*, $\theta : I \rightarrow \mathbb{R}$, to report the angle that $\mathbf{v}(t)$ makes with the positive x -axis. But we should not insist that $\theta(t) \in [0, 2\pi]$ for all t , because that would make θ become discontinuous each time the moving object completed a loop (three times in Fig. 1.22), since the outputs would approach 2π but then jump instantaneously to 0. Rather, to construct a *continuous* angle function, we must arrange the definition so that in Fig. 1.22, θ increases by 6π when the entire three-loop path is traversed. The challenge is to show that this is possible.

PROPOSITION 1.39.

If $\gamma : I \rightarrow \mathbb{R}^2$ is a unit-speed plane curve, then there exists a smooth **angle function**, $\theta : I \rightarrow \mathbb{R}$, such that for all $t \in I$, we have

$$(1.11) \quad \mathbf{v}(t) = (\cos \theta(t), \sin \theta(t)).$$

This function is unique up to adding an integer multiple of 2π .

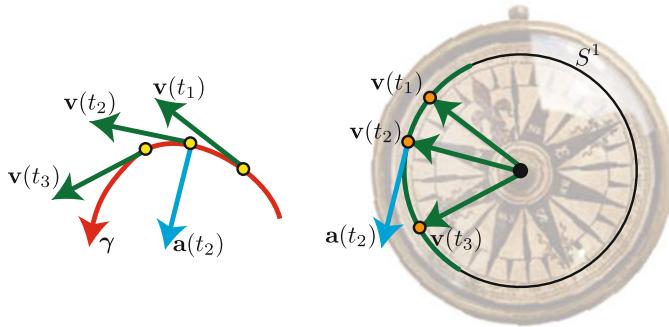


FIGURE 1.21. For a unit-speed plane curve, $\kappa_s(t)$ measures the compass needle's counterclockwise speed at time t



FIGURE 1.22. A curve with three loops

derstand how, denote the components of the velocity function by $\mathbf{v}(t) = (v_x(t), v_y(t))$. In a neighborhood where $v_x > 0$, the function $\theta(t) = \arcsin(v_y(t))$ will work. Similarly, in a neighborhood where $v_y > 0$, the function $\theta(t) = \arccos(v_x(t))$ will work. It is similarly easy to find formulas that work in neighborhoods where $v_x < 0$ or $v_y < 0$. Thus, the real challenge involves defining a single *global angle function* that works on all of I . Notice that a local angle function will still work if you alter it by adding any integer multiple of 2π .

One more crucial observation before we begin the proof: if θ is a local angle function that works in a neighborhood, then for all t in this neighborhood, we have

$$(1.12) \quad \boxed{\theta'(t) = \kappa_s(t)}.$$

That is, “signed curvature equals the rate at which the angle changes.” This important fact is clear from Fig. 1.21, or it can be proven algebraically by

Before beginning the proof, notice that for every $t_0 \in I$, it is relatively easy to construct a *local angle function* that works in a small neighborhood of t_0 in I (“works” means that Eq. 1.11 is valid in this neighborhood). To un-

differentiating Eq. 1.11 to get

$$\mathbf{a}(t) = \theta'(t)(-\sin \theta(t), \cos \theta(t)) = \theta'(t)R_{90}(\mathbf{v}(t)).$$

Equation 1.12 suggests that we might define a global angle function by integrating the function that is supposed to turn out to be its derivative, namely κ_s .

PROOF OF PROPOSITION 1.39.

Choose any $t_0 \in I$, and choose $\theta_0 \in \mathbb{R}$ such that $\mathbf{v}(t_0) = (\cos(\theta_0), \sin(\theta_0))$. Next define our global angle function, $\theta : I \rightarrow \mathbb{R}$, as follows:

$$\theta(t) = \int_{t_0}^t \kappa_s(u) du + \theta_0.$$

Notice that by the fundamental theorem of calculus, $\theta'(t) = \kappa_s(t)$ for all $t \in I$.

To show that this function works, define $\hat{I} \subset I$ to be the set on which Eq. 1.11 holds:

$$\hat{I} = \{t \in I \mid \mathbf{v}(t) = (\cos \theta(t), \sin \theta(t))\}.$$

To complete the proof, it will suffice to show that $\hat{I} = I$.

It is obvious that $t_0 \in \hat{I}$. We claim that \hat{I} also contains every subinterval of I containing t_0 on which a local angle function exists. To see this, add the correct multiple of 2π to the local angle function so that the local and global angle functions agree at t_0 . But since they have the same derivative, the local and global angle functions must agree on the entire subinterval, so this subinterval lies in \hat{I} . Thus, there is a neighborhood³ of t_0 in I that lies in \hat{I} .

The same argument verifies that *every* element of \hat{I} has a neighborhood in I that lies in \hat{I} . Thus, \hat{I} is open in I . It is also straightforward to see that \hat{I} is closed in I . But the interval I is connected, so it does not have any nonempty subsets that are both open and closed, other than all of I (as explained in Sect. 3 of the appendix). Thus, $\hat{I} = I$.

If $\Theta : I \rightarrow \mathbb{R}$ is another smooth function for which Eq. 1.11 is true, then $\Theta(t_0)$ and $\theta(t_0)$ differ by a multiple of 2π . Since θ and Θ have the same derivative on I , namely k_s , they must in fact differ by this multiple of 2π on all of I . \square

Proposition 1.39 provides a language for rigorously defining the *rotation index*, which measures the net number of counterclockwise loops that a unit-speed curve performs.

³This proof uses the relative definitions of “neighborhood,” “open,” and “closed” explained in Sect. 1 of the appendix.

DEFINITION 1.40.

The **rotation index** of a unit-speed closed plane curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ equals $\frac{1}{2\pi}(\theta(b) - \theta(a))$, where θ is the angle function from Proposition 1.39. The rotation index of a regular closed plane curve (not necessarily of unit speed) means the rotation index of an orientation-preserving unit-speed reparametrization of it.

Exercise 1.55 provides methods to compute the rotation index that don't require a unit-speed reparametrization.

Some examples are shown in Fig. 1.23. The rotation index is an integer because $\mathbf{v}(b) = \mathbf{v}(a)$, so $\theta(b) - \theta(a)$ is a multiple of 2π .

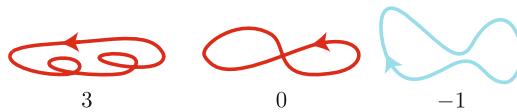


FIGURE 1.23. The rotation indices of some closed plane curves

EXAMPLE 1.41 (The Tire Tracks of a Chariot). Imagine driving a chariot at unit speed along an arbitrary path on a flat surface (the plane). Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be the curve that the center of the chariot follows (the midpoint on the axle between its wheels). When the chariot is turning to the left, its right wheel moves faster than its left wheel. When it is turning to the right, its left wheel moves faster. To understand this phenomenon more quantitatively, observe that the left and right wheels respectively traverse the following paths:

$$\gamma_L(t) = \gamma(t) + cR_{90}(\gamma'(t)), \quad \gamma_R(t) = \gamma(t) - cR_{90}(\gamma'(t)),$$

where $2c$ is the length of the axle; see Fig. 1.24. The velocity of the left wheel is $\gamma'_L = \gamma' + c(R_{90}(\gamma'))' = \gamma' + c(R_{90}(\gamma'')) = \gamma' + c(R_{90}(\kappa_s R_{90}(\gamma'))) = (1 - c\kappa_s)\gamma'$.

Similarly, the right wheel's velocity is $\gamma'_R = (1 + c\kappa_s)\gamma'$. Therefore, the right wheel's speed minus the left wheel's speed equals $2c\kappa_s$ (the length of the axle times the signed curvature).

Let's assume that $1 \pm c\kappa_s > 0$ (the axle is small relative to the turning radius), for otherwise our formulas do not appropriately model the real physical situation here. The length of the tire track is obtained by integrating:

$$\begin{aligned} \text{length}(\gamma_L) &= \int_a^b |\gamma'_L(t)| dt = \int_a^b (1 - c\kappa_s(t)) dt \\ &= \text{length}(\gamma) - c \int_a^b \kappa_s(t) dt = \text{length}(\gamma) - c(\theta(b) - \theta(a)), \end{aligned}$$

where θ is an angle function of γ . In summary, the speed difference between the right and left wheels at any instant is proportional to the signed curvature,

so the length difference between the paths traversed by the right and left wheels is proportional to the net change in the angle function. The constant of proportionality is just the length of the axle.

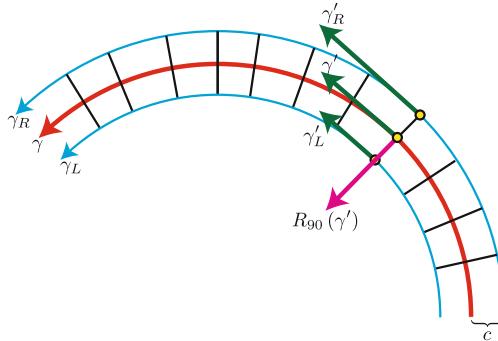


FIGURE 1.24. The tire tracks of a chariot

Figure 1.25 illustrates a mechanical device, called a **south-pointing chariot**, whose operation is based on the principles of Example 1.41. When this chariot is pulled along an arbitrary curve on a flat surface (a plane curve), the statue on top remains always pointing in the same direction. If the statue begins pointing south, then it remains pointing south, even as the chariot curves left or right. This trick is achieved by clever gearing. For example, when the chariot turns left, its right wheel moves faster than its left, and the gearing causes the statue to compensate by spinning right. In fact, the gearing causes the statue's clockwise angular speed (relative to the chariot) to be proportional to the right wheel's speed minus the left wheel's speed. According to Example 1.41, this is exactly what is needed to keep the statue always pointing in the same direction, provided that it is calibrated so the constant of proportionality equals the axle length.

Therefore, the total clockwise rotation of the statue (as observed by someone driving the chariot) is an angle function for the curve that the chariot traverses. So if the charioteer seated next to the statue sees the statue complete three full clockwise rotations before the chariot returns to the starting point of a closed path (the charioteer must duck her head three times as the statue's arm swings clockwise over her head), then the rotation index of the path is 3. We recommend viewing YouTube videos that illustrate how these gears work, and also searching for 3D printable models of south-pointing chariots.

We'll see in Chap. 5 that a south-pointing chariot's behavior on a curved surface illustrates some fundamental geometric principles. Impatient readers can find an excellent general-audience account of these ideas in [6].



FIGURE 1.25. A south-pointing chariot

EXERCISES

EXERCISE 1.44. Describe a regular plane curve with constant speed 3 and constant signed curvature -4 .

EXERCISE 1.45. For the parabola $\gamma(t) = (t, t^2)$, at an arbitrary time $t \in \mathbb{R}$:

- (1) Compute $\kappa(t)$ from Definition 1.30 by decomposing $\mathbf{a}(t) = \mathbf{a}(t)^\parallel + \mathbf{a}(t)^\perp$.
- (2) Compute $\kappa(t)$ using Proposition 1.34.
- (3) Compute $\kappa_s(t)$ using Definition 1.38.

EXERCISE 1.46. Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular plane curve. At a time $t \in I$ with $\kappa(t) \neq 0$, show that Proposition 1.35 implies that $\mathbf{n}' = -\kappa|\mathbf{v}|\mathbf{t}$. Is this true for curves in \mathbb{R}^n with $n > 2$?

EXERCISE 1.47. Prove that the trace of every regular plane curve with constant nonzero signed curvature must equal a circle or a segment of a circle.

EXERCISE 1.48. Prove that the signed curvature function of a regular plane curve described as $\gamma(t) = (x(t), y(t))$ is

$$\kappa_s(t) = \frac{x'(t)y''(t) - x''(t)y'(t)}{(x'(t)^2 + y'(t)^2)^{\frac{3}{2}}}.$$

EXERCISE 1.49. Suppose that $f : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth function. Prove that the signed curvature of the graph of f (oriented left to right) at $(x, f(x))$ equals

$$\kappa_s = \frac{f''(x)}{(1 + f'(x)^2)^{3/2}}.$$

In particular, if $(x, f(x))$ is a critical point, then $\kappa_s = f''(x)$.

EXERCISE 1.50. Prove that the formula for κ_s in Definition 1.38 is unchanged by orientation-preserving reparametrizations.

EXERCISE 1.51. Find a plane curve $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$ whose global angle function is $\theta(t) = 2t$. How unique is the answer?

EXERCISE 1.52. If $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is a closed plane curve and r denotes its rotation index, prove that $\int_a^b \kappa_s(t) dt = 2\pi r$.

EXERCISE 1.53. Use a computer algebra system to graph the signed curvature function of each of the curves that you plotted in Exercise 1.12 on page 8.

EXERCISE 1.54. Prove that the signed curvature of a polar coordinate function $r(\theta)$ is

$$\kappa_s = \frac{2(r')^2 - rr'' + r^2}{((r')^2 + r^2)^{3/2}}.$$

EXERCISE 1.55 (Rotation Index of Variable-Speed Curves). Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a regular (not necessarily unit-speed) closed plane curve.

- (1) Prove there exists a smooth function $\theta : [a, b] \rightarrow \mathbb{R}$ such that for all $t \in [a, b]$, the unit tangent vector satisfies

$$\mathbf{t}(t) = (\cos \theta(t), \sin \theta(t)),$$

and that the rotation index of γ equals $\frac{1}{2\pi} (\theta(b) - \theta(a))$.

- (2) Prove that the rotation index of γ equals

$$\text{rotation index} = \frac{1}{2\pi} \int_a^b \kappa_s(t) |\gamma'(t)| dt.$$

HINT: For both parts, verify that the formula is valid for unit-speed curves and is unchanged by orientation-preserving reparametrizations.

EXERCISE 1.56. Let m, n be positive integers. Find the rotation index of the **Lissajous curve** $\gamma : [0, 2\pi] \rightarrow \mathbb{R}^2$ (see Fig. 1.26) defined as

$$\gamma(t) = (\cos(mt), \sin(nt)).$$

EXERCISE 1.57. In Exercise 1.12(4) on page 8, you graphed the **epitrochoid**

$$\begin{aligned} \gamma(t) &= (\cos t, \sin t) \\ &- c(\cos(nt), \sin(nt)), \quad t \in [0, 2\pi], \end{aligned}$$

for several choices of the integer $n > 1$ and the real number $c \in (0, 1)$. For several choices of n, c , use a computer algebra system to calculate its rotation index.

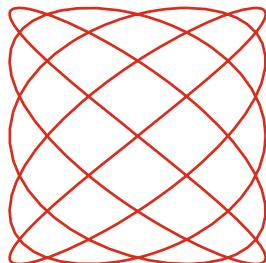


FIGURE 1.26. The Lissajous curve $\gamma(t) = (\cos(5t), \sin(4t))$, $t \in [0, 2\pi]$

EXERCISE 1.58. Let $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$ be a unit-speed plane curve. Let κ_s be its signed curvature function, and let θ be a global angle function for γ . Prove or disprove:

- (1) If θ is periodic, then γ is periodic.
- (2) If κ_s is periodic, then θ is periodic.

EXERCISE 1.59. How are the conclusions of Example 1.41 affected if γ is only regular rather than of unit speed?

EXERCISE 1.60. In Example 1.41, express the signed curvature functions of γ_L and γ_R in terms of c and the signed curvature function of γ .

EXERCISE 1.61. Research the mathematics and the applications of the Reuleaux triangle and other **curves of constant width**.



7. Space Curves

In this section, we will explore regular space curves (regular curves in \mathbb{R}^3). What is unique about the case $n = 3$ within our general study of curves in \mathbb{R}^n ? One answer is that only \mathbb{R}^3 has a **cross product** operation, which is the algebraic heart of most of its special geometric properties. The cross product is defined as follows:

DEFINITION 1.42.

If $\mathbf{a} = (a_1, a_2, a_3), \mathbf{b} = (b_1, b_2, b_3) \in \mathbb{R}^3$, then

$$\mathbf{a} \times \mathbf{b} = (a_2 b_3 - a_3 b_2, a_3 b_1 - a_1 b_3, a_1 b_2 - a_2 b_1) \in \mathbb{R}^3.$$

The familiar geometric properties are given in the following lemma.

LEMMA 1.43.

Let $\mathbf{a}, \mathbf{b} \in \mathbb{R}^3$.

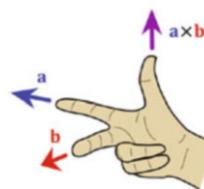
- (1) $\mathbf{a} \times \mathbf{b}$ is orthogonal to both \mathbf{a} and \mathbf{b} .
- (2) $|\mathbf{a} \times \mathbf{b}| = |\mathbf{a}||\mathbf{b}| \sin(\theta) = \sqrt{|\mathbf{a}|^2|\mathbf{b}|^2 - \langle \mathbf{a}, \mathbf{b} \rangle^2} =$ the area of the parallelogram spanned by \mathbf{a} and \mathbf{b} (where $\theta = \angle(\mathbf{a}, \mathbf{b})$).
- (3) The direction of $\mathbf{a} \times \mathbf{b}$ is given by the right-hand rule.

The familiar algebraic properties are given in the following lemma.

LEMMA 1.44.

If $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^3$ and $\lambda, \mu \in \mathbb{R}$, then:

- (1) $\mathbf{a} \times \mathbf{b} = -(\mathbf{b} \times \mathbf{a})$.
- (2) $(\lambda \mathbf{a} + \mu \mathbf{b}) \times \mathbf{c} = \lambda(\mathbf{a} \times \mathbf{c}) + \mu(\mathbf{b} \times \mathbf{c})$, and
 $\mathbf{a} \times (\lambda \mathbf{b} + \mu \mathbf{c}) = \lambda(\mathbf{a} \times \mathbf{b}) + \mu(\mathbf{a} \times \mathbf{c})$.



In addition to the algebraic and geometric properties above, we have a product rule analogous to Lemma 1.16 on page 13:

LEMMA 1.45.

If $\gamma, \beta : I \rightarrow \mathbb{R}^3$ is a pair of space curves, then

$$\frac{d}{dt} (\gamma(t) \times \beta(t)) = \gamma'(t) \times \beta(t) + \gamma(t) \times \beta'(t).$$

The cross product makes possible a convenient formula for the curvature of a space curve:

PROPOSITION 1.46.

If $\gamma : I \rightarrow \mathbb{R}^3$ is a regular space curve, then for all $t \in I$,

$$\kappa(t) = \frac{|\mathbf{v}(t) \times \mathbf{a}(t)|}{|\mathbf{v}(t)|^3}.$$

PROOF. Suppressing the input variable, Fig. 1.27 shows that $|\mathbf{a}^\perp| = |\mathbf{a}| \sin(\theta)$, so

$$\kappa = \frac{|\mathbf{a}^\perp|}{|\mathbf{v}|^2} = \frac{|\mathbf{a}| \sin(\theta)}{|\mathbf{v}|^2} = \frac{|\mathbf{v}| |\mathbf{a}| \sin(\theta)}{|\mathbf{v}|^3} = \frac{|\mathbf{v} \times \mathbf{a}|}{|\mathbf{v}|^3}.$$

□



FIGURE 1.27. $|\mathbf{a}^\perp| = |\mathbf{a}| \sin(\theta) = \frac{1}{|\mathbf{v}|} \cdot (\text{area of parallelogram})$

The cross product also enables us to add one more vector to the frame $\{\mathbf{t}, \mathbf{n}\}$ that was described in Sect. 5.

DEFINITION 1.47.

Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve. Let $t \in I$ with $\kappa(t) \neq 0$. The **Frenet frame** at t is the basis $\{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}$ of \mathbb{R}^3 defined as

$$\mathbf{t}(t) = \frac{\mathbf{v}(t)}{|\mathbf{v}(t)|}, \quad \mathbf{n}(t) = \frac{\mathbf{a}^\perp(t)}{|\mathbf{a}^\perp(t)|} = \frac{\mathbf{t}'(t)}{|\mathbf{t}'(t)|}, \quad \mathbf{b}(t) = \mathbf{t}(t) \times \mathbf{n}(t).$$

Individually they are called the **unit tangent**, **unit normal**, and **unit binormal** vectors at t .

Whenever it won't cause confusion, we will omit the input parameter and just write $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$. By construction, this frame is an *orthonormal* basis of \mathbb{R}^3 .

The osculating plane is spanned by \mathbf{t} and \mathbf{n} , so their cross product \mathbf{b} is a unit-length normal vector to the osculating plane. You might picture the osculating plane as a table top, and \mathbf{b} as the umbrella pole that protrudes

from it and thereby encodes the table's tilt; see Fig. 1.28. Regard $t \mapsto \mathbf{b}(t)$ as a path on the sphere S^2 , and visualize S^2 as a physical globe—a “tilt-meter” device that allows you at the origin to remotely monitor the faraway moving object because its internal arrow, $\mathbf{b}(t)$, always encodes the tilt of the osculating plane. Thus, $|\mathbf{b}'|$ measures the rate at which the osculating plane's tilt is changing.

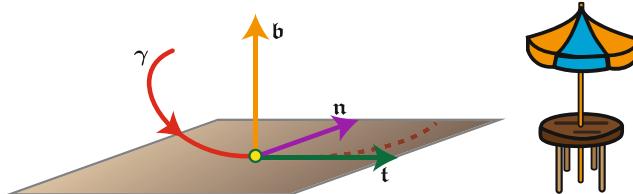


FIGURE 1.28. The Frenet frame and the osculating plane (shown translated to the object's position)

The next key idea is to define the *torsion* of γ as a measurement of how quickly the tilt of the osculating plane changes. The value $|\mathbf{b}'|$ almost works for this, but there are two insufficiencies. First, we'd prefer a measurement that is independent of parametrization. What we really want is the rate at which the osculating plane's tilt would change if the curve were traversed at unit speed. Second, we can do better by defining a *signed* measurement of the rate at which the tilt changes. To understand how, observe that \mathbf{b}' is orthogonal to \mathbf{b} (by Proposition 1.17(1) on page 13) and is also orthogonal to \mathbf{t} because (using Proposition 1.17(2) and Eq. 1.7 on page 28)

$$(1.13) \quad \langle \mathbf{b}', \mathbf{t} \rangle = -\langle \mathbf{t}', \mathbf{b} \rangle = -\langle |\mathbf{t}'| \mathbf{n}, \mathbf{b} \rangle = -|\mathbf{t}'| \langle \mathbf{n}, \mathbf{b} \rangle = 0.$$

Therefore, \mathbf{b}' is parallel to \mathbf{n} , so the absolute value of the measurement $\langle \mathbf{b}', \mathbf{n} \rangle$ equals $|\mathbf{b}'|$, but this measurement carries a sign that will turn out to be geometrically significant. To make this measurement independent of parametrization, we must divide by the speed, like this:⁴

DEFINITION 1.48.

Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve. Let $t \in I$ with $\kappa(t) \neq 0$. The **torsion** of γ at t , denoted by $\tau(t)$, is

$$\tau(t) = \frac{-\langle \mathbf{b}'(t), \mathbf{n}(t) \rangle}{|\mathbf{v}(t)|}.$$

LEMMA 1.49.

Torsion is independent of parametrization.

⁴Some authors define torsion as the negative of this definition.

PROOF. Suppose first that $\tilde{\gamma} = \gamma \circ \phi$ is an orientation-preserving reparametrization ($\phi' > 0$). We'll use tildes to denote functions associated to $\tilde{\gamma}$. First notice that the Frenet frame is unchanged: $\tilde{\mathbf{t}} = \mathbf{t} \circ \phi$, $\tilde{\mathbf{n}} = \mathbf{n} \circ \phi$, and $\tilde{\mathbf{b}} = \mathbf{b} \circ \phi$. Thus

$$\tilde{\tau}(t) = \frac{-\langle \tilde{\mathbf{b}}'(t), \tilde{\mathbf{n}}(t) \rangle}{|\tilde{\mathbf{v}}(t)|} = \frac{-\langle \phi'(t)\mathbf{b}'(\phi(t)), \mathbf{n}(\phi(t)) \rangle}{|\phi'(t)\mathbf{v}(\phi(t))|} = \frac{-\langle \mathbf{b}'(\phi(t)), \mathbf{n}(\phi(t)) \rangle}{|\mathbf{v}(\phi(t))|} = \tau(\phi(t)).$$

On the other hand, if $\tilde{\gamma}$ is an orientation-reversing reparametrization ($\phi' < 0$), then $\tilde{\mathbf{t}} = -\mathbf{t} \circ \phi$, $\tilde{\mathbf{n}} = \mathbf{n} \circ \phi$, and $\tilde{\mathbf{b}} = -\mathbf{b} \circ \phi$. The required sign changes in the above formula cancel, yielding the same conclusion: $\tilde{\tau} = \tau \circ \phi$. \square

EXAMPLE 1.50 (Torsion of a Plane Curve). Let $\gamma : I \rightarrow \mathbb{R}^3$ be a space curve confined to the xy -plane; in other words, $\gamma(t) = (x(t), y(t), 0)$. At every time $t \in I$ for which $\kappa(t) = 0$, the values $\mathbf{n}(t)$, $\mathbf{b}(t)$, and $\tau(t)$ are undefined. At every time $t \in I$ for which $\kappa(t) \neq 0$, notice that $\mathbf{t}(t)$ and $\mathbf{n}(t)$ lie in the xy -plane, so their cross product is $\mathbf{b}(t) = (0, 0, \pm 1)$. According to the right-hand rule, the sign reflects whether the plane curve $t \mapsto (x(t), y(t))$ is turning counterclockwise or clockwise; in other words, it encodes the sign of the signed curvature of this plane curve. Notice that \mathbf{b} is constant on every interval on which it is defined, so $\tau = 0$ wherever it is defined; see Fig. 1.29.

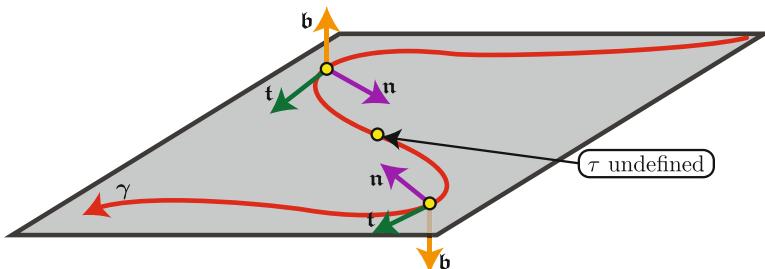


FIGURE 1.29. A space curve constrained to a plane has $\tau = 0$ wherever τ is defined

Zero torsion means that the tilt of the osculating plane is not changing. As in the above example, this phenomenon occurs when the trace of the curve is constrained to a plane. The following converse is also true:

PROPOSITION 1.51.

Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve with $\kappa(t) \neq 0$ for all $t \in I$. The trace of γ is constrained to a plane if and only if $\tau(t) = 0$ for all $t \in I$.

PROOF. First suppose that the trace of γ is constrained to the plane, P , with equation $ax + by + cz = d$. Let $\mathbf{w} = (a, b, c) \in \mathbb{R}^3$, and notice that this plane can be redescribed as

$$P = \{\mathbf{p} \in \mathbb{R}^3 \mid \langle \mathbf{p}, \mathbf{w} \rangle = d\}.$$

Since $\langle \gamma(t), \mathbf{w} \rangle = d$ is a constant function of t , its derivatives vanish:

$$0 = \frac{d}{dt} \langle \gamma(t), \mathbf{w} \rangle = \langle \mathbf{v}(t), \mathbf{w} \rangle, \quad 0 = \frac{d^2}{dt^2} \langle \gamma(t), \mathbf{w} \rangle = \langle \mathbf{a}(t), \mathbf{w} \rangle.$$

It follows from this that $\mathbf{t}(t)$ and $\mathbf{n}(t)$ are both orthogonal to \mathbf{w} , so their cross product must be parallel to \mathbf{w} : $\mathbf{b}(t) = \pm \frac{\mathbf{w}}{|\mathbf{w}|}$. Since \mathbf{b} is continuous, this sign cannot change abruptly, so it must be constant on I . More rigorously, the function $t \mapsto \left\langle \mathbf{b}(t), \frac{\mathbf{w}}{|\mathbf{w}|} \right\rangle$ attains only the two values $+1$ and -1 on I , so by Proposition A.19 from the appendix (on page 353), it must be constant on I . Thus, \mathbf{b} is constant, so $\tau = 0$.

For the other direction, suppose that $\tau(t) = 0$ for all $t \in I$. This implies that $\mathbf{b}'(t) = 0$ for all $t \in I$, so $\mathbf{b}(t) = \mathbf{w}$ (a constant) for all $t \in I$. Notice that

$$\frac{d}{dt} \langle \mathbf{w}, \gamma(t) \rangle = \langle \mathbf{w}, \mathbf{v}(t) \rangle = 0,$$

because \mathbf{b} and \mathbf{t} are orthogonal. Thus $\langle \mathbf{w}, \gamma(t) \rangle = d$ is a constant function. In other words, the trace of γ lies on the plane defined as $P = \{\mathbf{p} \in \mathbb{R}^3 \mid \langle \mathbf{p}, \mathbf{w} \rangle = d\}$. \square

Thus, torsion roughly measures the failure of the trace of the curve to remain in a single plane. To formulate this idea more precisely, we must first compute the derivatives of the vectors in the Frenet frame:

PROPOSITION 1.52 (The Frenet Equations).

Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve. At every time $t \in I$ with $\kappa(t) \neq 0$, the derivatives of the vectors in the Frenet frame are

$$\begin{aligned} \mathbf{t}' &= & |\mathbf{v}| \kappa \mathbf{n}, \\ \mathbf{n}' &= -|\mathbf{v}| \kappa \mathbf{t} & + |\mathbf{v}| \tau \mathbf{b}, \\ \mathbf{b}' &= & -|\mathbf{v}| \tau \mathbf{n}. \end{aligned}$$

PROOF. Recall from Sect. 2 that when a vector is written as a linear combination of members of an orthonormal basis, the coefficients are simply the components, computed with the inner product. The first equation, $\mathbf{t}' = |\mathbf{v}| \kappa \mathbf{n}$, is Proposition 1.35 (on page 28). The second equation, $\mathbf{n}' = -|\mathbf{v}| \kappa \mathbf{t} + |\mathbf{v}| \tau \mathbf{b}$, is proven as follows:

$$\langle \mathbf{n}', \mathbf{t} \rangle = -\langle \mathbf{t}', \mathbf{n} \rangle = -|\mathbf{v}| \kappa \quad (\text{Proposition 1.17(2) and Proposition 1.35})$$

$$\langle \mathbf{n}', \mathbf{n} \rangle = 0 \quad (\text{Proposition 1.17(1)})$$

$$\langle \mathbf{n}', \mathbf{b} \rangle = -\langle \mathbf{b}', \mathbf{n} \rangle = |\mathbf{v}| \tau \quad (\text{Proposition 1.17(2) and the definition of } \tau)$$

The third equation, $\mathbf{b}' = -|\mathbf{v}|\tau\mathbf{n}$, follows from the definition of τ together with the previously mentioned fact that \mathbf{b}' is orthogonal to both \mathbf{b} and \mathbf{t} . \square

The Frenet equations can be written symbolically and compactly with matrix notation:

$$\begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}' = |\mathbf{v}| \begin{pmatrix} 0 & \kappa & 0 \\ -\kappa & 0 & \tau \\ 0 & -\tau & 0 \end{pmatrix} \begin{pmatrix} \mathbf{t} \\ \mathbf{n} \\ \mathbf{b} \end{pmatrix}.$$

The fact that the matrix is skew-symmetric could have been predicted from Proposition 1.17 (on page 13).

We can now describe more precisely how torsion measures the failure of the trace of the curve to remain in a single plane. For the remainder of this section, assume that $\gamma : I \rightarrow \mathbb{R}^3$ is a *unit-speed* space curve, and that $t_0 \in I$ with $\kappa(t_0) \neq 0$. We saw in Sect. 5 that the trace of the *second*-order Taylor polynomial for γ at t_0 is a parabola in the (translated) osculating plane. But computing torsion involves taking three derivatives. If $\tau(t_0) \neq 0$, we will show that the *third*-order Taylor polynomial for γ at t_0 leaves this osculating plane. In fact, the sign of $\tau(t_0)$ will signify whether it leaves by dropping below or rising above this osculating plane.

Adding one more (third-degree) term to the Taylor polynomial at $t_0 \in I$ from Sect. 5 gives

$$\mathbf{D}(h) = \gamma(t_0 + h) - \gamma(t_0) \approx h\gamma'(t_0) + \frac{h^2}{2}\gamma''(t_0) + \frac{h^3}{6}\gamma'''(t_0).$$

When we suppress input variables in the following calculations, for example by writing $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$, we always mean at time t_0 . Notice that

$$\begin{aligned} \gamma' &= \textcolor{teal}{1}\mathbf{t}, \\ \gamma'' &= \textcolor{violet}{\kappa}\mathbf{n}, \\ \gamma''' &= (\kappa\mathbf{n})' = \kappa'\mathbf{n} + \kappa\mathbf{n}' = \kappa'\mathbf{n} + \kappa(-\kappa\mathbf{t} + \tau\mathbf{b}) \\ &= \textcolor{violet}{\kappa}'\mathbf{n} - \textcolor{violet}{\kappa}^2\mathbf{t} + \textcolor{blue}{\kappa}\tau\mathbf{b}. \end{aligned}$$

So the third-order Taylor polynomials for the components of $\mathbf{D}(h)$ in the directions of the vectors of the Frenet frame are

$$(1.14) \quad \begin{aligned} x(h) &= \langle \mathbf{D}(h), \mathbf{t} \rangle \approx \textcolor{teal}{h} & -\frac{\kappa^2}{6}h^3, \\ y(h) &= \langle \mathbf{D}(h), \mathbf{n} \rangle \approx & \frac{\kappa}{2}h^2 + \frac{\kappa'}{6}h^3, \\ z(h) &= \langle \mathbf{D}(h), \mathbf{b} \rangle \approx & \frac{\kappa\tau}{6}h^3. \end{aligned}$$

The labels $x(h)$, $y(h)$, and $z(h)$ are appropriate if you imagine repositioning and tilting your head to a vantage point from which it appears that $\gamma(t_0) = \mathbf{0}$, $\mathbf{t} = (1, 0, 0)$, $\mathbf{n} = (0, 1, 0)$, and $\mathbf{b} = (0, 0, 1)$, which seems roughly to be the case in Fig. 1.30.

Since $\kappa > 0$, the above Taylor polynomial for $z(h)$ implies that if $\tau > 0$, then $z(h) > 0$ for sufficiently small positive h . On the other hand, if $\tau < 0$, then $z(h) < 0$ for sufficiently small positive h . Figure 1.30 illustrates positive torsion, while Fig. 1.28 illustrates negative torsion. In summary, *positive torsion at t_0 implies that γ is passing through the (translated) osculating plane at t_0 from below*. Negative torsion implies from above. Here “above” really means in the direction of \mathbf{b} ; returning to the umbrella metaphor, it’s the side shaded by the umbrella. If the curve in Fig. 1.30 were reparametrized with the opposite orientation, it would still have positive torsion, because \mathbf{b} would change sign, so the umbrella would shade the other side of the osculating plane, so the curve would still pass through from the unshaded side into the shaded side.

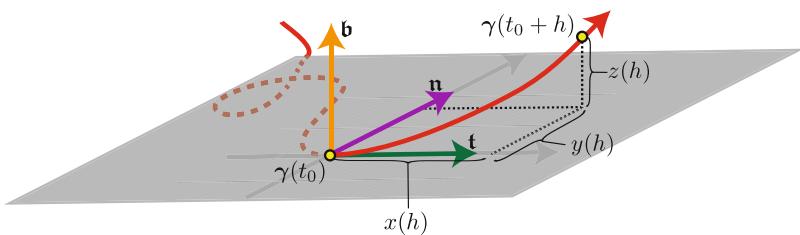


FIGURE 1.30. Positive torsion implies that the curve passes through the osculating plane from below

Every equation above containing an approximation symbol “ \approx ” is a *third-order* Taylor polynomial, so the left and right sides of each such an equation differ by an error term, $E(h)$, for which $\lim_{h \rightarrow 0} \frac{|E(h)|}{h^3} = 0$. However, if we ignore all terms of order three, then the equations 1.14 simplify to the following *second-order* approximations:

$$x(h) \approx h, \quad y(h) \approx \frac{\kappa}{2}h^2, \quad z(h) \approx 0,$$

which just repeats our discovery from Sect. 5 that the trace of the second-order Taylor polynomial for γ at t_0 is a parabola in the (translated) osculating plane with concavity κ .

EXERCISES

EXERCISE 1.62. Calculate the curvature function and torsion function for the curve $\gamma(t) = (t, t^2, t^3)$, $t \in \mathbb{R}$.

EXERCISE 1.63. The *unsigned* curvature of a plane curve $\gamma(t) = (x(t), y(t))$ can be computed with Proposition 1.46 by considering it to have a vanishing third component function: $\gamma(t) = (x(t), y(t), 0)$. Use this method to compute the curvature function of the parabola $\gamma(t) = (t, t^2)$. How can the *signed* curvature be determined from this approach?

EXERCISE 1.64. For the helix in Example 1.3, defined as $\gamma(t) = (\cos t, \sin t, t)$, $t \in \mathbb{R}$, compute the torsion function. Do the same for the helix defined as $\beta(t) = (\cos t, \sin t, -t)$, $t \in \mathbb{R}$. Describe the visual difference between a helix with positive torsion and a helix with negative torsion.

EXERCISE 1.65. Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve, not necessarily parametrized by arc length. Prove that at every $t \in I$ with $\kappa(t) \neq 0$, the torsion is given by the formula

$$\tau = \frac{\langle \gamma' \times \gamma'', \gamma''' \rangle}{|\gamma' \times \gamma''|^2}.$$

HINT: First prove the result for a unit-speed curve, and then prove that the formula is independent of parametrization.

EXERCISE 1.66. For a space curve $\gamma : I \rightarrow \mathbb{R}^3$ at a time $t_0 \in I$ with $\kappa(t_0) \neq 0$, the **rectifying plane** means the plane P passing through $\gamma(t_0)$ spanned by $\mathbf{t}(t_0)$ and $\mathbf{b}(t_0)$. Prove that there exists $\epsilon > 0$ such that the trace of the restriction of γ to $(t_0 - \epsilon, t_0 + \epsilon)$ lies on a single side of P .

EXERCISE 1.67. Use a computer algebra system to graph the curvature and torsion function of each space curve that you plotted in Exercise 1.13 on page 8.

EXERCISE 1.68. Let $\gamma : I \rightarrow \mathbb{R}^3$ be a space curve, and let $t_0 \in I$ with $\kappa(t_0) \neq 0$. Let P denote the osculating plane at t_0 (translated to $\gamma(t_0)$). For $t \in I$ near t_0 , let $\beta(t)$ denote the point of P closest to $\gamma(t)$. Prove that γ and β have the same curvature at time t_0 .

EXERCISE 1.69. Let $\gamma : I \rightarrow \mathbb{R}^3$ be a regular space curve (possibly with points where $\kappa = 0$ and hence where τ is undefined). *Prove or disprove:*

- (1) If the trace of γ lies in a plane, then τ equals zero everywhere it is defined.
- (2) If τ equals zero everywhere it is defined, then the trace of γ lies in a plane.

8. Rigid Motions

A recurring strategy in this chapter is to “tilt your head” to another frame of reference; that is, to use an orthonormal set adapted to the problem at hand. A concrete way to implement this strategy is to apply a *rigid motion*. For example, we could redescribe the Taylor polynomial computation from the previous section by beginning like this: “After applying a rigid motion of \mathbb{R}^3 , we can assume without loss of generality that $\gamma(t_0) = \mathbf{0}$, $\mathbf{t} = (1, 0, 0)$, and $\mathbf{n} = (0, 1, 0)$.”

This section is devoted to background on rigid motions of \mathbb{R}^n , which is crucial for our study of curves now and our study of surfaces later. We will end this section with the fundamental theorems, which essentially say that a curve is rigidly determined by its signed curvature function (for a plane curve) or by its curvature and torsion functions (for a space curve).

DEFINITION 1.53.

A *rigid motion* of \mathbb{R}^n means a function $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ that preserves distances; that is, $\text{dist}(f(\mathbf{p}), f(\mathbf{q})) = \text{dist}(\mathbf{p}, \mathbf{q})$ for all pairs $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$.

As described at the beginning of the appendix, $\text{dist}(\mathbf{p}, \mathbf{q}) = |\mathbf{p} - \mathbf{q}|$, so we could also express this property as

$$|f(\mathbf{p}) - f(\mathbf{q})| = |\mathbf{p} - \mathbf{q}| \quad \text{for all } \mathbf{p}, \mathbf{q} \in \mathbb{R}^n.$$

An important class of rigid motions is the class of linear ones. Since linear transformations are represented by matrices, we must establish some notation for discussing matrices. Let M_n denote the set of all $n \times n$ real matrices. If $\mathbf{A} \in M_n$, let

$$L_{\mathbf{A}} : \mathbb{R}^n \rightarrow \mathbb{R}^n$$

denote the “left multiplication by \mathbf{A} ” function. That is, $L_{\mathbf{A}}(\mathbf{p}) = \mathbf{A} \cdot \mathbf{p}$, where the element $\mathbf{p} \in \mathbb{R}^n$ is thought of as an $n \times 1$ column matrix here, so that this matrix multiplication $\mathbf{A} \cdot \mathbf{p}$ makes sense. In other words, \mathbf{A} is the matrix that represents the linear transformation $L_{\mathbf{A}}$ with respect to the standard orthonormal basis of \mathbb{R}^n .

What special property must \mathbf{A} have in order for $L_{\mathbf{A}}$ to be a rigid motion? We at least need that $|L_{\mathbf{A}}(\mathbf{p}) - L_{\mathbf{A}}(\mathbf{0})| = |\mathbf{p} - \mathbf{0}|$ for all $\mathbf{p} \in \mathbb{R}^n$. Since linear transformations fix the origin ($L_{\mathbf{A}}(\mathbf{0}) = \mathbf{0}$), this becomes $|L_{\mathbf{A}}(\mathbf{p})| = |\mathbf{p}|$. In other words, we at least need for \mathbf{A} to be an *orthogonal matrix*:

DEFINITION 1.54.

A matrix $\mathbf{A} \in M_n$ is called **orthogonal** if $|L_{\mathbf{A}}(\mathbf{p})| = |\mathbf{p}|$ for all $\mathbf{p} \in \mathbb{R}^n$. The set of all orthogonal matrices in M_n is denoted by $O(n)$.

By definition, \mathbf{A} is called orthogonal if $L_{\mathbf{A}}$ preserves norms, which is the same as preserving distances to the origin; in other words, $L_{\mathbf{A}}$ sends each point of the **sphere**,

$$S^{n-1} = \{\mathbf{p} \in \mathbb{R}^n \mid |\mathbf{p}| = 1\},$$

to another point of this sphere. Together with the fact that $L_{\mathbf{A}}$ is a *linear* transformation, we will show that this forces $L_{\mathbf{A}}$ to preserve *all* distances. This claim, among others, is established by the following proposition:

PROPOSITION 1.55.

Let $\mathbf{A} \in M_n$. The following are equivalent:

- (1) $L_{\mathbf{A}}$ is a rigid motion.
- (2) \mathbf{A} is orthogonal.
- (3) $\langle L_{\mathbf{A}}(\mathbf{p}), L_{\mathbf{A}}(\mathbf{q}) \rangle = \langle \mathbf{p}, \mathbf{q} \rangle$ for all pairs $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$.
- (4) $L_{\mathbf{A}}$ preserves orthonormal bases; i.e., if $\{\mathbf{p}_1, \mathbf{p}_2, \dots, \mathbf{p}_n\}$ is an orthonormal basis of \mathbb{R}^n , then so is $\{L_{\mathbf{A}}(\mathbf{p}_1), L_{\mathbf{A}}(\mathbf{p}_2), \dots, L_{\mathbf{A}}(\mathbf{p}_n)\}$.
- (5) The columns of \mathbf{A} form an orthonormal basis of \mathbb{R}^n .
- (6) $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{I}$ (the identity matrix), where \mathbf{A}^T denotes the transpose of \mathbf{A} .

To interpret this proposition, recall from Sect. 2 that a collection of vectors is called *orthonormal* if each has norm 1 and they are mutually orthogonal. Also recall that the *standard orthonormal basis* of \mathbb{R}^n is

$$\{\mathbf{e}_1 = (1, 0, \dots, 0), \mathbf{e}_2 = (0, 1, 0, \dots, 0), \dots, \mathbf{e}_n = (0, 0, \dots, 0, 1)\}.$$

PROOF. (1) \iff (2) If $L_{\mathbf{A}}$ is a rigid motion, then for all $\mathbf{p} \in \mathbb{R}^n$,

$$|L_{\mathbf{A}}(\mathbf{p})| = |L_{\mathbf{A}}(\mathbf{p}) - \mathbf{0}| = |L_{\mathbf{A}}(\mathbf{p}) - L_{\mathbf{A}}(\mathbf{0})| = |\mathbf{p} - \mathbf{0}| = |\mathbf{p}|,$$

so \mathbf{A} is orthogonal. Conversely, if \mathbf{A} is orthogonal, then for all $\mathbf{p}, \mathbf{q} \in \mathbb{R}^n$,

$$|L_{\mathbf{A}}(\mathbf{p}) - L_{\mathbf{A}}(\mathbf{q})| = |L_{\mathbf{A}}(\mathbf{p} - \mathbf{q})| = |\mathbf{p} - \mathbf{q}|,$$

so $L_{\mathbf{A}}$ is a rigid motion.

(2) \iff (3) If (3) is true, then for all $\mathbf{p} \in \mathbb{R}^n$, we have

$$|L_{\mathbf{A}}(\mathbf{p})| = \langle L_{\mathbf{A}}(\mathbf{p}), L_{\mathbf{A}}(\mathbf{p}) \rangle^{\frac{1}{2}} = \langle \mathbf{p}, \mathbf{p} \rangle^{\frac{1}{2}} = |\mathbf{p}|,$$

so (2) is true. In other words, if $L_{\mathbf{A}}$ preserves inner products, then it preserves norms.

For the converse, we must show that the inner product is completely determined by the norm. Solving the equation

$$|\mathbf{p} - \mathbf{q}|^2 = \langle \mathbf{p} - \mathbf{q}, \mathbf{p} - \mathbf{q} \rangle = \langle \mathbf{p}, \mathbf{p} \rangle + \langle \mathbf{q}, \mathbf{q} \rangle - 2 \langle \mathbf{p}, \mathbf{q} \rangle$$

for $\langle \mathbf{p}, \mathbf{q} \rangle$ gives

$$(1.15) \quad \langle \mathbf{p}, \mathbf{q} \rangle = \frac{1}{2} ((|\mathbf{p}|^2 + |\mathbf{q}|^2 - |\mathbf{p} - \mathbf{q}|^2)).$$

From this, it is straightforward to show that if $L_{\mathbf{A}}$ preserves norms, then it preserves inner products.

(3) \implies (4) is obvious.

(4) \implies (5) because the columns of \mathbf{A} are $\{L_{\mathbf{A}}(\mathbf{e}_1), L_{\mathbf{A}}(\mathbf{e}_2), \dots, L_{\mathbf{A}}(\mathbf{e}_n)\}$.

(5) \iff (6) because the (i, j) -entry of $(\mathbf{A}^T \cdot \mathbf{A})$ is

$$\begin{aligned} (\mathbf{A}^T \cdot \mathbf{A})_{ij} &= \langle (\text{row } i \text{ of } \mathbf{A}^T), (\text{column } j \text{ of } \mathbf{A}) \rangle \\ &= \langle (\text{column } i \text{ of } \mathbf{A}), (\text{column } j \text{ of } \mathbf{A}) \rangle. \end{aligned}$$

(5) \implies (3) because if the columns of \mathbf{A} are orthonormal, then for all pairs $\mathbf{p} = (p_1, \dots, p_n), \mathbf{q} = (q_1, \dots, q_n) \in \mathbb{R}^n$, we have

$$\begin{aligned} \langle L_{\mathbf{A}}(\mathbf{p}), L_{\mathbf{A}}(\mathbf{q}) \rangle &= \left\langle L_{\mathbf{A}} \left(\sum_{l=1}^n p_l \mathbf{e}_l \right), L_{\mathbf{A}} \left(\sum_{s=1}^n q_s \mathbf{e}_s \right) \right\rangle \\ &= \sum_{l,s=1}^n p_l q_s \langle L_{\mathbf{A}}(\mathbf{e}_l), L_{\mathbf{A}}(\mathbf{e}_s) \rangle \\ &= \sum_{l,s=1}^n p_l q_s \langle (\text{column } l \text{ of } \mathbf{A}), (\text{column } s \text{ of } \mathbf{A}) \rangle \\ &= p_1 q_1 + \cdots + p_n q_n = \langle \mathbf{p}, \mathbf{q} \rangle. \end{aligned}$$

□

An orthogonal matrix is invertible because its columns form a basis. In fact, we have the following:

LEMMA 1.56.

The inverse of an orthogonal matrix is orthogonal, and the product of two orthogonal matrices is orthogonal.

PROOF. Exercise 1.70

□

We now have a good understanding of the *linear* rigid motions. We next show that every rigid motion that fixes the origin must be linear.

PROPOSITION 1.57.

If f is a rigid motion of \mathbb{R}^n that fixes the origin ($f(\mathbf{0}) = \mathbf{0}$), then $f = L_{\mathbf{A}}$ for some $\mathbf{A} \in O(n)$. In particular, f is linear.

PROOF. Equation 1.15 can be reexpressed as a description of the inner product completely in terms of distances:

$$\langle \mathbf{p}, \mathbf{q} \rangle = \frac{1}{2} (\text{dist}(\mathbf{p}, \mathbf{0})^2 + \text{dist}(\mathbf{q}, \mathbf{0})^2 - \text{dist}(\mathbf{p}, \mathbf{q})^2).$$

Since f preserves distances and fixes the origin, it is straightforward to show using this that it must also preserve the inner product:

$$\langle f(\mathbf{p}), f(\mathbf{q}) \rangle = \langle \mathbf{p}, \mathbf{q} \rangle \text{ for all } \mathbf{p}, \mathbf{q} \in \mathbb{R}^n.$$

Let \mathbf{A} be the matrix whose i th column is $f(\mathbf{e}_i)$, so $f(\mathbf{e}_i) = L_{\mathbf{A}}(\mathbf{e}_i)$ for all $i = 1, \dots, n$. Notice that $\mathbf{A} \in O(n)$, since its columns are orthonormal. We will prove that $f = L_{\mathbf{A}}$ (and thus that f is linear) by showing that

$g = (L_{\mathbf{A}})^{-1} \circ f$ is the identity function. Notice that g is a rigid motion with $g(\mathbf{0}) = \mathbf{0}$ (so g preserves norms and inner products, as above) and $g(\mathbf{e}_i) = \mathbf{e}_i$ for all $i = 1, \dots, n$. Let $\mathbf{p} \in \mathbb{R}^n$. Write $\mathbf{p} = \sum a_i \mathbf{e}_i$ and $g(\mathbf{p}) = \sum b_i \mathbf{e}_i$. Then,

$$b_i = \langle g(\mathbf{p}), \mathbf{e}_i \rangle = \langle g(\mathbf{p}), g(\mathbf{e}_i) \rangle = \langle \mathbf{p}, \mathbf{e}_i \rangle = a_i,$$

which proves $g(\mathbf{p}) = \mathbf{p}$, so g is the identity function. \square

What about rigid motions that do not fix the origin? For any $\mathbf{q} \in \mathbb{R}^n$, the **translation** map $T_{\mathbf{q}} : \mathbb{R}^n \rightarrow \mathbb{R}^n$, defined as $T_{\mathbf{q}}(\mathbf{p}) = \mathbf{p} + \mathbf{q}$, is clearly a rigid motion that sends the origin to \mathbf{q} . We'll prove next that there are no rigid motions other than compositions of translations with linear rigid motions.

PROPOSITION 1.58.

If f is a rigid motion of \mathbb{R}^n , then $f = T_{\mathbf{q}} \circ L_{\mathbf{A}}$ for a unique choice of $\mathbf{q} \in \mathbb{R}^n$ and $\mathbf{A} \in O(n)$.

PROOF. Define $\mathbf{q} = f(\mathbf{0})$. Notice that $(T_{\mathbf{q}})^{-1} \circ f$ is a rigid motion that fixes the origin. By Proposition 1.57, $(T_{\mathbf{q}})^{-1} \circ f = L_{\mathbf{A}}$ for some $\mathbf{A} \in O(n)$. Thus, $f = T_{\mathbf{q}} \circ L_{\mathbf{A}}$. The uniqueness claim is left to the reader. \square

There are two qualitatively different types of rigid motions, corresponding to the two possibilities for the determinant of an orthogonal matrix:

LEMMA 1.59.

If $\mathbf{A} \in O(n)$, then $\det(\mathbf{A}) = 1$ or $\det(\mathbf{A}) = -1$.

PROOF. Since $\mathbf{A}^T \cdot \mathbf{A} = \mathbf{I}$,

$$1 = \det(\mathbf{I}) = \det(\mathbf{A}^T \cdot \mathbf{A}) = \det(\mathbf{A}^T) \cdot \det(\mathbf{A}) = (\det(\mathbf{A}))^2.$$

\square

DEFINITION 1.60.

The rigid motion $f = T_{\mathbf{q}} \circ L_{\mathbf{A}}$ (as in Proposition 1.58) is called **proper** if $\det(\mathbf{A}) = 1$, and **improper** if $\det(\mathbf{A}) = -1$.

EXAMPLE 1.61. If $\mathbf{A} \in O(2)$, then its two columns form an orthonormal basis of \mathbb{R}^2 . Its first column is an arbitrary unit-length vector of \mathbb{R}^2 , which can be written as $\mathbf{x} = (\cos \theta, \sin \theta)$ for some $\theta \in [0, 2\pi)$. The second column is of unit length and orthogonal to the first, which leaves two choices: $R_{90}(\mathbf{x}) = (-\sin \theta, \cos \theta)$ or $-R_{90}(\mathbf{x}) = (\sin \theta, -\cos \theta)$. So we learn that

$$O(2) = \underbrace{\left\{ \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix} \mid \theta \in [0, 2\pi) \right\}}_{\det=1} \cup \underbrace{\left\{ \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix} \mid \theta \in [0, 2\pi) \right\}}_{\det=-1}.$$

If $\mathbf{A} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$, then $L_{\mathbf{A}} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a rotation by the angle θ about the origin, while if $\mathbf{A} = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}$, then $L_{\mathbf{A}}$ is a reflection over the line through the origin that makes the angle $\frac{\theta}{2}$ with the positive x -axis (Exercise 1.71).

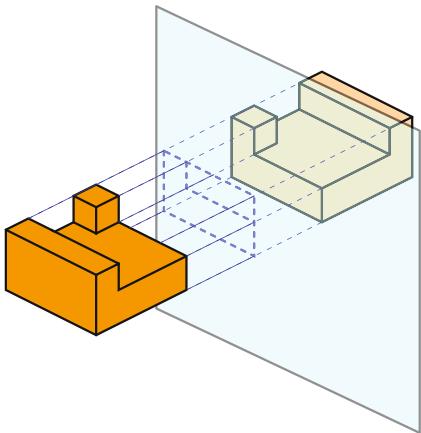


FIGURE 1.31. The reflection across a plane is an improper rigid motion of \mathbb{R}^3 . It cannot be performed to a physical object

rigid motions of \mathbb{R}^3 cannot be performed on solid objects. Reflecting a right boot across a plane would yield a left boot, but there is nothing you can physically do to a right boot to make it become a left boot; see Fig. 1.31

We will next describe an alternative characterization of proper versus improper rigid motions, based on the effect that the motion has on the *orientation* of a basis.

DEFINITION 1.62.

An ordered orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ of \mathbb{R}^n is called **positively oriented** if the orthogonal matrix whose columns are $\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n$ (in that order) has determinant 1, and **negatively oriented** if it has determinant -1 .

In \mathbb{R}^2 , this notion is related to the clockwise/counterclockwise distinction. The ordered orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2\}$ is positively oriented if $\mathbf{v}_2 = R_{90}(\mathbf{v}_1)$, and negatively oriented if $\mathbf{v}_2 = -R_{90}(\mathbf{v}_1)$. One of these must be the case, because \mathbf{v}_2 is of unit length and orthogonal to \mathbf{v}_1 , as discussed in Example 1.61. For example, at a point of a plane curve with positive signed curvature, the

The above example provides a complete description of $O(2)$, and thus a complete description of the rigid motions of \mathbb{R}^2 fixing the origin. What about $O(3)$? It turns out that every *proper* rigid motion of \mathbb{R}^3 fixing the origin equals a rotation about some axis (Exercise 1.73). The *improper* rigid motions fixing the origin include the **antipodal map**, defined as $(x, y, z) \mapsto (-x, -y, -z)$, which can be visualized as the reflection through the origin. They also include reflections across planes; for example, $(x, y, z) \mapsto (x, y, -z)$ is the reflection across the xy -plane. Think of proper rigid motions of \mathbb{R}^3 as physically performable; if you are holding a solid object, you can rotate it or translate it. Improper

orthonormal basis $\{\mathbf{t}, \mathbf{n}\}$ is positively oriented. At a point with negative signed curvature, this basis is negatively oriented.

In \mathbb{R}^3 , this notion is related to the cross product. The ordered orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ is positively oriented (also called *right-handed*) if $\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2$, and negatively oriented (*left-handed*) if $\mathbf{v}_3 = -(\mathbf{v}_1 \times \mathbf{v}_2)$ (Exercise 1.72). One of these choices must be the case, because \mathbf{v}_3 is of unit length and orthogonal to \mathbf{v}_1 and \mathbf{v}_2 . For example, the ordered orthonormal basis $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$ is positively oriented at all times for which it is defined for a regular space curve.

The following lemma says that proper rigid motions preserve the orientation of a basis, while improper rigid motions reverse it:

LEMMA 1.63.

Let $\mathbf{A} \in O(n)$. The orientation of an ordered orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n\}$ of \mathbb{R}^n (whether it is positively or negatively oriented) agrees with that of $\{\mathbf{Av}_1, \mathbf{Av}_2, \dots, \mathbf{Av}_n\}$ if and only if $\det(\mathbf{A}) = 1$.

PROOF. Let \mathbf{B} denote the orthogonal matrix whose columns are $\mathbf{v}_1, \dots, \mathbf{v}_n$. Notice that $\mathbf{A} \cdot \mathbf{B}$ is the orthogonal matrix whose columns are $\mathbf{Av}_1, \dots, \mathbf{Av}_n$. The result now follows from the fact that $\det(\mathbf{A} \cdot \mathbf{B}) = \det(\mathbf{A}) \det(\mathbf{B})$. \square

Applying a rigid motion to a curve might translate it, rotate it, or reflect it, but it should not affect measurements of its essential shape, such as curvature:

PROPOSITION 1.64.

For a regular curve, the following measurements are invariant under proper rigid motions:

- (1) *the curvature of a curve in \mathbb{R}^n ,*
- (2) *the torsion of a space curve,*
- (3) *the signed curvature of a plane curve.*

Improper rigid motions also preserve curvature, but multiply torsion and signed curvature by -1 .

To interpret this proposition, suppose that $\gamma : I \rightarrow \mathbb{R}^n$ is a regular curve, $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is a rigid motion, and $\hat{\gamma} = f \circ \gamma : I \rightarrow \mathbb{R}^n$. If we use hats to denote functions associated with $\hat{\gamma}$, the first claim says that $\hat{\kappa} = \kappa$ (that is, $\hat{\gamma}$ and γ have the same curvature function). When $n = 3$, the second claim says that $\hat{\tau} = \pm \tau$ (and the two curves agree about the times at which torsion is undefined), where the sign depends on whether f is proper or improper. When $n = 2$, the third claim says that $\hat{\kappa}_s = \pm \kappa_s$.

PROOF. Let $\gamma : I \rightarrow \mathbb{R}^n$ be a regular curve. Assume without loss of generality that γ is parametrized by arc length. Let $f = T_{\mathbf{q}} \circ L_{\mathbf{A}}$ be a rigid motion of \mathbb{R}^n (as in Proposition 1.58). Define $\hat{\gamma} = f \circ \gamma : I \rightarrow \mathbb{R}^n$. In other words,

$$\hat{\gamma}(t) = \mathbf{A} \cdot \gamma(t) + \mathbf{q}.$$

We'll use hats to denote functions associated with $\hat{\gamma}$. Pause to verify componentwise that for all $t \in I$, we have

$$\frac{d}{dt}(\mathbf{A} \cdot \gamma(t)) = \mathbf{A} \cdot \gamma'(t).$$

This “constant multiple rule” will be used repeatedly. First notice that

$$\begin{aligned}\hat{\mathbf{v}}(t) &= \frac{d}{dt}\hat{\gamma}(t) = \frac{d}{dt}(\mathbf{A} \cdot \gamma(t) + \mathbf{q}) = \mathbf{A} \cdot \gamma'(t) + \mathbf{0} = \mathbf{A} \cdot \mathbf{v}(t), \\ \hat{\mathbf{a}}(t) &= \frac{d}{dt}(\mathbf{A} \cdot \mathbf{v}(t)) = \mathbf{A} \cdot \mathbf{a}(t).\end{aligned}$$

Suppressing the input variable, this can be summarized as $\hat{\mathbf{v}} = \mathbf{A} \cdot \mathbf{v}$ and $\hat{\mathbf{a}} = \mathbf{A} \cdot \mathbf{a}$. In particular, $|\hat{\mathbf{v}}| = |\mathbf{A} \cdot \mathbf{v}| = |\mathbf{v}| = 1$, so $\hat{\gamma}$ is also parametrized by arc length. The curvature is

$$\hat{\kappa} = |\hat{\mathbf{a}}| = |\mathbf{A} \cdot \mathbf{a}| = |\mathbf{a}| = \kappa,$$

which proves the first claim. For the second claim, first observe that

$$\hat{\mathbf{t}} = \frac{\hat{\mathbf{v}}}{|\hat{\mathbf{v}}|} = \frac{\mathbf{A} \cdot \mathbf{v}}{|\mathbf{A} \cdot \mathbf{v}|} = \frac{\mathbf{A} \cdot \mathbf{v}}{|\mathbf{v}|} = \mathbf{A} \cdot \frac{\mathbf{v}}{|\mathbf{v}|} = \mathbf{A} \cdot \mathbf{t}.$$

Therefore, at every time when $\kappa \neq 0$,

$$\hat{\mathbf{n}} = \frac{\hat{\mathbf{t}}'}{|\hat{\mathbf{t}}'|} = \frac{(\mathbf{A} \cdot \mathbf{t})'}{|(\mathbf{A} \cdot \mathbf{t})'|} = \frac{\mathbf{A} \cdot \mathbf{t}'}{|\mathbf{A} \cdot \mathbf{t}'|} = \frac{\mathbf{A} \cdot \mathbf{t}'}{|\mathbf{t}'|} = \mathbf{A} \cdot \frac{\mathbf{t}'}{|\mathbf{t}'|} = \mathbf{A} \cdot \mathbf{n}.$$

Assume that $n = 3$. Notice that $\{\mathbf{t}, \mathbf{n}, \mathbf{b}\}$ is a positively oriented orthonormal basis of \mathbb{R}^3 . Since \mathbf{A} is orthogonal,

$$\{\mathbf{A} \cdot \mathbf{t}, \mathbf{A} \cdot \mathbf{n}, \mathbf{A} \cdot \mathbf{b}\} = \{\hat{\mathbf{t}}, \hat{\mathbf{n}}, \mathbf{A} \cdot \mathbf{b}\}$$

is an orthonormal basis that is positively oriented if and only if $\det(\mathbf{A}) = 1$. Comparing to the positively oriented orthonormal basis $\{\hat{\mathbf{t}}, \hat{\mathbf{n}}, \hat{\mathbf{b}}\}$, we learn that $\hat{\mathbf{b}} = \pm \mathbf{A} \cdot \mathbf{b}$, with the sign depending on whether $\det(\mathbf{A}) = \pm 1$. Therefore,

$$\langle \hat{\mathbf{b}}', \hat{\mathbf{n}} \rangle = \langle (\pm \mathbf{A} \cdot \mathbf{b})', \mathbf{A} \cdot \mathbf{n} \rangle = \pm \langle \mathbf{A} \cdot \mathbf{b}', \mathbf{A} \cdot \mathbf{n} \rangle = \pm \langle \mathbf{b}', \mathbf{n} \rangle.$$

Thus, $\hat{\tau} = \tau$ if $\det(\mathbf{A}) > 0$, and $\hat{\tau} = -\tau$ if $\det(\mathbf{A}) < 0$.

For the third claim, assume that $n = 2$. Since $\{\mathbf{v}, R_{90}(\mathbf{v})\}$ is a positively oriented orthonormal basis and \mathbf{A} is orthogonal,

$$\{\mathbf{A} \cdot \mathbf{v}, \mathbf{A} \cdot R_{90}(\mathbf{v})\} = \{\hat{\mathbf{v}}, \mathbf{A} \cdot R_{90}(\mathbf{v})\}$$

is an orthonormal basis that is positively oriented if and only if $\det(\mathbf{A}) > 0$. Comparing to the positively oriented orthonormal basis $\{\hat{\mathbf{v}}, R_{90}(\hat{\mathbf{v}})\}$, we learn

that $R_{90}(\hat{\mathbf{v}}) = \pm \mathbf{A} \cdot R_{90}(\mathbf{v})$, with the sign depending on whether $\det(\mathbf{A}) = \pm 1$. So the signed curvature is

$$\hat{\kappa}_s = \langle \hat{\mathbf{a}}, R_{90}(\hat{\mathbf{v}}) \rangle = \langle \mathbf{A} \cdot \mathbf{a}, \pm \mathbf{A} \cdot R_{90}(\mathbf{v}) \rangle = \pm \langle \mathbf{a}, R_{90}(\mathbf{v}) \rangle = \pm \kappa_s.$$

Thus, $\hat{\kappa}_s = \kappa_s$ if $\det(\mathbf{A}) > 0$, and $\hat{\kappa}_s = -\kappa_s$ if $\det(\mathbf{A}) < 0$. \square

Figure 1.32 illustrates how an improper rigid motion of \mathbb{R}^2 will change the sign of the signed curvature. To visualize how an improper rigid motion of \mathbb{R}^3 will change the sign of the torsion, look back at Fig. 1.28 or Fig. 1.30 and imagine reflecting across the osculating plane.

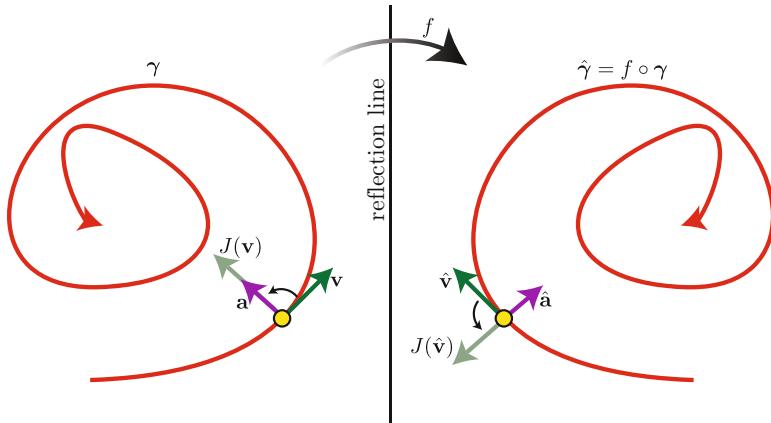


FIGURE 1.32. A reflection reverses the sign of κ_s

We end this section with two fundamental theorems:

THEOREM 1.65 (The Fundamental theorems of plane and space curves).

- (1) *If $I \subset \mathbb{R}$ is an interval and $\kappa_s : I \rightarrow \mathbb{R}$ is a smooth function, then there exists a unit-speed plane curve $\gamma : I \rightarrow \mathbb{R}^2$ whose signed curvature function equals κ_s . If $\gamma, \hat{\gamma} : I \rightarrow \mathbb{R}^2$ are two such curves, then there exists a proper rigid motion, f , of \mathbb{R}^2 such that $\hat{\gamma} = f \circ \gamma$.*
- (2) *If $I \subset \mathbb{R}$ is an interval and $\kappa, \tau : I \rightarrow \mathbb{R}$ is a pair of a smooth functions with $\kappa > 0$, then there exists a unit-speed space curve $\gamma : I \rightarrow \mathbb{R}^3$ whose curvature function equals κ and whose torsion function equals τ . If $\gamma, \hat{\gamma} : I \rightarrow \mathbb{R}^3$ are two such curves, then there exists a proper rigid motion, f , of \mathbb{R}^3 such that $\hat{\gamma} = f \circ \gamma$.*

We will only prove (1), since the proof of (2) requires differential equation prerequisites.

PROOF OF (1). Choose any $t_0 \in I$. For convenience, we'll construct γ with initial conditions $\gamma(t_0) = (0, 0)$ and $\gamma'(t_0) = (1, 0)$. For this, define the angle function $\theta : I \rightarrow \mathbb{R}$ as

$$\theta(t) = \int_{t_0}^t \kappa_s(u) du + 0.$$

Next define the velocity function as $\mathbf{v}(t) = (\cos \theta(t), \sin \theta(t))$. Finally, define $\gamma : I \rightarrow \mathbb{R}^2$ as

$$\gamma(t) = \int_{t_0}^t \mathbf{v}(u) du + (0, 0).$$

This vector equation just means that the components of γ are defined to be the integrals of the components of \mathbf{v} .

As desired, γ is a unit-speed plane curve for which θ is an angle function. Its signed curvature function is $\kappa_s = \theta'$, and its initial conditions are $\gamma(t_0) = (0, 0)$ and $\mathbf{v}(t_0) = (1, 0)$. But notice that the red terms above could be altered in order to give γ arbitrary initial conditions.

For the second statement, let $\gamma, \hat{\gamma} : I \rightarrow \mathbb{R}^2$ be two unit-speed plane curves each of whose signed curvature function equals κ_s . Each angle function (denoted by θ and $\hat{\theta}$ respectively) is an antiderivative of κ_s ; since antiderivatives are unique up to an additive constant, we have $\hat{\theta} = \theta + \theta_0$ for some $\theta_0 \in \mathbb{R}$. Let $\mathbf{v}, \hat{\mathbf{v}}$ denote their velocity functions. Notice that

$$\begin{aligned} \hat{\mathbf{v}}(t) &= \left(\cos \hat{\theta}(t), \sin \hat{\theta}(t) \right) = (\cos(\theta(t) + \theta_0), \sin(\theta(t) + \theta_0)) \\ &= \underbrace{L_{\mathbf{A}}(\cos \theta(t), \sin \theta(t))}_{\text{Exercise 1.71(1)}} = L_{\mathbf{A}}(\mathbf{v}(t)), \end{aligned}$$

where $\mathbf{A} = \begin{pmatrix} \cos(\theta_0) & -\sin(\theta_0) \\ \sin(\theta_0) & \cos(\theta_0) \end{pmatrix}$. Consider the rigid motion $f = T_{\mathbf{q}} \circ L_{\mathbf{A}}$ and the curve $\beta = f \circ \gamma$, where $\mathbf{q} \in \mathbb{R}^2$ is chosen such that $\beta(t_0) = \hat{\gamma}(t_0)$. Since β and $\hat{\gamma}$ have the same derivative function and the same initial value, the uniqueness of antiderivatives implies that $\beta = \hat{\gamma}$. \square

EXERCISES

EXERCISE 1.70. Prove Lemma 1.56.

EXERCISE 1.71. Prove the following claims from Example 1.61:

- (1) If $\mathbf{A} = \begin{pmatrix} \cos \theta & -\sin \theta \\ \sin \theta & \cos \theta \end{pmatrix}$, then $L_{\mathbf{A}} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a rotation by the angle θ about the origin.
- (2) If $\mathbf{A} = \begin{pmatrix} \cos \theta & \sin \theta \\ \sin \theta & -\cos \theta \end{pmatrix}$, then $L_{\mathbf{A}}$ is a reflection over the line through the origin that makes the angle $\frac{\theta}{2}$ with the positive x -axis.

EXERCISE 1.72. Prove that the ordered orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of \mathbb{R}^3 is positively oriented if $\mathbf{v}_3 = \mathbf{v}_1 \times \mathbf{v}_2$, and negatively oriented if $\mathbf{v}_3 = -\mathbf{v}_1 \times \mathbf{v}_2$.

EXERCISE 1.73. Prove that every proper rigid motion, f , of \mathbb{R}^3 that fixes the origin is a rotation about some axis.

HINT: Write $f = L_{\mathbf{A}}$, where $\mathbf{A} \in O(3)$ with $\det(\mathbf{A}) = 1$. Notice that \mathbf{A} has a real eigenvalue $\lambda \in \mathbb{R}$, because its characteristic polynomial is cubic. Since \mathbf{A} is orthogonal, $\lambda = \pm 1$. Let \mathbf{v}_1 denote a corresponding unit-length eigenvector, and complete it to an orthonormal basis $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$ of \mathbb{R}^3 . Show that the matrix representing f with respect to this basis has the form $\begin{pmatrix} \lambda & 0 & 0 \\ 0 & a & b \\ 0 & c & d \end{pmatrix}$, where $\mathbf{B} = \begin{pmatrix} a & b \\ c & d \end{pmatrix} \in O(2)$. If $\lambda = 1$, then $\det(\mathbf{B}) = 1$, and Example 1.61 applied to \mathbf{B} gives that f is a rotation about $\text{span}\{\mathbf{v}_1\}$. On the other hand, if $\lambda = -1$, then $\det(\mathbf{B}) = -1$, and Example 1.61 applied to \mathbf{B} gives that a vector of $\text{span}\{\mathbf{v}_2, \mathbf{v}_3\}$ in the direction of the reflection line is an eigenvector with eigenvalue +1, so the above argument can be started over with this vector playing the role of \mathbf{v}_1 .

EXERCISE 1.74. Let $\mathbf{B} \in O(n)$ with $\det(\mathbf{B}) = -1$. Prove that every member of $O(n)$ with negative determinant has the form $\mathbf{A} \cdot \mathbf{B}$ for some $\mathbf{A} \in O(n)$ with $\det(\mathbf{A}) = 1$. In other words, you get all of the improper linear rigid motions by composing a single one with all of the proper linear rigid motions.

EXERCISE 1.75. For practice with the concepts involved, redo the proof of Proposition 1.64 without assuming that γ is parametrized by arc length.

EXERCISE 1.76. Let $\gamma : \mathbb{R} \rightarrow \mathbb{R}^3$ be a regular space curve with component functions denoted by $\gamma(t) = (x(t), y(t), z(t))$. Define $\hat{\gamma} : \mathbb{R} \rightarrow \mathbb{R}^3$ as $\hat{\gamma}(t) = (z(t), x(t), -y(t))$. Describe how the curvature and torsion functions of $\hat{\gamma}$ are related to the curvature and torsion functions of γ .

EXERCISE 1.77. Show that the helices denoted by γ and β in Exercise 1.64 (on page 48) are related by an improper rigid motion, and use this to account for the sign difference in their torsion functions.

EXERCISE 1.78. For $c > 0$, consider the *dilation* map $d_c : \mathbb{R}^n \rightarrow \mathbb{R}^n$ defined as $d_c(\mathbf{p}) = c\mathbf{p}$. If $\gamma : I \rightarrow \mathbb{R}^n$ is a regular curve and $\hat{\gamma} = d_c \circ \gamma$, how are the curvature functions of $\hat{\gamma}$ and γ related? In the case $n = 3$, how are the torsion functions related? In the case $n = 2$, how are the signed curvature functions related? Would any of these answers change if $c < 0$?

EXERCISE 1.79. For $\mathbf{A} \in M_n$, prove that the following are equivalent:

- (1) \mathbf{A} is orthogonal.
- (2) The rows of \mathbf{A} form an orthonormal basis of \mathbb{R}^n .
- (3) $\mathbf{A} \cdot \mathbf{A}^T = \mathbf{I}$.

EXERCISE 1.80. A **permutation matrix** means an orthogonal matrix whose entries are all 0 or 1. For example, $\mathbf{A} = \begin{pmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \in O(3)$ is a permutation matrix; since $L_{\mathbf{A}}(x, y, z) = (z, x, y)$, this matrix permutes the symbols x, y, z , or equivalently, it permutes the members $\mathbf{e}_1, \mathbf{e}_2, \mathbf{e}_3$ of the standard orthonormal basis of \mathbb{R}^3 . Prove that for every integer $n > 1$, there are exactly $n!$ different $n \times n$ permutation matrices, and that exactly half of them have positive determinant.

EXERCISE 1.81. Give examples of a regular plane curve $\gamma : I \rightarrow \mathbb{R}^2$ and a rigid motion f of \mathbb{R}^2 such that $f \circ \gamma$ is a reparametrization of γ . Include an example in which the curve is closed and one in which it is not closed. Include examples in which the rigid motion is proper and in which it is improper.

EXERCISE 1.82. Let $\kappa_s : \mathbb{R} \rightarrow \mathbb{R}$ be a smooth function that is periodic with period L . Let $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$ be a plane curve whose signed curvature function equals κ_s , constructed as in the proof of Theorem 1.65.

- (1) Give examples in which γ is not periodic.
- (2) Find necessary and sufficient conditions for γ on $[0, L]$ to be a closed curve.

EXERCISE 1.83.

- (1) If $\mathbf{A}, \mathbf{B} : I \rightarrow M_n$ are parametrized curves (which makes sense if you make the identification $M_n \cong \mathbb{R}^{n^2}$), prove the following product rule for matrix multiplication:

$$\frac{d}{dt}(\mathbf{A}(t) \cdot \mathbf{B}(t)) = \mathbf{A}'(t) \cdot \mathbf{B}(t) + \mathbf{A}(t) \cdot \mathbf{B}'(t),$$

where all derivatives are interpreted as componentwise derivatives.

- (2) A matrix $\mathbf{M} \in M_n$ is called **skew-symmetric** if $\mathbf{M} = -\mathbf{M}^T$. Suppose that $\mathbf{A} : I \rightarrow O(n)$ is a parametrized curve of orthogonal matrices with $\mathbf{A}(0) = \mathbf{I}$ (the identity matrix). Prove that $\mathbf{A}'(0)$ is skew-symmetric. *HINT: Differentiate the equation $\mathbf{A}(t)^T \cdot \mathbf{A}(t) = \mathbf{I}$.*

EXERCISE 1.84. With a computer algebra system, implement the construction in the proof of Theorem 1.65 to graph a plane curve whose signed curvature function is:

- (1) $\kappa_s(t) = -t$.
- (2) $\kappa_s(t) = -2t^2$.
- (3) $\kappa_s(t) = c \cdot \sin t$ for several choices of $c > 0$ (and find a value of c for which the curve appears to be periodic).
- (4) $\kappa_s(t) = t \cdot \sin t$.
- (5) $\kappa_s(t) = e^t$.

EXERCISE 1.85. With a computer algebra system:

- (1) Write a procedure for generating a random element of $O(3)$.
- (2) The curve $\gamma(t) = (\cos t, \sin t, 0)$, $t \in [0, 2\pi]$, parameterizes the equator of S^2 . Plot together curves of the form $L_{\mathbf{A}} \circ \gamma$ for a large number of randomly chosen $A \in O(3)$.

HINT for (1): Construct a matrix whose entries are random numbers in $[-1, 1]$ and then apply the Gram–Schmidt process to its columns, as described in Exercise 1.20 on page 16.



9. Overview of Curvature Formulas

Below is an overview of our various formulas for computing the curvature or signed curvature of a regular curve (a few of which were found in the exercises):

- (1) For a GENERAL CURVE $\gamma : I \rightarrow \mathbb{R}^n$:

$$\kappa = \frac{|\mathbf{t}'|}{|\mathbf{v}|} = \frac{|\mathbf{a}^\perp|}{\underbrace{|\mathbf{v}|^2}_{\text{if unit speed}}} = |\mathbf{a}|.$$

- (2) For a PLANE CURVE $\gamma = (x, y) : I \rightarrow \mathbb{R}^2$:

$$\kappa_s = \theta' = \frac{\langle \mathbf{a}, R_{90}(\mathbf{v}) \rangle}{|\mathbf{v}|^3} = \frac{x'y'' - x''y'}{((x')^2 + (y')^2)^{\frac{3}{2}}} = \underbrace{\frac{f''}{(1 + (f')^2)^{3/2}}}_{\text{if } \gamma(t) = (t, f(t))}.$$

- (3) For a SPACE CURVE $\gamma : I \rightarrow \mathbb{R}^3$:

$$\kappa = \frac{|\mathbf{v} \times \mathbf{a}|}{|\mathbf{v}|^3}.$$



A planimeter is a drafting instrument that measures the area enclosed in a region. Its design is based on Green's theorem, one of several global results about curves presented in this chapter.

Additional Topics in Curves

This chapter presents several excursions that delve more deeply into the geometry of curves, including some of the famous theorems in the field. The theory of curves is an old and extremely well developed mathematical topic. Our aim is simply to describe a few fundamental and interesting highlights.

1. Theorems of Hopf and Jordan

This section is devoted to proving the following two historically significant global theorems:

THEOREM 2.1.

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let $C = \gamma([a, b])$ denote its trace.

- (1) (**Hopf's Umlaufsatz**) The rotation index of γ is either 1 or -1 .
- (2) (**The Jordan Curve Theorem**) $\mathbb{R}^2 - C = \{\mathbf{p} \in \mathbb{R}^2 \mid \mathbf{p} \notin C\}$ has exactly two path-connected components. Their common boundary is C . One component (which we call the **interior**) is bounded, while the other (which we call the **exterior**) is unbounded.

Each theorem provides a method to meaningfully distinguish between the two possible orientations of γ , and the methods they provide are equivalent:

DEFINITION 2.2.

A simple closed plane curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is called **positively oriented** if it satisfies the following equivalent conditions:

- (1) The rotation index of γ equals 1.
- (2) The interior is on one's left as one traverses γ ; more precisely, for each $t \in [a, b]$, $R_{90}(\gamma'(t))$ points toward the interior in the sense that there exists $\delta > 0$ such that $\gamma(t) + sR_{90}(\gamma'(t))$ lies in the interior for all $s \in (0, \delta)$.

Otherwise, γ is **negatively oriented**, in which case its rotation index equals -1 , and $R_{90}(\gamma'(t))$ points toward the exterior for all $t \in [a, b]$.

The equivalence of these two conditions will follow from ideas in the proofs of Hopf's Umlaufsatz and the Jordan curve theorem (Exercise 2.1).

The curve in Fig. 2.1 hints that Theorem 2.1 is not as obvious as it might at first appear. Although this curve performs many full clockwise and many full counterclockwise turns, most of them cancel each other, leaving a net counterclockwise rotation of one turn. This curve is therefore positively oriented, which is more easily verified by observing that its interior is on its left.

The remainder of this section is devoted to (1) sketching the proofs of these two fundamental theorems, which could be skipped on a first read, and (2) generalizing Hopf's Umlaufsatz to piecewise-regular curves, which is an important prerequisite for Chap. 6.

Recall that the velocity function of a unit-speed closed plane curve $\gamma : [a, b] \rightarrow \mathbb{R}^2$ can be regarded as a function $\mathbf{v} : [a, b] \rightarrow S^1$ with $\mathbf{v}(a) = \mathbf{v}(b)$, where $S^1 = \{(\cos \theta, \sin \theta) \mid \theta \in \mathbb{R}\}$. This viewpoint allowed us in the previous chapter to construct a global angle function $\theta : [a, b] \rightarrow \mathbb{R}$, contrived so that $\mathbf{v}(t) = (\cos \theta(t), \sin \theta(t))$ for all $t \in [a, b]$. From this, we defined the rotation index of the plane curve as $\frac{1}{2\pi}(\theta(b) - \theta(a))$.

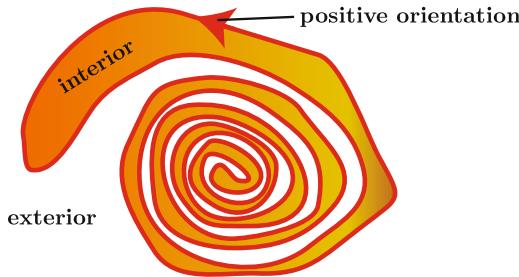


FIGURE 2.1. Perhaps it's not obvious how to prove Theorem 2.1

Now forget about the plane curve, and imagine instead that you began with an arbitrary continuous function from $[a, b]$ to S^1 whose values at a and b agree. Don't assume that it is the velocity function of anything. We claim that the above steps still apply:

PROPOSITION AND DEFINITION 2.3.

If $\mathbf{f} : [a, b] \rightarrow S^1$ is a continuous function with $\mathbf{f}(a) = \mathbf{f}(b)$, then there exists a continuous **angle function** $\theta : [a, b] \rightarrow \mathbb{R}$ such that for all $t \in [a, b]$, we have

$$\mathbf{f}(t) = (\cos \theta(t), \sin \theta(t)).$$

This function is unique up to adding an integer multiple of 2π . The **degree** of \mathbf{f} is defined as the integer $\frac{1}{2\pi}(\theta(b) - \theta(a))$.

If \mathbf{f} is smooth, then the claim follows from Proposition 1.39 (on page 35), since integrating \mathbf{f} yields a unit-speed plane curve whose velocity function is \mathbf{f} . The proof for continuous functions is outlined in Exercise 2.2. In summary, the *degree* of a continuous function $\mathbf{f} : [a, b] \rightarrow S^1$ with $\mathbf{f}(a) = \mathbf{f}(b)$ is an integer that represents roughly the number of times the domain is wrapped counterclockwise around the circle. Notice that the rotation index of a closed plane curve γ (as defined in Exercise 1.55 on page 40) equals the degree of its unit tangent function $t \mapsto \mathbf{t}(t)$.

We will repeatedly use the idea that two functions from $[a, b]$ into S^1 with sufficiently close outputs must have the same degree. In fact, if their outputs never point in opposite directions, then they must have the same degree:

LEMMA 2.4.

Let $\mathbf{f}_1, \mathbf{f}_2 : [a, b] \rightarrow S^1$ be continuous functions with $\mathbf{f}_1(a) = \mathbf{f}_1(b)$ and $\mathbf{f}_2(a) = \mathbf{f}_2(b)$. If \mathbf{f}_1 and \mathbf{f}_2 have different degrees, then $\mathbf{f}_1(t_0) = -\mathbf{f}_2(t_0)$ for some $t_0 \in [a, b]$.

PROOF. Let $\theta_1, \theta_2 : [a, b] \rightarrow \mathbb{R}$ be angle functions for \mathbf{f}_1 and \mathbf{f}_2 . Consider the difference $\delta(t) = \theta_2(t) - \theta_1(t)$. Since the degrees are different,

$$|\delta(b) - \delta(a)| = \left| \underbrace{(\theta_2(b) - \theta_2(a))}_{2\pi(\text{degree } \mathbf{f}_2)} - \underbrace{(\theta_1(b) - \theta_1(a))}_{2\pi(\text{degree } \mathbf{f}_1)} \right| \geq 2\pi.$$

Since δ has a net change of at least 2π , there must be an *odd* integer multiple of π between $\delta(a)$ and $\delta(b)$. The intermediate value theorem implies that δ achieves this value for some $t_0 \in [a, b]$, so $\mathbf{f}_1(t_0) = -\mathbf{f}_2(t_0)$. \square

PROOF OF HOPF'S UMLAUFSSATZ. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let C denote its trace. Let $\mathbf{p} \in C$ be a point such that C is entirely on one side of the tangent line, L , to C at \mathbf{p} . One can find such a point by considering a circle (centered anywhere in \mathbb{R}^2) with radius large enough to contain C , and then shrinking the radius until the circle first touches C . The point at which it first touches C will have the desired property.

We can assume without loss of generality that γ is parametrized by arc length with $\gamma(a) = \mathbf{p}$. Consider the triangle

$$T = \{(t_1, t_2) \mid a \leq t_1 \leq t_2 \leq b\}.$$

Define the function $\psi : T \rightarrow S^1$ as follows:

$$\psi(t_1, t_2) = \begin{cases} \gamma'(t_1) & \text{if } t_1 = t_2 \\ \frac{\gamma(t_2) - \gamma(t_1)}{|\gamma(t_2) - \gamma(t_1)|} & \text{if } t_1 \neq t_2 \text{ and } \{t_1, t_2\} \neq \{a, b\}, \\ -\gamma'(a) & \text{if } \{t_1, t_2\} = \{a, b\}. \end{cases}$$

For most inputs, $\psi(t_1, t_2)$ is the unit vector pointing in the direction from $\gamma(t_1)$ to $\gamma(t_2)$. The rest of the definition just ensures that ψ is continuous. For example, according to Proposition 1.7 (on page 4), the correct way to extend ψ continuously to a point (t, t) on the hypotenuse of T is $\psi(t, t) = \gamma'(t)$.

Let $\alpha_0 : [0, 1] \rightarrow T$ be a parametrization of the line segment from (a, a) to (b, b) . Let $\alpha_1 : [0, 1] \rightarrow T$ be a parametrization of the line segment from (a, a) to (a, b) followed by the line segment from (a, b) to (b, b) . It is possible to interpolate continuously between α_0 and α_1 by a family of paths, $\alpha_s : [0, 1] \rightarrow T$, $s \in [0, 1]$, each of which goes from (a, a) to (b, b) ; see Fig. 2.2 (left). Here “continuously” means that $(s, t) \mapsto \alpha_s(t)$ is a continuous function from $[0, 1] \times [0, 1]$ to T .

For each $s \in [0, 1]$, let $D(s)$ denote the degree of $\psi \circ \alpha_s : [0, 1] \rightarrow S^1$. Lemma 2.4 can be used to show that $s \mapsto D(s)$ is locally constant and therefore continuous on $[0, 1]$. Since D is integer-valued and continuous, it follows from Proposition A.19 of the appendix (on page 353) that D must be constant on $[0, 1]$, so $D(1) = D(0)$.

By definition, $D(0)$ equals the degree of the unit tangent function of γ , which equals the rotation index of γ . It remains to explain why $D(1)$ equals 1 or -1 . In Fig. 2.2 (right), as α_1 first goes from (a, a) to (a, b) , the path $\psi \circ \alpha_1$ follows the blue vectors, tracing the top half of S^1 counterclockwise. Then as α_1 goes from (a, b) to (b, b) , the path $\psi \circ \alpha_1$ follows the negatives

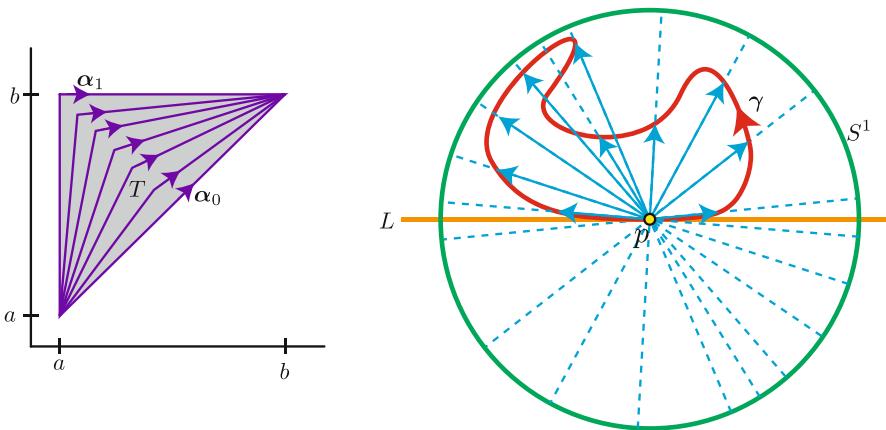


FIGURE 2.2. The proof of Hopf's Umlaufsatz

of the blue vectors, tracing the bottom half of S^1 counterclockwise. Thus, $D(1) = 1$. If γ had the other orientation, then $D(1) = -1$. This completes the proof.¹ \square

For the next proof, we require the idea of a *tubular neighborhood*. Suppose that $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is a simple closed plane curve. For small $\epsilon > 0$, consider the function $\varphi : (-\epsilon, \epsilon) \times [a, b] \rightarrow \mathbb{R}^2$ defined as

$$\varphi(s, t) = \gamma(t) + s \cdot R_{90}(\mathbf{v}(t)).$$

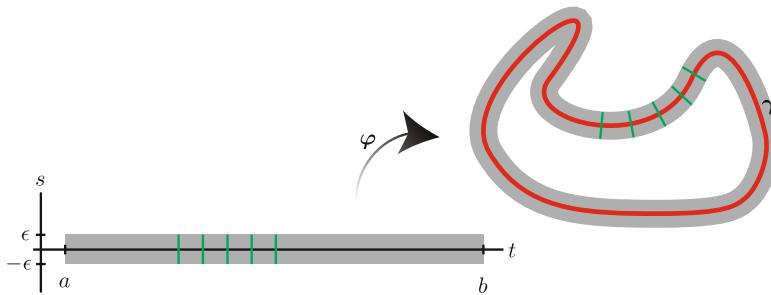


FIGURE 2.3. A tubular neighborhood

For fixed t_0 , the function $s \mapsto \varphi(t_0, s)$ parametrizes a small line segment that crosses the trace of γ orthogonally; in Fig. 2.3, it is shown in green for five choices of t_0 . The important fact is that these green lines do not intersect each other. In other words, we have the following:

PROPOSITION 2.5.

For sufficiently small $\epsilon > 0$, φ is injective.

¹To help visualize the function $\psi \circ \alpha_s$ and the degree of this function, a nice animation is available at <http://www.mathematik.com/Hopf/index.html>

We will postpone the proof of this claim until Exercise 3.11, after we discuss the inverse function theorem. The phenomenon is both local and global. The value ϵ must be chosen small enough to ensure that the green lines remain disjoint locally as the curve bends sharply, and also globally as the curve loops back close to itself. The image of φ is called a **tubular neighborhood** of γ .

If the trace of γ is removed from the tubular neighborhood, then what is left has two path-connected components, namely $\varphi((-\epsilon, 0) \times [a, b])$ and $\varphi((0, \epsilon) \times [a, b])$. Each is path-connected, because φ identifies it with a rectangle.

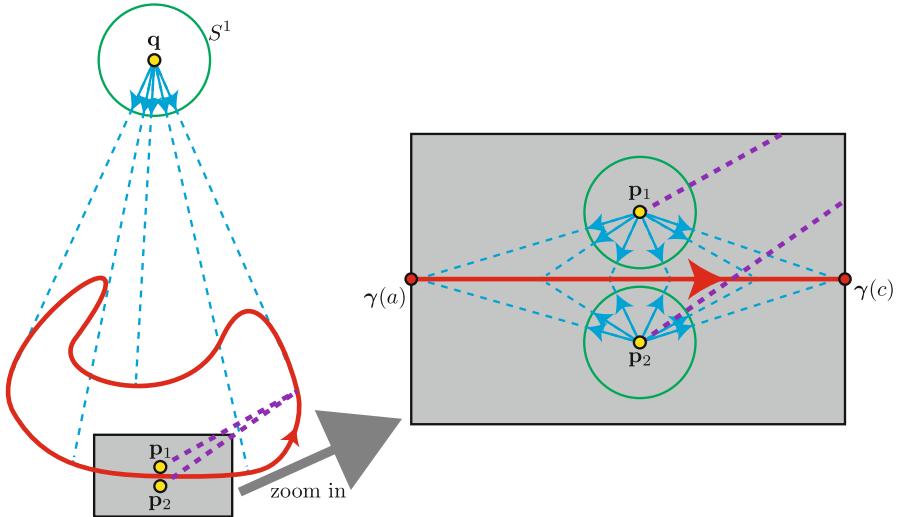
PROOF OF THE JORDAN CURVE THEOREM. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a simple closed plane curve. Let C denote its trace. For any $\mathbf{p} \in \mathbb{R}^2 - C$, consider the function $\mathbf{f}_\mathbf{p} : [a, b] \rightarrow S^1$ defined as

$$\mathbf{f}_\mathbf{p}(t) = \frac{\gamma(t) - \mathbf{p}}{|\gamma(t) - \mathbf{p}|}.$$

Let $W(\mathbf{p})$ denote the degree of $\mathbf{f}_\mathbf{p}$. Intuitively, if you stand at \mathbf{p} while keeping your finger pointing at a friend who traverses C , then $W(\mathbf{p})$ is the net number of counterclockwise rotations that this activity forces you to perform. Lemma 2.4 can be used to verify that W is locally constant and therefore continuous on its domain $\mathbb{R}^2 - C$. It is constant on every path-connected component of this domain, because along every path in the domain, W changes continuously but is also integer-valued, so Proposition A.19 from the appendix (on page 353) implies that it must be constant. Our aim is to show that $\mathbb{R}^2 - C$ has exactly two path-connected components, one on which $W = 0$ and the other on which $W = 1$ or $W = -1$ (depending on the orientation of γ).

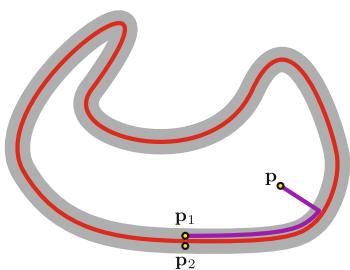
If \mathbf{q} is sufficiently far from C , like the point labeled “ \mathbf{q} ” in Fig. 2.4, then $\mathbf{f}_\mathbf{q}$ is not surjective, because its image is constrained to an arc of S^1 , so Lemma 2.4 implies that $W(\mathbf{q})$ equals the degree of a constant function, which is 0. We will demonstrate next that W also attains at least one nonzero value, so $\mathbb{R}^2 - C$ has at least two connected components.

Imagine zooming in with sufficient magnification at a point of C so that the zooming window lies within a tubular neighborhood, and C is well approximated by its tangent line within this window. Choose a pair of points $\mathbf{p}_1, \mathbf{p}_2$ in this window that are close to each other but are on opposite sides of C . We claim that the window and the points can be chosen such that $|W(\mathbf{p}_1) - W(\mathbf{p}_2)| = 1$. To prove this, let ϵ denote any quantity that approaches zero as the zooming window shrinks and as $|\mathbf{p}_1 - \mathbf{p}_2|$ becomes small relative to the size of the window. Just to make the discussion more specific, we assume that the curve is positioned and oriented as in Fig. 2.4 (with C horizontal traversed left to right, \mathbf{p}_1 above, and \mathbf{p}_2 below) and is parametrized so that $\gamma(t)$ lies in this window for time parameters $t \in [a, c]$ (where $a < c < b$). Restricted to $[a, c]$, both $\mathbf{f}_{\mathbf{p}_1}$ and $\mathbf{f}_{\mathbf{p}_2}$ begin within distance ϵ of $(-1, 0)$ and end within distance ϵ of $(1, 0)$; $\mathbf{f}_{\mathbf{p}_1}$ approximately covers the bottom half of

FIGURE 2.4. $W(\mathbf{q}) = 0$, while $|W(\mathbf{p}_1) - W(\mathbf{p}_2)| = 1$

S^1 counterclockwise, while $\mathbf{f}_{\mathbf{p}_2}$ approximately covers the top half of S^1 clockwise. Furthermore, for $t \in [c, b]$ we have $|\mathbf{f}_{\mathbf{p}_1}(t) - \mathbf{f}_{\mathbf{p}_2}(t)| < \epsilon$, as indicated by the dashed purple lines in Fig. 2.4.

We will perform some small perturbations to $\mathbf{f}_{\mathbf{p}_1}$ and $\mathbf{f}_{\mathbf{p}_2}$ that (by Lemma 2.4) do not alter the degree of either function. First, we can modify both functions on $[a, c]$ so that they begin exactly at $\mathbf{f}_{\mathbf{p}_1}(a) = \mathbf{f}_{\mathbf{p}_2}(a) = (-1, 0)$ and end exactly at $\mathbf{f}_{\mathbf{p}_1}(c) = \mathbf{f}_{\mathbf{p}_2}(c) = (1, 0)$. Next, we can redefine $\mathbf{f}_{\mathbf{p}_2}$ to equal $\mathbf{f}_{\mathbf{p}_1}$ on $[c, b]$. After these modifications, their degrees differ by one. To understand why, imagine traversing $\mathbf{f}_{\mathbf{p}_1}$ followed by the reverse orientation of $\mathbf{f}_{\mathbf{p}_2}$. This path, denoted by $\mathbf{f}_{\mathbf{p}_1} - \mathbf{f}_{\mathbf{p}_2}$, traverses the bottom half of S^1 counterclockwise, then does something else, then does that same something else in reverse, then traverses the top half of S^1 counterclockwise. The net result is one counterclockwise rotation. Since the degree of $\mathbf{f}_{\mathbf{p}_1} - \mathbf{f}_{\mathbf{p}_2}$ equals 1, it follows that $\text{degree}(\mathbf{f}_{\mathbf{p}_1}) - \text{degree}(\mathbf{f}_{\mathbf{p}_2}) = 1$.

FIGURE 2.5. Every $\mathbf{p} \in \mathbb{R}^2 - C$ can be joined to either \mathbf{p}_1 or \mathbf{p}_2 with a path that avoids C

We now know that \mathbf{p}_1 and \mathbf{p}_2 are in different path-connected components of $\mathbb{R}^2 - C$. We claim that these are the only components. In other words, every other point $\mathbf{p} \in \mathbb{R}^2 - C$ can be connected to either \mathbf{p}_1 or \mathbf{p}_2 by a continuous path in $\mathbb{R}^2 - C$. To see this, choose a shortest path from \mathbf{p} to C . Before reaching C , this path will reach a fixed tubular neighborhood of C , inside of which it can be connected to \mathbf{p}_1 or \mathbf{p}_2 ; see Fig. 2.5.

Thus, $\mathbb{R}^2 - C$ has exactly two path-connected components. Let $B \subset \mathbb{R}^2$ denote any ball large enough to contain C . Clearly, one component of $\mathbb{R}^2 - C$ contains the complement of B , and is therefore unbounded, so the other component is contained in B and is therefore bounded. \square

The remainder of this section is devoted to generalizing its main theorems. The Jordan curve theorem remains true if $\gamma : [a, b] \rightarrow \mathbb{R}^2$ is only a *continuous* function with $\gamma(a) = \gamma(b)$ that is one-to-one on $[a, b]$, but the proof is more difficult in this setting.

Hopf's Umlaufsatz does not make sense when γ is only continuous, but it can at least be generalized to *piecewise-regular* curves:

DEFINITION 2.6.

A *piecewise-regular curve* in \mathbb{R}^n is a continuous function $\gamma : [a, b] \rightarrow \mathbb{R}^n$ with a partition, $a = t_0 < t_1 < \dots < t_n = b$, such that the restriction, γ_i , of γ to each subinterval $[t_i, t_{i+1}]$ is a regular curve. It is called **closed** if additionally $\gamma(a) = \gamma(b)$, and **simple** if γ is one-to-one on the domain $[a, b]$. It is said to be **of unit speed** if each γ_i is of unit speed.

In other words, there might be finitely many times at which γ is only continuous but not smooth. The definition of “closed” does not require the derivatives of γ to agree at a and b ; this allows the possibility that $t = a$ might correspond to one of the nonsmooth points.

A piecewise-regular simple closed *plane* curve γ is called **positively oriented** if $R_{90}(\gamma'(t))$ points toward the interior for all values of t that correspond to smooth points (all values except the partition endpoints), as in Fig. 2.6. Otherwise, it is called **negatively oriented**.

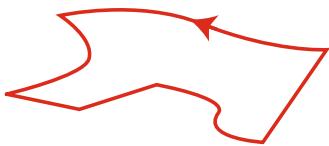


FIGURE 2.6. A positively oriented piecewise-regular simple closed plane curve

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a piecewise-regular *plane* curve with partition denoted by $a = t_0 < t_1 < \dots < t_n = b$. Each nonsmooth point $\gamma(t_i)$ is called a **corner** of γ . At this corner, there are *two* velocity vectors, coming from the left- and right-hand limits:

$$\mathbf{v}^-(t_i) = \lim_{h \rightarrow 0^-} \frac{\gamma(t_i + h) - \gamma(t_i)}{h} = \lim_{t \rightarrow t_i^-} \gamma'(t),$$

$$\mathbf{v}^+(t_i) = \lim_{h \rightarrow 0^+} \frac{\gamma(t_i + h) - \gamma(t_i)}{h} = \lim_{t \rightarrow t_i^+} \gamma'(t).$$

By the regularity hypothesis, neither is zero. The **signed angle** at $\gamma(t_i)$, denoted by $\alpha_i \in [-\pi, \pi]$, is defined such that its absolute value equals the smallest determination of the angle between $\mathbf{v}^-(t_i)$ and $\mathbf{v}^+(t_i)$. The sign of

α_i is defined to be positive if $\mathbf{v}^+(t_i)$ is a counterclockwise rotation of $\mathbf{v}^-(t_i)$ through this angle (and to be negative if it is clockwise); see Fig. 2.7. Notice that reversing the orientation of γ would change the sign of the signed angle at each corner, but would not affect the absolute value.

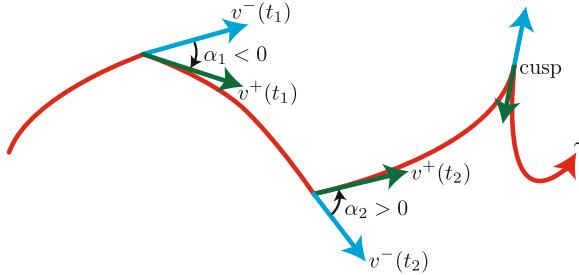


FIGURE 2.7. A piecewise-regular plane curve with three corners

The corner $\gamma(t_i)$ is called a **cusp** if $\mathbf{v}^+(t_i)$ is a negative scalar multiple of $\mathbf{v}^-(t_i)$. The specification of whether the signed angle at a cusp equals π or $-\pi$ is easiest to describe when γ is simple, closed, and positively oriented. Under these added hypotheses, $\alpha_i = \pi$ if $\mathbf{v}^-(t_i)$ points toward the exterior, or $\alpha_i = -\pi$ if $\mathbf{v}^-(t_i)$ points toward the interior; see Fig. 2.8. The sign convention is the opposite if γ is negatively oriented.

If γ is closed and $\gamma'(a) \neq \gamma'(b)$, then $\gamma(a)$ counts as a corner, and the corresponding signed angle is defined exactly as above, but with $\mathbf{v}^-(a)$ replaced by $\mathbf{v}^-(b)$.

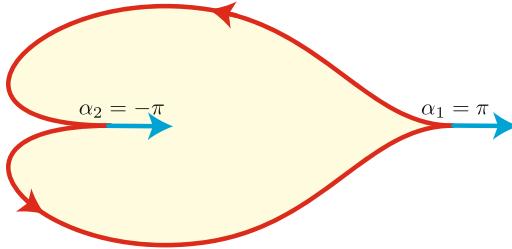


FIGURE 2.8. The sign convention at a cusp

THEOREM 2.7 (Generalized Hopf's Umlaufsatz).

Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a unit-speed positively oriented piecewise-regular simple closed plane curve. Let κ_s denote its signed curvature function, and let $\{\alpha_i\}$ be the list of signed angles at its corners. Then

$$\int_a^b \kappa_s(t) dt + \sum_i \alpha_i = 2\pi.$$

Here “ $\int_a^b \kappa_s(t) dt$ ” is shorthand for $\sum_i \left(\int_{t_i}^{t_{i+1}} \kappa_s(t) dt \right)$, which means the sum of the integral of κ_s over the smooth segments of γ .

If γ is regular (no corners) and θ denotes a global angle function of γ , recall from Sect. 6 of Chap. 1 that $\kappa_s = \theta'$, so

$$\int_a^b \kappa_s(t) dt = \int_a^b \theta'(t) dt = \theta(b) - \theta(a) = 2\pi \cdot (\text{rotation index}).$$

So in this case, Theorem 2.7 says that the rotation index equals 1, which we knew from Theorem 2.1.

When γ has corners, it is still possible to define an “angle function” θ that has a jump discontinuity at each corner by an amount equal to the corresponding signed angle, and that elsewhere satisfies $\kappa_s = \theta'$. The expression $\int_a^b \kappa_s(t) dt + \sum_i \alpha_i$ equals the net change in this (discontinuous) angle function. The proof of Theorem 2.7 involves smoothing the corners so that this angle function becomes continuous:

PROOF IDEA. The visual idea of the proof is to smooth γ in neighborhoods of the corners, as illustrated in Fig. 2.9. If $\tilde{\gamma}$ denotes a smoothed version of γ (in which neighborhoods of the corners have been replaced with the dashed lines shown in the figure), and $\tilde{\kappa}_s$ is the signed curvature function of $\tilde{\gamma}$, then our original version of Hopf's Umlaufsatz says that $\int_a^b \tilde{\kappa}_s(t) dt = 2\pi$. Although we will not discuss the analytic details, it is visually believable that the smoothing can be constructed such that

$$\int_a^b \kappa_s(t) dt + \sum_i \alpha_i = \int_a^b \tilde{\kappa}_s(t) dt = 2\pi.$$

□

The above proof helps explain our previous definition of the signed angle at a corner. The definition was essentially contrived to match the net change in the angle function after smoothing; in other words, α_i is the net counterclockwise rotation of the smoothed corner (interpreted as clockwise if α_i is negative), in the limit as the smoothing occurs in a smaller and smaller neighborhood of the corner. This description applies equally well at a cusp.

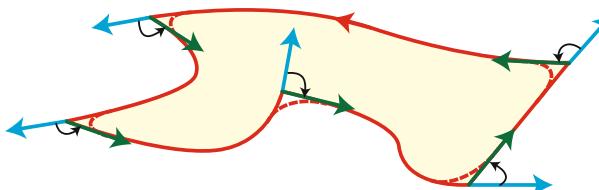


FIGURE 2.9. Hopf's Umlaufsatz is generalized by smoothing the curve at the corners

It is sometimes convenient to rephrase the previous theorem in terms of *interior angles*. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a piecewise-smooth simple closed plane curve with signed angles denoted by $\{\alpha_i\}$. The *i*th **interior angle** of γ , denoted by $\beta_i \in [0, 2\pi]$, is defined as in Fig. 2.10. Notice that changing the orientation of γ would change the sign of each signed angle, but would not affect the interior angles. Interior angles are related to signed angles as follows:

$$\beta_i = \begin{cases} \pi - \alpha_i & \text{if } \gamma \text{ is positively oriented,} \\ \pi + \alpha_i & \text{if } \gamma \text{ is negatively oriented.} \end{cases}$$

In Fig. 2.8, for example, $\beta_1 = 0$ and $\beta_2 = 2\pi$.

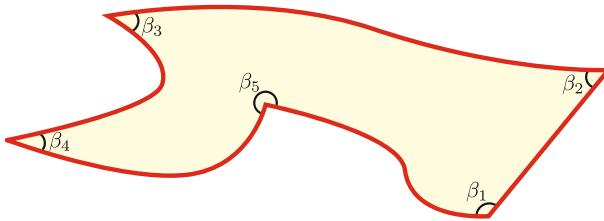


FIGURE 2.10. Interior angles

In Theorem 2.7, γ is assumed to be positively oriented, so the theorem becomes

$$\boxed{\int_a^b \kappa_s(t) dt = \sum_i \beta_i - (n - 2)\pi},$$

where n is the number of corners. If the smooth segments of γ are straight line segments, then this becomes

$$\boxed{\sum_i \beta_i = (n - 2)\pi},$$

which is a well-known formula for the sum of the interior angles of a polygon.

EXERCISES

EXERCISE 2.1. Prove that the two conditions in Definition 2.2 are equivalent, as claimed. *HINT: Using the existence of a tubular neighborhood, prove that $R_{90}(\gamma'(t))$ either points to the interior for all $t \in [a, b]$ or points to the exterior for all $t \in [a, b]$, so it suffices to consider a single value of t . Choose the value corresponding to the point labeled “p” in Fig. 2.2.*

EXERCISE 2.2. Prove Proposition 2.3. *HINT: Use a compactness argument to divide $[a, b]$ into finitely many subintervals, on each of which the image of \mathbf{f} is completely contained in one of the following four half-circles: top, bottom, right, left. Define a local angle function on each subinterval.*

Working from left to right, add the correct integer multiple of 2π to each local angle function so they match to form a global angle function.

EXERCISE 2.3. Let $\mathbf{f} : [a, b] \rightarrow S^1$ be a continuous function with $\mathbf{f}(a) = \mathbf{f}(b)$, and let $\mathbf{p} \in S^1$. If the degree of \mathbf{f} equals n , what is the minimal possible size of the set $\{t \in [a, b] \mid \mathbf{f}(t) = \mathbf{p}\}$? □

2. Convexity and the Four Vertex Theorem (Optional)

In this section, we describe one of the earliest global results in differential geometry, which provides a restriction on the number of vertices of a simple closed plane curve.

DEFINITION 2.8.

Let $\gamma : I \rightarrow \mathbb{R}^2$ be a regular plane curve. A point, $\gamma(t)$, on its trace is called a **vertex** if the signed curvature function has a local maximum or local minimum at t .

This definition is independent of parametrization (Exercise 2.5). As calculated in Exercise 2.4, an ellipse has exactly four vertices, two at which κ_s is maximal and two at which κ_s is minimal. The polar coordinate graph of $r = 1 - 2 \sin(\theta)$ has exactly two vertices; see Fig. 2.11. Two is the smallest number of vertices that a closed curve could have, because the signed curvature function must achieve its global maximum and global minimum on its compact domain (according to Corollary A.25 on page 356 of the appendix). Notice that every point of a circle qualifies as a vertex that is both a local maximum and a local minimum, so a circle has infinitely many vertices. For the same reason, so does a straight line.

The main theorem of this section says that a *simple* closed plane curve must have at least as many vertices as an ellipse has.

THEOREM 2.9 (The Four Vertex Theorem).

Every simple closed plane curve has at least four vertices.

This section is devoted to the proof of this theorem, but only in the special in which the curve is *convex*:

DEFINITION 2.10.

A simple closed plane curve is called **convex** if its trace lies entirely on one closed side of each of its tangent lines.

The term “closed side” means that the tangent line itself is considered part of either “side” into which it divides the plane, since of course the tangent line at \mathbf{p} intersects the trace at \mathbf{p} (and possibly also at nearby points if the trace is a straight line segment in a neighborhood of \mathbf{p}); see Fig. 2.12.

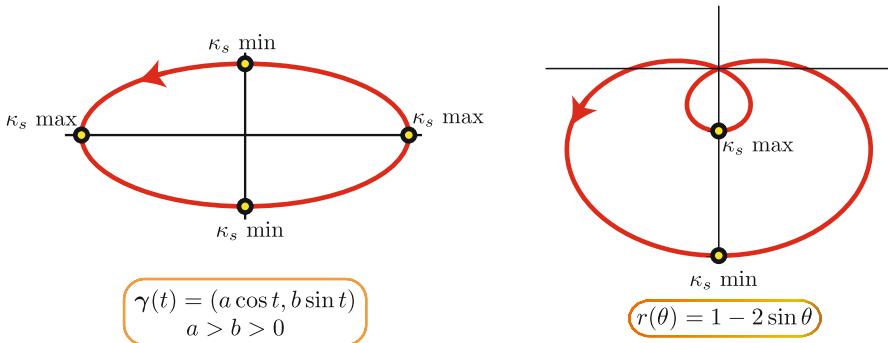


FIGURE 2.11. An ellipse has four vertices (*left*), while a nonsimple curve can have two (*right*)

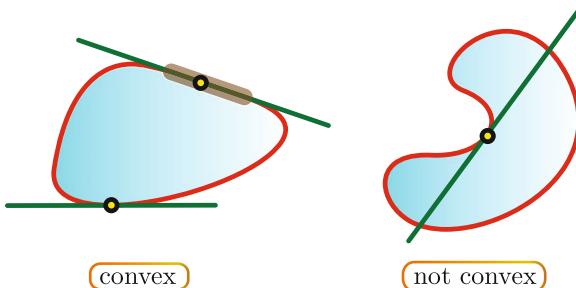


FIGURE 2.12. A convex curve lies on one closed side of each of its tangent lines

We'll require the following consequence of convexity:

LEMMA 2.11.

Let C be the trace of a simple closed convex plane curve, and let L be a line.

- (1) If L is tangent to C at two distinct points, then C contains the entire segment of L between these two points.
- (2) If $C \cap L$ contains more than two points, then it contains the entire segment of L between any pair of these points.

PROOF. Part (1) is left to the reader in Exercise 2.7 (with hints). For part (2), suppose that $L \cap C$ contains at least three points, and order them $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3\}$ so that \mathbf{p}_2 is between \mathbf{p}_1 and \mathbf{p}_3 along L . Notice that the tangent line to C at \mathbf{p}_2 must equal L , for otherwise, the other two points would lie on opposite sides of this tangent line. Since L is now a tangent line, convexity implies that C cannot cross L at either of the other two points, so L must be tangent to C at all three points. The result now follows from part (1). \square

The final ingredient that we'll require for the proof is a formulation of the familiar definition of signed curvature in terms of the separate x - and y -components of the curve. For a regular plane curve $\gamma(t) = (x(t), y(t))$, recall that

$$\mathbf{v}(t) = (x'(t), y'(t)), \quad \mathbf{a}(t) = (x''(t), y''(t)), \quad R_{90}(\mathbf{v}(t)) = (-y'(t), x'(t)).$$

When γ is of unit speed, its signed curvature is defined by $\mathbf{a}(t) = \kappa_s(t)R_{90}(\mathbf{v}(t))$ (Eq. 1.9 on page 33). The separate x - and y -components of this equation are

$$(2.1) \quad x''(t) = -\kappa_s(t)y'(t), \quad y''(t) = \kappa_s(t)x'(t).$$

PROOF OF THE FOUR VERTEX THEOREM FOR CONVEX CURVES. Let γ be a simple closed convex plane curve and let C denote its trace. Assume without loss of generality that γ is positively oriented. We can assume that κ_s is not constant on any interval, since every element of such an interval would correspond to a vertex, so C would have infinitely many vertices. In particular, we can assume that no segment of C is a straight line segment.

If γ had exactly three vertices, then two consecutive ones along C would be of the same type (both local maxima or both local minima). But this is impossible, because a non-locally-constant smooth real-valued function cannot have two consecutive local extrema of the same type.

Now suppose that γ has fewer than three vertices; in other words, its only vertices are the point $\mathbf{p} \in C$ at which κ_s attains its global maximum and the point $\mathbf{q} \in C$ at which κ_s attains its global minimum. Notice that $\mathbf{p} \neq \mathbf{q}$, for otherwise, κ_s would be constant. Let L denote the line through \mathbf{p} and \mathbf{q} . Lemma 2.11(2) implies that L intersects C only at \mathbf{p} and \mathbf{q} .

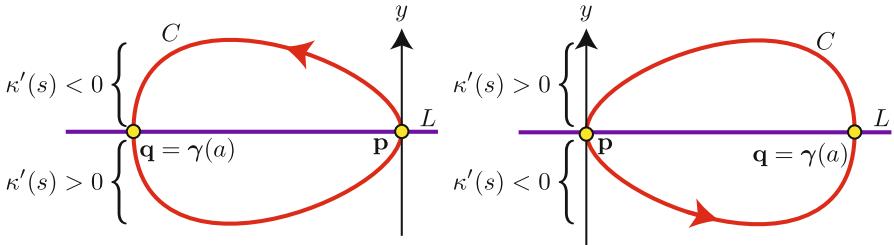
Choose an orientation-preserving unit-speed parametrization, $\gamma : [0, l] \rightarrow \mathbb{R}^2$, such that $\gamma(0) = \gamma(l) = \mathbf{p}$ and $\gamma(a) = \mathbf{q}$ for some $a \in (0, l)$. Notice that $\kappa'_s \leq 0$ on $(0, a)$ (since κ_s decreases from its maximum to its minimum) and $\kappa'_s \geq 0$ on (a, l) (since κ_s increases from its minimum to its maximum). We can assume without loss of generality (by applying a proper rigid motion) that \mathbf{p} is the origin and L is the x -axis.

Write γ in terms of its component functions: $\gamma(t) = (x(t), y(t))$. Notice that $y(t)$ changes sign only at $t = a$, since γ lies above L on $(0, a)$ and below L on (a, l) , or possibly vice versa, depending on whether \mathbf{p} lies to the right or left of \mathbf{q} . In either case, notice that $y(t)\kappa'_s(t)$ never changes sign; this expression is either ≤ 0 on all of $[0, l]$ or it is ≥ 0 on all of $[0, l]$; see Fig. 2.13.

Furthermore, the expression $y(t)\kappa'_s(t)$ equals zero only when $\kappa'_s = 0$, which does not occur on any interval of nonzero length. Thus, this expression has a nonzero average value:

$$\int_0^l y(t)\kappa'_s(t) dt \neq 0.$$

However, integrating by parts and using Eq. 2.1 together with the fact that the functions involved are periodic (they have the same values at 0 and l), we get

FIGURE 2.13. Either $y(t)\kappa'_s(t) \leq 0$ (left), or $y(t)\kappa'_s(t) \geq 0$ (right)

$$\int_0^l y(t)\kappa'_s(t) dt = y(t)\kappa_s(t) \Big|_{t=0}^{t=l} - \int_0^l \kappa_s(t)y'(t) dt = 0 + \int_0^l x''(t) dt = x'(t) \Big|_{t=0}^{t=l} = 0.$$

This contradiction shows that γ must have at least four vertices. \square

Since the concept of convexity is of independent importance, we end this section with some equivalent formulations of its definition:

PROPOSITION 2.12.

Let γ be a simple closed plane curve. Let C denote its trace. Let \mathcal{I} denote its interior. The following are equivalent:

- (1) γ is convex; that is, C lies on one closed side of each of its tangent lines.
- (2) The line segment joining any two points of \mathcal{I} lies entirely in \mathcal{I} ; see Fig. 2.14
- (3) κ_s does not change sign; that is, either $\kappa_s \geq 0$ on the whole domain or $\kappa_s \leq 0$ on the whole domain.

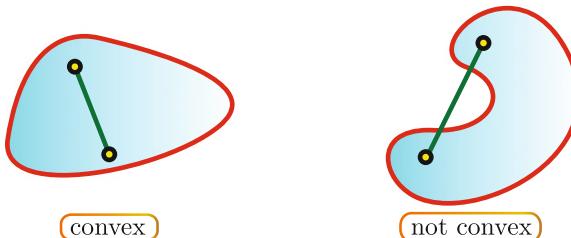


FIGURE 2.14. The line segment joining any two points of the interior of a convex curve must lie entirely in the interior

PROOF. $\boxed{(1) \implies (2)}$ (by contradiction): Suppose that γ is convex yet there is a pair $\mathbf{p}, \mathbf{q} \in \mathcal{I}$ such that the line segment joining them does *not* lie in \mathcal{I} . The (infinite) line, L , containing \mathbf{p} and \mathbf{q} must intersect C in at least three points (colored purple in Fig. 2.15). Lemma 2.11(2) implies that C contains the corresponding segment of L , so $\mathbf{p}, \mathbf{q} \in C$, contradicting the fact that they are interior points.

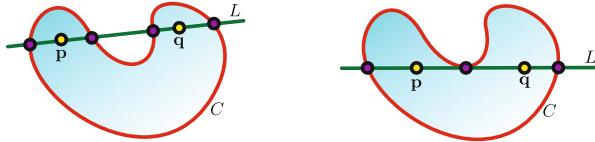


FIGURE 2.15. $L \cap C$ could contain four points (left) or three points (right), but no fewer

(2) \implies (1) Exercise 2.9.

(3) \implies (1) (by contradiction): Assume that κ_s does not change sign yet C lies on both sides of its tangent line at some point $\mathbf{p} = \gamma(t_0)$. The vector $\mathbf{n} = -R_{90}(\mathbf{v}(t_0))$ is orthogonal to C at \mathbf{p} . Consider the function

$$h(t) = \langle \gamma(t) - \gamma(t_0), \mathbf{n} \rangle.$$

Intuitively, $h(t)$ is the “height” of $\gamma(t)$ above the tangent line to C at $\gamma(t_0)$, with \mathbf{n} considered the “up” direction. Notice that $h(t_0) = 0$. Since C lies on both sides of the tangent line, h attains positive and negative values, so its global minimum and maximum occur at time values, called t_1 and t_2 respectively, that are distinct from each other and from t_0 . It is straightforward to show that the velocity vectors $\{\mathbf{v}(t_0), \mathbf{v}(t_1), \mathbf{v}(t_2)\}$ are mutually parallel; these velocity vectors are purple in Fig. 2.16.

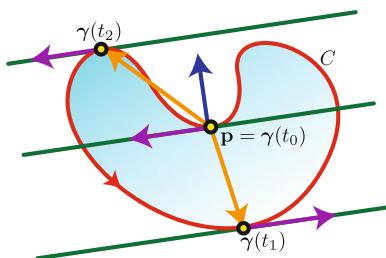


FIGURE 2.16. If C lies on both sides of its tangent line at \mathbf{p} , then the three purple velocity vectors are parallel

stant (and hence C is a straight line segment) on the segment of C between some pair of $\{\gamma(t_0), \gamma(t_1), \gamma(t_2)\}$. But this contradicts the fact that h has different values at all three points.

(1) \implies (3) (by contradiction): Assume that γ is convex yet $\kappa_s = \theta'$ changes sign. Then it is possible to choose nearby times $t_1 \neq t_2$, between which θ' changes sign, such that $\theta(t_1) = \theta(t_2)$, which means that $\mathbf{v}(t_1)$ and $\mathbf{v}(t_2)$ point in the same direction. By Hopf’s Umlaufsatz, there exists a time t_3 such that $\mathbf{v}(t_3)$ points in the opposite of this direction. The tangent lines to C at the three points $\{t_1, t_2, t_3\}$ are parallel. If these three tangent

Thus, some pair of these three velocity vectors must point in the same direction (rather than opposite directions). It doesn’t matter which pair, so let’s say that $\mathbf{v}(t_0)$ and $\mathbf{v}(t_2)$ point in the same direction, as in the figure. This means that a global angle function, θ , changes by an integer multiple of 2π between times t_0 and t_2 . By Hopf’s Umlaufsatz, θ changes by exactly $\pm 2\pi$ on the entire domain. Since $\kappa_s = \theta'$ does not change sign, θ does all of this changing monotonically. This is possible only if θ is constant.

lines were all distinct, then the middle one would contradict convexity, so some two of them must coincide (here “middle” means with respect to their positions as subsets of \mathbb{R}^2). Lemma 2.11 implies that the trace of γ is a straight line segment between these times. But γ can’t be straight between t_1 and t_2 , because θ' changes sign between them. Nor can γ be straight between t_3 and either other time, since θ changes by π between them. This is a contradiction. \square

EXERCISES

EXERCISE 2.4. Suppose $p > q > 0$ and consider the ellipse $\gamma(t) = (p \cos(t), q \sin(t))$. The *foci* of this ellipse are the two points on the x -axis with x -coordinates $\pm\sqrt{p^2 - q^2}$, colored purple in Fig. 2.17.

- (1) Prove that the sum of the distances from $\gamma(t)$ to these two foci is independent of t .
- (2) Prove that the signed curvature function of the ellipse is

$$\kappa_s(t) = \frac{pq}{(p^2 \sin^2(t) + q^2 \cos^2(t))^{\frac{3}{2}}}.$$

- (3) Prove that the critical points of the signed curvature function occur at $t \in \{0, \frac{\pi}{2}, \pi, \frac{3\pi}{2}\}$. The corresponding points on the ellipse are its intersections with the x - and y -axes.

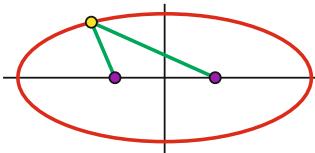


FIGURE 2.17. An ellipse is the set of points with constant summed distance to its two foci

EXERCISE 2.5. Prove that the definition of *vertex* is independent of parametrization.

EXERCISE 2.6. Let $f : (a, b) \rightarrow \mathbb{R}$ be a smooth function, and let $\gamma(t) = (t, f(t))$ be the natural parametrization of its graph. *Prove or disprove:*

- (1) A critical point of f is a vertex of γ .
- (2) A vertex of γ is a critical point of f .

EXERCISE 2.7. Prove Lemma 2.11(1). *HINT: Let $\mathbf{p}_1, \mathbf{p}_2$ denote the two points at which L is tangent to C . Let \mathbf{s} denote the last point of L past \mathbf{p}_1 that is contained in C (which could be \mathbf{p}_1 itself). If \mathbf{s} comes before \mathbf{p}_2 , show that just past \mathbf{s} , there would be a point, \mathbf{q} , such that \mathbf{p}_1 and \mathbf{p}_2 lie on different sides of its tangent line, contradicting convexity; see Fig. 2.18.*

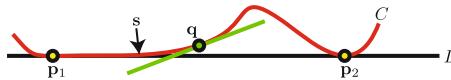


FIGURE 2.18. Moving away from \mathbf{p}_1 toward \mathbf{p}_2 along L , as soon as C were to separate from L , there would be a point $\mathbf{q} \in C$ with \mathbf{p}_1 and \mathbf{p}_2 on different sides of its tangent line. Thus, C cannot separate from L

EXERCISE 2.8. In the proof of the four vertex theorem, we chose an orientation-preserving unit-speed reparametrization and a proper rigid motion. Why did the reparametrization need to be orientation-preserving? Why did the rigid motion need to be proper?

EXERCISE 2.9. Prove $(2) \implies (1)$ in Proposition 2.12.

EXERCISE 2.10. Let γ be a closed plane curve whose signed curvature does not change sign (either $\kappa_s \geq 0$ on its whole domain or $\kappa_s \leq 0$ on its whole domain). If the rotation index of γ equals ± 1 , prove that γ is simple.

EXERCISE 2.11 (Convexity for a Piecewise-Regular Curve). Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a positively oriented piecewise-regular simple closed plane curve, with interior denoted by \mathcal{I} . Prove that the following are equivalent characterizations of what it means for γ to be **convex**:

- (1) The line segment joining any two points of \mathcal{I} lies entirely in \mathcal{I} .
- (2) $\kappa_s \geq 0$ and all signed angles are positive.

EXERCISE 2.12. Describe the history of the four vertex theorem and its converse. Discuss the ideas behind the proof in the nonconvex case. An excellent reference is [4].

□

3. Fenchel's Theorem (*Optional*)

It is natural to consider the total amount that a curve curves, measured as follows:

DEFINITION 2.13.

The **total curvature** of a regular curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ is defined as

$$\text{total curvature} = \int_a^b \kappa(t) |\gamma'(t)| dt.$$

The total curvature of γ is unchanged by reparametrization, so we will usually assume that γ is of unit speed, in which case its total curvature is $\int_a^b \kappa(t) dt$.

In order for a curve to be closed, it must return to where it started. How much total curvature does this require? The answer for a simple closed plane curves is a quick consequence of some previous theorems:

LEMMA 2.14.

The total curvature of a simple closed plane curve is $\geq 2\pi$, with equality if and only if it is convex.

PROOF. Let γ be a unit-speed simple closed plane curve. The integral of the *signed* curvature comes from Hopf's Umlaufsatz:

$$\int_a^b \kappa_s(t) dt = \int_a^b \theta'(t) dt = \theta(b) - \theta(a) = 2\pi \cdot (\text{rotation index}) = \pm 2\pi.$$

This is related to the integral of the *unsigned* curvature as follows:

$$\int_a^b \kappa(t) dt = \int_a^b |\kappa_s(t)| dt \geq \left| \int_a^b \kappa_s(t) dt \right| = 2\pi,$$

with equality if and only if κ_s does not change sign, which by Proposition 2.12 occurs if and only if γ is convex. \square

The goal of this section is to prove the following generalization to curves in \mathbb{R}^n :

THEOREM 2.15 (Fenchel's Theorem).

The total curvature of a closed curve in \mathbb{R}^n is $\geq 2\pi$, with equality if and only if it is a simple closed convex curve contained in a plane in \mathbb{R}^n (which means a translate of a two-dimensional subspace of \mathbb{R}^n).

The term “convex” was previously defined only for curves in \mathbb{R}^2 , but it also makes perfect sense for a curve contained in an arbitrary plane. For simplicity, we will prove Fenchel's theorem in the case $n = 3$. The general case follows from essentially the same argument.

We will require some vocabulary and facts related to the geometry of the sphere:

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}.$$

A “curve in S^2 ” means a space curve whose trace is a subset of S^2 . A “great circle” means the intersection of S^2 with a two-dimensional subspace of \mathbb{R}^3 . For example, the equator $E = \{(x, y, 0) \mid x^2 + y^2 = 1\}$ is a great circle. If $\mathbf{p}, \mathbf{q} \in S^2$, then $\overline{\mathbf{pq}}$ will denote their intrinsic distance in S^2 , which means the smallest possible arc length of a regular curve in S^2 between \mathbf{p} and \mathbf{q} . We will require the following fact:

LEMMA 2.16.

If $\mathbf{p}, \mathbf{q} \in S^2$ is a pair of distinct points, then $\overline{\mathbf{pq}} \leq \pi$, with equality if and only if $\mathbf{p} = -\mathbf{q}$. There exists a segment of a great circle from \mathbf{p} to \mathbf{q} with arc length $\overline{\mathbf{pq}}$, and this segment is unique if $\mathbf{p} \neq -\mathbf{q}$. The trace of every curve in S^2 from \mathbf{p} to \mathbf{q} with arc length $\overline{\mathbf{pq}}$ is a segment of a great circle.

This lemma will be easily proven when we discuss shortest paths in general curved surfaces in Chap. 5. For now, we'll assume that it is likely familiar to most readers. It essentially just says that segments of great circles are the unique shortest paths in S^2 , as all transatlantic pilots know.

Most of the work of proving Fenchel's theorem lies in proving the following fact:

LEMMA 2.17.

Let β be a regular closed curve in S^2 . If the trace of β intersects every great circle of S^2 , then the arc length of β is $\geq 2\pi$, with equality if and only if β is a simple parametrization of a great circle.

PROOF OF LEMMA 2.17. Assume without loss of generality that $\beta : [0, l] \rightarrow S^2$ is parametrized by arc length. Define $\mathbf{p} = \beta(0) = \beta(l)$ and $\mathbf{q} = \beta(l/2)$. Let β_1, β_2 denote the restrictions of β to the domains $[0, l/2]$ and $[l/2, l]$ respectively.

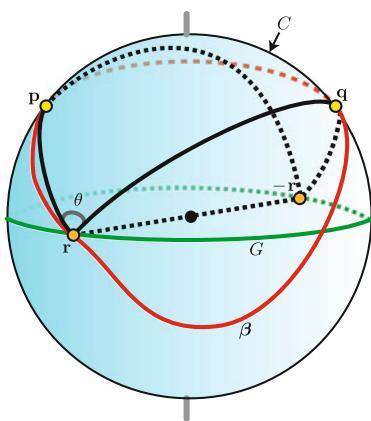
If $\mathbf{p} = -\mathbf{q}$, then $\overline{\mathbf{pq}} = \pi$, so β_1 and β_2 each have arc length $\geq \pi$, with simultaneous equality if and only the trace of each is half of a great circle. Thus the arc length of β is $\geq 2\pi$, with equality if and only if the trace of β equals two halves of great circles, which must be halves of the same great circle because β is smooth.

If $\mathbf{p} \neq -\mathbf{q}$, then there is a unique great circle C containing \mathbf{p} and \mathbf{q} . Let G denote the great circle that is orthogonal to C and equidistant to \mathbf{p} and \mathbf{q} ; see Fig. 2.19. The trace of β intersects G at some point \mathbf{r} (because it intersects every great circle). Notice that $\overline{\mathbf{rp}} = \overline{-\mathbf{rq}}$, because the 180-degree rotation about the illustrated axis is a rigid motion mapping $\mathbf{p} \mapsto \mathbf{q}$ and $\mathbf{r} \mapsto -\mathbf{r}$. It follows that

$$\overline{\mathbf{rp}} + \overline{\mathbf{rq}} = \overline{-\mathbf{rq}} + \overline{\mathbf{rq}} = \pi.$$

Either β_1 or β_2 travels between \mathbf{p} and \mathbf{q} via \mathbf{r} , so its arc length must be at least $\overline{\mathbf{rp}} + \overline{\mathbf{rq}} = \pi$. In fact, its arc length must be $> \pi$ (because equality would force the angle labeled θ to equal 180° in order for β to be smooth, contradicting the assumption that $\mathbf{p} \neq -\mathbf{q}$). Since this is half of β , the full arc length of β must be $> 2\pi$, as desired. \square

PROOF OF FENCHEL'S THEOREM FOR SPACE CURVES. Let $\gamma : [0, l] \rightarrow \mathbb{R}^3$ be a unit-speed closed curve. Since γ is of unit speed, its velocity function, \mathbf{v} , is a path in S^2 (visualized as in Fig. 1.17 from Sect. refch1:sec5 of Chap. 1).



We will use the fact that γ is closed to prove that the trace of \mathbf{v} intersects every great circle. For this, let $\mathcal{P} \subset \mathbb{R}^3$ be an arbitrary two-dimensional subspace, so that $G = \mathcal{P} \cap S^2$ is an arbitrary great circle. Let \mathbf{n} be a normal vector to \mathcal{P} . Notice that a point of S^2 lies on G if and only if it is orthogonal to \mathbf{n} . Since $\frac{d}{dt} \langle \gamma(t), \mathbf{n} \rangle = \langle \mathbf{v}(t), \mathbf{n} \rangle$, the fundamental theorem of calculus gives

FIGURE 2.19. The proof of Lemma 2.17

$$\int_0^l \langle \mathbf{v}(t), \mathbf{n} \rangle = \langle \gamma(l), \mathbf{n} \rangle - \langle \gamma(0), \mathbf{n} \rangle = 0 \quad (\text{because } \gamma \text{ is periodic}).$$

Since the average value of $\langle \mathbf{v}(t), \mathbf{n} \rangle$ equals zero, we must have $\langle \mathbf{v}(t_0), \mathbf{n} \rangle = 0$ for some $t_0 \in [0, l]$. Thus, the trace of \mathbf{v} intersects every great circle. By Lemma 2.17, the arc length of \mathbf{v} is $\geq 2\pi$.

Since γ is of unit speed, $\kappa(t) = |\mathbf{v}'(t)|$, so we have

$$(\text{total curvature of } \gamma) = \int_0^l \kappa(t) dt = \int_0^l |\mathbf{v}'(t)| dt = (\text{arc length of } \mathbf{v}) \geq 2\pi.$$

If equality holds, then \mathbf{v} must be a simple parametrization of a great circle, say the great circle $G = S^2 \cap \mathcal{P}$. Since $\gamma(t) = \int_0^t \mathbf{v}(u) du + \gamma(0)$, the trace of γ must lie in the plane $\{\gamma(0) + z \mid z \in \mathcal{P}\}$. After applying a rigid motion, we can assume that this plane is the xy -plane, so we can consider γ to be a plane curve. Its velocity function \mathbf{v} is a *simple* parametrization of the unit circle S^1 . This implies that γ has rotation index ± 1 and has a monotonic global angle function (or equivalently, its signed curvature does not change sign). Exercise 2.10 (on page 78) implies that γ is simple, and then Lemma 2.14 implies that γ is convex. \square

EXERCISES

EXERCISE 2.13. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a (not necessarily closed) regular plane curve with $\gamma(a) = \gamma(b)$. What is the minimal possible total curvature of γ ?

EXERCISE 2.14. Prove the $n = 2$ (plane curve) case of Fenchel's theorem.

Hint: The $n = 3$ proof involved showing that \mathbf{v} intersects every great circle of S^2 (that is, it meets the intersection of S^2 with every two-dimensional subspace of \mathbb{R}^3). The $n = 2$ analogue is that \mathbf{v} intersects every pair of antipodal points of S^1 (that is, it meets the intersection of S^1 with every one-dimensional subspace of \mathbb{R}^2).

EXERCISE 2.15. Prove that every great circle of S^2 is the image of the equator under a rigid motion.

EXERCISE 2.16. Prove that the length, L , of a regular closed curve in \mathbb{R}^n with nowhere vanishing curvature satisfies

$$L \geq \frac{2\pi}{\kappa_{\max}},$$

where $\kappa_{\max} > 0$ is the global maximum of its curvature function.

4. Green's Theorem (*Calculus Background*)

This section is devoted to Green's theorem, a powerful global theorem about vector fields on \mathbb{R}^2 . As a consequence, we will prove in the next section that a circle is the least-perimeter way to enclose a given area in the plane.

For consistency with other sources, in the remainder of this chapter we will sometimes use the term “curve” to mean the trace of a regular parametrized curve, and the term “oriented curve” to mean such a trace together with a choice of one of the two possible directions in which the trace could be traversed. A more formal and precise way to formulate this notion was discussed in Exercise 1.35 (on page 24), but the informal version is sufficient for our purposes.

We begin with some basic facts about vector fields on \mathbb{R}^n , which is also good preparation for our later study of vector fields on curved surfaces.

DEFINITION 2.18.

A **vector field** on an open set $U \subset \mathbb{R}^n$ is a smooth function $\mathbf{F} : U \rightarrow \mathbb{R}^n$.

Thus, \mathbf{F} associates to each point $\mathbf{p} \in U$ a vector $\mathbf{F}(\mathbf{p})$, which should be visualized with its tail drawn at \mathbf{p} . For example, a vector field on $U \subset \mathbb{R}^2$ will have the form $\mathbf{F}(x, y) = (P(x, y), Q(x, y))$, where $P, Q : U \rightarrow \mathbb{R}$ are called the **component functions** of \mathbf{F} (we’ll write this as $\mathbf{F} = (P, Q)$); see Fig. 2.20. Smoothness of \mathbf{F} means that the component functions are smooth in the sense that all partial derivatives of all orders exist. Similarly, a vector field on $U \subset \mathbb{R}^3$ has the form $\mathbf{F}(x, y, z) = (P(x, y, z), Q(x, y, z), R(x, y, z))$, shorthanded as $\mathbf{F} = (P, Q, R)$.

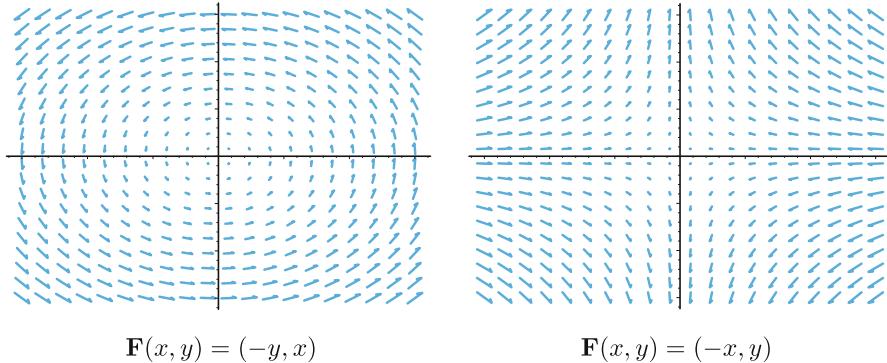


FIGURE 2.20. Two vector fields on \mathbb{R}^2 , with all vectors drawn about one-tenth their correct lengths for legibility

In physics applications, a vector field often represents a **force field**; that is, $\mathbf{F}(\mathbf{p})$ represents the force that would act on an object placed at position \mathbf{p} . The object might be a satellite acted on by gravitational forces of nearby planets, a paperclip acted on by the forces of nearby magnets, or a piece of tumbleweed acted on by wind currents. But the tumbleweed example is a stretch, because wind currents tend to change with time, while vector fields model forces that change only with position.

When a constant (vector) force, \mathbf{F} , moves an object along the displacement vector \mathbf{D} (which points from the starting to the ending position), the **work** done is defined as

$$W = |\mathbf{F}| |\mathbf{D}| \cos(\theta) = \langle \mathbf{F}, \mathbf{D} \rangle.$$

To understand why this definition is reasonable, first imagine lifting a five-pound statue three feet off the ground. This requires 15 foot-pounds of work against gravity (\mathbf{F} and \mathbf{D} point in the same direction here—straight up—so their inner product is their regular product). If you instead move the five-pound statue along a diagonal line so that it ends up three feet up and seven feet to the right, as in Fig. 2.21, then $\mathbf{F} = (0, 5)$ and $\mathbf{D} = (7, 3)$, so $W = \langle (0, 5), (7, 3) \rangle = 15$. It's not surprising that the answer stayed the same—only the component of the displacement in the direction of the force is relevant (the horizontal component of the displacement requires no work against gravity). This is the same as saying that only the component of the force in the direction of the displacement is relevant. In any case, this example should help explain why our definition of work is reasonable.

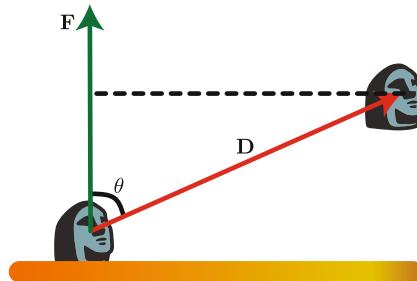


FIGURE 2.21. $W = |\mathbf{F}| |\mathbf{D}| \cos(\theta) = \langle \mathbf{F}, \mathbf{D} \rangle$

How much work is required to send the statue to the Moon? It would need to travel a curved path, γ , along which the gravitational force vector, \mathbf{F} , would change as it moves farther from Earth's tug and closer to the Moon's tug. In situations like this, the work is calculated with a line integral.

DEFINITION 2.19.

If C is an oriented plane curve parametrized as $\gamma : [a, b] \rightarrow \mathbb{R}^2$ and \mathbf{F} is a vector field whose domain contains C , then the **line integral** of \mathbf{F} along C is defined as

$$\int_C \mathbf{F} \cdot d\gamma = \int_a^b \langle \mathbf{F}(\gamma(t)), \gamma'(t) \rangle dt.$$

When C is a simple closed curve, the line integral is also denoted by $\oint_C \mathbf{F} \cdot d\gamma$, and is called the **circulation** of \mathbf{F} around C .

EXAMPLE 2.20. Consider the vector field on \mathbb{R}^2 defined as $\mathbf{F}(x, y) = (-y, x)$, previously illustrated in Fig. 2.20. Let C denote the counterclockwise circle of radius 3 about the origin of \mathbb{R}^2 . To compute $\oint_C \mathbf{F} \cdot d\gamma$, we first parametrize C as $\gamma(t) = (\underbrace{3 \cos(t)}_{x(t)}, \underbrace{3 \sin(t)}_{y(t)})$, $t \in [0, 2\pi]$, and write

$$\begin{aligned}\oint_C \mathbf{F} \cdot d\gamma &= \int_0^{2\pi} \langle \mathbf{F}(x(t), y(t)), (x'(t), y'(t)) \rangle \\ &= \int_0^{2\pi} \langle (-3 \sin t, 3 \cos t), (-3 \sin t, 3 \cos t) \rangle = \int_0^{2\pi} 9 = 18\pi.\end{aligned}$$

The line integral represents the work done by the force field \mathbf{F} in moving the object along the curve C . This interpretation is reasonable, because a Riemann sum for the line integral has the form

$$\sum_i \langle \mathbf{F}(\gamma(t_i)), \gamma'(t_i) \rangle \Delta t_i = \sum_i \left\langle \mathbf{F}(\gamma(t_i)), \underbrace{\gamma'(t_i) \Delta t_i}_{\approx \text{displacement}} \right\rangle,$$

where t_i is a sample point from the i th subinterval into which $[a, b]$ is partitioned, and Δt_i is the length of this subinterval. The restriction of γ to each such subinterval is a subarc of C . When Δt_i is small, the forces along the i th subarc are approximately constant at the sample value $\mathbf{F}(\gamma(t_i))$, and this subarc is itself an approximately straight displacement by $\gamma'(t_i) \Delta t_i$ (because this vector points in the right direction and has the right length). Thus, the i th term of this Riemann sum approximates the work done in moving the object along the i th subarc, so the entire Riemann sum approximates the work done in moving the object along all of C ; see Fig. 2.22.

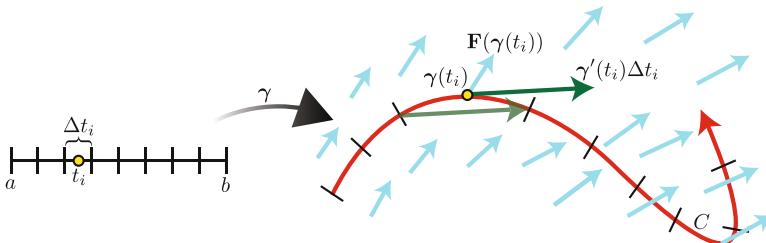


FIGURE 2.22. A Riemann sum for $\int_C \mathbf{F} \cdot d\gamma$ approximates the work done by the force field \mathbf{F} in moving the object along C

If C is only a piecewise-regular curve, then $\int_C \mathbf{F} \cdot d\gamma$ is defined as the sum of the line integrals of \mathbf{F} along the smooth segments of C . Line integrals are unchanged by orientation-preserving reparametrizations (Exercise 2.17). An orientation-reversing reparametrization would change the sign of the line integral; that is, $\int_{-C} \mathbf{F} \cdot d\gamma = -\int_C \mathbf{F} \cdot d\gamma$, where “ $-C$ ” represents C with opposite orientation (traversed in the opposite direction).

If the vectors of \mathbf{F} (encountered along the way as C is traversed) mostly point in the direction of motion, then the line integral will be positive, and is interpreted as the work done by \mathbf{F} in moving the object along the curve. This was the case in Example 2.20. If \mathbf{F} mostly points against the direction of motion, then the line integral will be negative, and its absolute value is interpreted as the work required for some independent force to move the object against \mathbf{F} along the curve.

If C is a small counterclockwise circle, you could think of \mathbf{F} as modeling the flow of a water current, and imagine C as the rim of a paddle wheel set in the current. The circulation is roughly the force with which the current spins the paddle wheel counterclockwise; see Fig. 2.23.

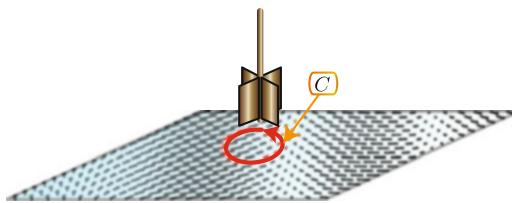


FIGURE 2.23. The circulation $\oint_C \mathbf{F} \cdot d\gamma$ is roughly the force with which the current spins the paddle wheel counterclockwise

Notice that $\int_C \mathbf{F} \cdot d\gamma$ is defined for arbitrary paths and arbitrary vector fields, which need not be related to each other. But the story becomes more natural when there are no forces other than \mathbf{F} . In this case, the path γ is completely determined by \mathbf{F} and by the initial conditions $\{\gamma(a), \gamma'(a)\}$; see Fig. 2.24. The next example provides a physical interpretation of the line integral in this situation.

EXAMPLE 2.21 (Line Integrals with No Other Forces). *If \mathbf{F} is the only force, then Newton's law says that $\mathbf{F}(\gamma(t)) = m\mathbf{a}(t)$ for all $t \in [a, b]$, where m is the object's mass. One could use this to solve for γ , given initial conditions and a formula for \mathbf{F} , but instead we will use it here to derive a general meaning for the line integral:*

$$\begin{aligned}\int_C \mathbf{F} \cdot d\gamma &= \int_a^b \langle \mathbf{F}(\gamma(t)), \gamma'(t) \rangle dt = \int_a^b \langle m\mathbf{a}(t), \mathbf{v}(t) \rangle dt = \frac{m}{2} \int_a^b \frac{d}{dt} \langle \mathbf{v}(t), \mathbf{v}(t) \rangle dt \\ &= \frac{m}{2} \int_a^b \frac{d}{dt} |\mathbf{v}(t)|^2 dt = \frac{m}{2} |\mathbf{v}(b)|^2 - \frac{m}{2} |\mathbf{v}(a)|^2.\end{aligned}$$

Since an object's **kinetic energy** is defined as $\frac{m}{2}|\mathbf{v}|^2$, we learn that $\int_C \mathbf{F} \cdot d\gamma$ equals the object's net change in kinetic energy between times a and b .

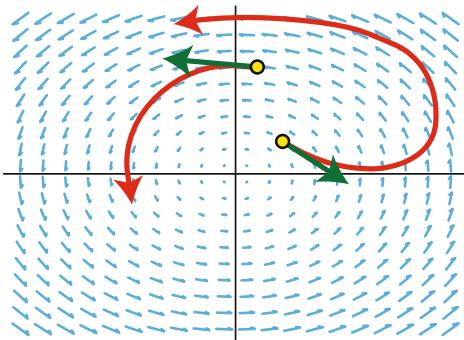


FIGURE 2.24. If there are no forces other than \mathbf{F} , then the object's path is determined by its initial position and initial velocity; it flows with the current

This physical interpretation agrees with our discussion earlier. If the force vectors are mostly in the direction of motion, then they push the object along, increasing its kinetic energy. If they are mostly against the direction of motion, then they slow the object down, decreasing its kinetic energy. Since work and energy are measured with the same units, this also agrees with the intuition that line integrals should represent work.

Another situation in which the line integral has a natural interpretation occurs when the vector field is *conservative*.

DEFINITION 2.22.

Let $U \subset \mathbb{R}^n$ be an open set, and let $f : U \rightarrow \mathbb{R}$ be a smooth function. The **gradient** of f , denoted by ∇f , is the vector field on U whose i th component function is the partial derivative of f with respect to the i th input variable. A vector field, \mathbf{F} , on U is called **conservative** if it is the gradient of some smooth function, f , on U . In this case, f is called a **potential function** of \mathbf{F} .

For example, the gradient of $f : \mathbb{R}^2 \rightarrow \mathbb{R}$ is $\nabla f = (f_x, f_y)$, where $f_x = \frac{\partial f}{\partial x}$ and $f_y = \frac{\partial f}{\partial y}$. Similarly, the gradient of $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ is $\nabla f = (f_x, f_y, f_z)$. The following should be familiar from multivariable calculus:

LEMMA 2.23.

If $U \subset \mathbb{R}^n$ is an open set, $f : U \rightarrow \mathbb{R}$ is a smooth function, and $\gamma : I \rightarrow U$ is a regular curve, then for all $t \in I$,

$$(2.2) \quad \frac{d}{dt} f(\gamma(t)) = \langle \nabla f(\gamma(t)), \gamma'(t) \rangle.$$

At a particular time $t_0 \in I$, if we let $\mathbf{p}_0 = \gamma(t_0)$ and $\mathbf{v}_0 = \gamma'(t_0)$, this derivative is denoted by $df_{\mathbf{p}_0}(\mathbf{v}_0)$:

$$(2.3) \quad df_{\mathbf{p}_0}(\mathbf{v}_0) = \left. \frac{d}{dt} \right|_{t=t_0} f(\gamma(t)) = \langle \nabla f(\mathbf{p}_0), \mathbf{v}_0 \rangle.$$

We call $df_{\mathbf{p}_0}(\mathbf{v}_0)$ the **directional derivative** of f at \mathbf{p}_0 in the direction of \mathbf{v}_0 (although some books reserve this term for the case that \mathbf{v}_0 is of unit length). It represents the initial rate at which f changes along *any* regular

curve passing through \mathbf{p}_0 with initial velocity vector \mathbf{v}_0 . Notice that the value $\frac{d}{dt}|_{t=t_0} f(\gamma(t))$ depends only on the vectors $\nabla f(\mathbf{p}_0)$ and $\mathbf{v}_0 = \gamma'(t_0)$. It does not depend on $\gamma''(t_0)$ or on any other higher-order derivative information.

Figure 2.25 shows a contour diagram (a collection of level curves) for the function $f(x, y) = -\frac{1}{2}x^2 + \frac{1}{2}y^2$ together with its gradient $\nabla f(x, y) = (-x, y)$. This figure illustrates the following general geometric relationship between gradients and contour diagrams:

LEMMA 2.24.

If $U \subset \mathbb{R}^n$ is open, $f : U \rightarrow \mathbb{R}$ is smooth, and $\mathbf{p} \in U$ is such that $\nabla f(\mathbf{p}) \neq \mathbf{0}$, then:

- (1) $\nabla f(\mathbf{p})$ is orthogonal to the level set $S_{\mathbf{p}} = \{\mathbf{q} \in U \mid f(\mathbf{q}) = f(\mathbf{p})\}$ in this sense: if γ is any regular curve with $\gamma(0) = \mathbf{p}$ whose trace lies in $S_{\mathbf{p}}$, then $\langle \gamma'(0), \nabla f(\mathbf{p}) \rangle = 0$.
- (2) $\nabla f(\mathbf{p})$ points in the direction of greatest increase of f . More precisely, the directional derivative $df_{\mathbf{p}}(\mathbf{u})$ is maximized among all unit vectors \mathbf{u} by the choice $\mathbf{u} = \frac{\nabla f(\mathbf{p})}{|\nabla f(\mathbf{p})|}$.
- (3) The norm of the gradient equals the rate of increase of f in this maximizing direction; that is, $|\nabla f(\mathbf{p})| = df_{\mathbf{p}}(\mathbf{u})$, where $\mathbf{u} = \frac{\nabla f(\mathbf{p})}{|\nabla f(\mathbf{p})|}$.

PROOF. For (1), since $f(\gamma(t))$ is constant, $0 = \frac{d}{dt}|_{t=0} f(\gamma(t)) = \langle \nabla f(\mathbf{p}), \gamma'(0) \rangle$. For (2) and (3), notice that $df_{\mathbf{p}}(\mathbf{u}) = \langle \nabla f(\mathbf{p}), \mathbf{u} \rangle = |\nabla f(\mathbf{p})| \cos(\theta)$ has maximal value $|\nabla f(\mathbf{p})|$ occurring when the angle θ between \mathbf{u} and $\nabla f(\mathbf{p})$ equals zero. \square

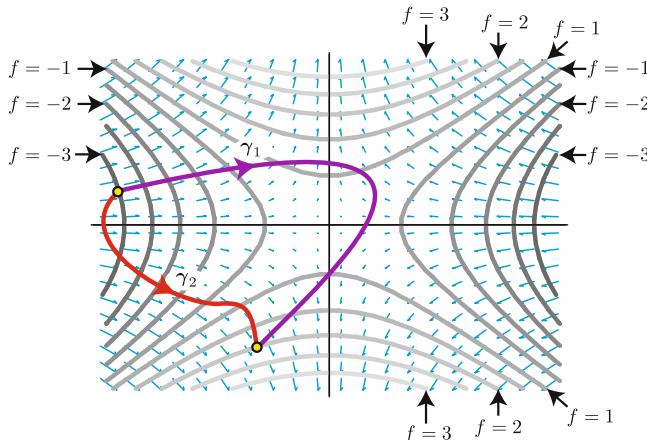


FIGURE 2.25. A contour diagram of $f(x, y) = -\frac{1}{2}x^2 + \frac{1}{2}y^2$ together with its gradient $\nabla f(x, y) = (-x, y)$. The line integral of ∇f along either γ_1 or γ_2 equals $2 - (-3) = 5$ (the net change in f)

The following proposition says that the line integral of a conservative vector field equals the net change of its potential function, as illustrated in Fig. 2.25:

PROPOSITION 2.25.

If $U \subset \mathbb{R}^n$ is open, $f : U \rightarrow \mathbb{R}$ is smooth, and $\gamma : [a, b] \rightarrow U$ is a parametrization of the piecewise-regular oriented curve C , then

$$\int_C \nabla f \cdot d\gamma = f(\gamma(b)) - f(\gamma(a)).$$

In particular, this line integral is **path-independent**—it would have the same value if C were replaced by any other piecewise-regular oriented curve in U with the same starting point $\gamma(a)$ and the same ending point $\gamma(b)$.

PROOF. We will prove this proposition in the case that C is smooth (the piecewise-regular case follows easily from this case). For this, we combine Eq. 2.2 with the fundamental theorem of calculus:

$$\int_C \nabla f \cdot d\gamma = \int_a^b \langle \nabla f(\gamma(t)), \gamma'(t) \rangle dt = \int_a^b \frac{d}{dt} f(\gamma(t)) dt = f(\gamma(b)) - f(\gamma(a)).$$

□

If γ is a closed path, then $\gamma(a) = \gamma(b)$, so $\oint_C \nabla f \cdot d\gamma = 0$. In fact, we have this:

PROPOSITION 2.26.

The following are equivalent properties for a vector field \mathbf{F} on an open path-connected set $U \subset \mathbb{R}^n$:

- (1) \mathbf{F} is conservative.
- (2) $\oint_C \mathbf{F} \cdot d\gamma = 0$ for every piecewise-regular closed curve C in U .
- (3) Line integrals of \mathbf{F} are path-independent; i.e., if C_1, C_2 are piecewise-regular curves in U with the same start and end points, then $\int_{C_1} \mathbf{F} \cdot d\gamma = \int_{C_2} \mathbf{F} \cdot d\gamma$.

PROOF. Exercise 2.18. □

We can now justify the term “conservative”—these are the vector fields for which there is a conservation of energy law.

EXAMPLE 2.27 (Conservation of Energy). Suppose that $\mathbf{F} = \nabla f$ is a conservative vector field. Physicists call $\rho = -f$ the “potential energy function.” If there are no forces other than \mathbf{F} , then the curve γ that an object will follow

is determined by the object's initial position $\gamma(0)$ and its initial velocity $\gamma'(0)$. In this case, Example 2.21 combines with Proposition 2.25 to yield

$$\int_C \mathbf{F} \cdot d\gamma = \frac{m}{2} |\mathbf{v}(b)|^2 - \frac{m}{2} |\mathbf{v}(a)|^2 = -\rho(b) + \rho(a).$$

Thus, the total energy (kinetic plus potential) is the same at times a and b :

$$\frac{m}{2} |\mathbf{v}(a)|^2 + \rho(a) = \frac{m}{2} |\mathbf{v}(b)|^2 + \rho(b).$$

In summary, for a conservative vector field, it is possible to define a position-dependent "potential energy" that obeys a conservation law: potential energy gets traded off against kinetic energy as an object moves under the influence of only the field.

EXAMPLE 2.28 (Gravity). According to Newton's law, the magnitude of the gravitational force between two objects (with masses denoted by m and M) equals $\frac{C}{r^2}$, where r is the distance between their centers of mass, $C = mMG$, and G is the universal gravitational constant.

Consider a large object (such as the Earth) with mass M centered at $(0, 0, 0)$. Let $\mathbf{F}(\mathbf{p})$ denote the force it exerts on a small object (such as a grapefruit) with mass m centered at $\mathbf{p} = (x, y, z)$. Since $\mathbf{F}(\mathbf{p})$ has magnitude $\frac{C}{|\mathbf{p}|^2}$ and points in the direction of the center-pointing unit vector $-\frac{\mathbf{p}}{|\mathbf{p}|}$, we have

$$\mathbf{F}(\mathbf{p}) = -\frac{C}{|\mathbf{p}|^3} \mathbf{p} = -C (x^2 + y^2 + z^2)^{-3/2} (x, y, z).$$

To demonstrate that \mathbf{F} is a conservative vector field on its domain, $\mathbb{R}^3 - \{(0, 0, 0)\}$, we must construct a smooth real-valued potential function f on this domain such that $\mathbf{F} = \nabla f$. Trial and error suffices to come up with

$$f(\mathbf{p}) = \frac{C}{|\mathbf{p}|} - K = C (x^2 + y^2 + z^2)^{-1/2} - K,$$

where $K \in \mathbb{R}$ is an arbitrary constant. Therefore, $\rho(\mathbf{p}) = -f(\mathbf{p}) = K - \frac{C}{|\mathbf{p}|}$ is the potential energy function. It is often convenient to choose K such that the potential energy equals zero on the surface of the Earth. The conservation law for this situation says that the total energy

$$\underbrace{\frac{1}{2} m |\mathbf{v}|^2}_{\text{kinetic}} + \underbrace{K - \frac{C}{|\mathbf{p}|}}_{\text{potential}}$$

remains constant for an object under the influence of only gravity (no air resistance, no smashing into the Earth's surface, etc.).

Now look back at the graph of the vector field $\mathbf{F}(x, y) = (-y, x)$ in Fig. 2.20. How could you verify that this vector field is *not* conservative? Visually, you could observe that the line integral is not zero around a circle centered at the origin (as confirmed in Example 2.20). Or algebraically, you could use the following:

LEMMA 2.29.

If $\mathbf{F} = (P, Q)$ is a conservative vector field on an open set $U \subset \mathbb{R}^2$, then $Q_x = P_y$ at every point of U .

PROOF. Since $\mathbf{F} = \nabla f$, we have $P = f_x$ and $Q = f_y$. Since mixed partial derivatives commute,

$$Q_x = (f_y)_x = (f_x)_y = P_y.$$

□

Thus, two things are true for a conservative vector field on \mathbb{R}^2 . First, the quantity $Q_x - P_y$ vanishes, and second, the circulation around closed curves vanish. The following definition hints at the relationship between these two things:

DEFINITION 2.30.

The **infinitesimal circulation** of the vector field $\mathbf{F} = (P, Q)$ is the real-valued function defined as $Q_x - P_y$ on the domain of \mathbf{F} .

This (nonstandard) term is appropriate because the infinitesimal circulation equals the limit circulation around smaller and smaller circles:

COROLLARY 2.31.

If \mathbf{F} is a vector field defined in a neighborhood of $\mathbf{p} \in \mathbb{R}^2$, and C_r denotes the counterclockwise circle of radius r centered at \mathbf{p} , then

$$(Q_x - P_y)(\mathbf{p}) = \lim_{r \rightarrow 0} \frac{1}{\pi r^2} \oint_{C_r} \mathbf{F} \cdot d\gamma.$$

This result is labeled a corollary because Green's theorem will be required to prove it. Nevertheless, the geometric meaning of Green's theorem will be easier to comprehend after understanding the content of this corollary. Returning to the paddle-wheel metaphor, Corollary 2.31 roughly says that the infinitesimal circulation at \mathbf{p} measures the counterclockwise force (per unit of paddle-wheel area) of the current on a small paddle wheel placed at \mathbf{p} . Why might you expect the expression $Q_x - P_y$ to measure such a thing? This expression is positive when Q_x is positive and P_y is negative. Both cause counterclockwise spin. $Q_x > 0$ means that the y -component of \mathbf{F} increases as x increases, so the current has more upward push against the right side of the paddle wheel than the left, causing counterclockwise spin. And $P_y < 0$ means that the x -component of \mathbf{F} decreases as y increases, so the current has more rightward push against the bottom side of the paddle wheel than the top, again causing counterclockwise spin. These two phenomena (the spin caused by vertical variations in the horizontal component of force, and the spin caused by horizontal variations in the vertical component of force) are separated out in the top two vector fields displayed in Fig. 2.26, while these phenomena combine additively in the bottom vector field. Is it visually

believable that a paddle wheel placed at any position in any of these three vector fields will experience the same force spinning it counterclockwise?

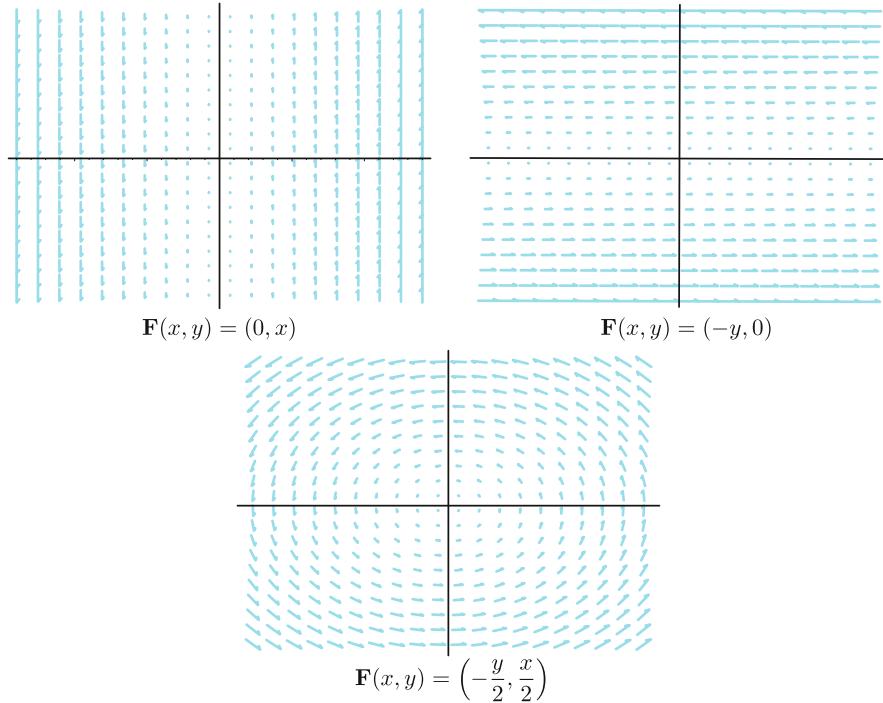


FIGURE 2.26. Three vector fields with constant infinitesimal circulation equal to 1

THEOREM 2.32 (Green's Theorem).

Let C be a positively oriented piecewise-regular simple closed curve in the plane. Let D denote the interior of C . Let $\mathbf{F} = (P, Q)$ be a vector field defined on an open set containing $D \cup C$. Then,

$$\oint_C \mathbf{F} \cdot d\gamma = \iint_D (Q_x - P_y) dA.$$

Green's theorem says that the circulation of \mathbf{F} around C equals the integral over D of the infinitesimal circulation. So if the average value over D of the infinitesimal circulation is positive (paddle wheels mostly spin counterclockwise), then the circulation around C is positive (the vectors encountered while traversing the curve mostly align with the direction of motion).

EXAMPLE 2.33. *Green's theorem provides an alternative way to compute the line integral of Example 2.20. Since the infinitesimal circulation of the vector field in this example is constant at 2, we have*

$$\oint_C \mathbf{F} \cdot d\gamma = \iint_D 2 \, dA = 2 \cdot \text{Area}(D) = 18\pi,$$

where D is the interior of C .

PROOF OF COROLLARY 2.31 USING GREEN'S THEOREM. Let D_r denote the interior of C_r . If r is sufficiently small, then $(Q_x - P_y)$ is approximately constant over D_r at the sample value $(Q_x - P_y)(\mathbf{p})$. Green's theorem gives

$$\oint_{C_r} \mathbf{F} \cdot d\gamma = \iint_{D_r} (Q_x - P_y) \, dA \approx \text{area}(D_r)((Q_x - P_y)(\mathbf{p})),$$

from which the result follows. \square

We will next prove Green's theorem in the special case that C is a rectangle, and then give a nonrigorous indication of how the general case follows.

PROOF OF GREEN'S THEOREM WHEN C IS A RECTANGLE. Suppose that D is the region $\{(x, y) \in \mathbb{R}^2 \mid a \leq x \leq b, c \leq y \leq d\}$, and C is its boundary. Denote the four segments of C as in Fig. 2.27, so $C = C_B + C_R - C_T - C_L$.

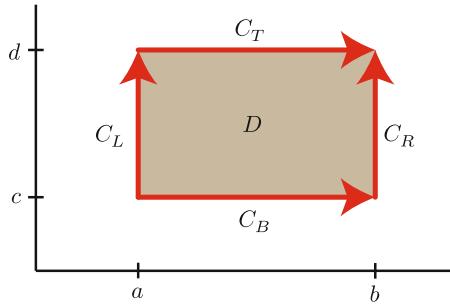


FIGURE 2.27. $C = C_B + C_R - C_T - C_L$

We parametrize each segment of C in the most natural manner. For example, we parametrize C_B as $\gamma(t) = (t, c)$, $t \in [a, b]$, so that

$$\int_{C_B} \mathbf{F} \cdot d\gamma = \int_a^b \langle \mathbf{F}(\gamma(t)), \gamma'(t) \rangle \, dt = \int_a^b \langle (P(t, c), Q(t, c)), (1, 0) \rangle \, dt = \int_a^b P(t, c) \, dt.$$

After similarly expressing the line integrals along the other three segments, we get

$$(2.4) \quad \int_C \mathbf{F} \cdot d\gamma = \underbrace{\int_c^d (Q(b, t) - Q(a, t)) \, dt}_{\int_{C_R} \mathbf{F} \cdot d\gamma - \int_{C_L} \mathbf{F} \cdot d\gamma} + \underbrace{\int_a^b (P(t, c) - P(t, d)) \, dt}_{\int_{C_B} \mathbf{F} \cdot d\gamma - \int_{C_T} \mathbf{F} \cdot d\gamma}.$$

On the other hand,

$$\begin{aligned}\iint_D (Q_x - P_y) dA &= \iint_D Q_x dA - \iint_D P_y dA \\ &= \int_{y=c}^{y=d} \left(\int_{x=a}^{x=b} Q_x dx \right) dy - \int_{x=a}^{x=b} \left(\int_{y=c}^{y=d} P_y dy \right) dx.\end{aligned}$$

Applying the fundamental theorem of calculus to both inner integrals turns this last expression into the expression in Eq. 2.4. \square

We now provide a (nonrigorous) indication of how the general case of Green's theorem follows from the rectangle case. The idea is to find a collection of rectangles, $\{R_1, R_2, \dots, R_k\}$, whose union closely approximates D (as in Fig. 2.28), so that the following is a good approximation:

$$\iint_D (Q_x - P_y) dA \approx \iint_{R_1 \cup R_2 \cup \dots \cup R_k} (Q_x - P_y) dA = \sum_i \iint_{R_i} (Q_x - P_y) dA.$$

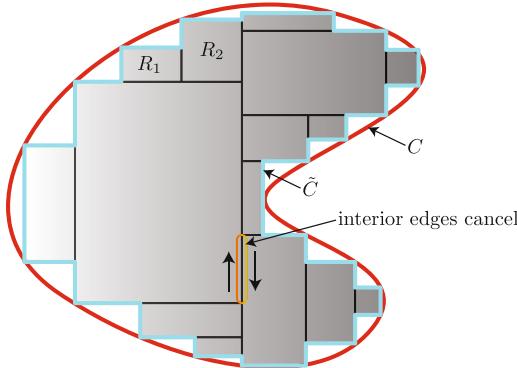


FIGURE 2.28. An “Etch A Sketch” approximation of C

Denote the boundaries of these rectangles by $\{C_1, C_2, \dots, C_k\}$, all oriented counterclockwise. Let \tilde{C} denote the outer edge (which is colored light blue in the figure and looks like something drawn on an Etch A Sketch). It is possible to ensure that the following is a good approximation: $\oint_{\tilde{C}} \mathbf{F} \cdot d\gamma \approx \oint_C \mathbf{F} \cdot d\gamma$. We won't prove this, but think about why you believe it; the arc lengths of \tilde{C} and C will *not* become close to each other as the picture is refined using more and more smaller rectangles, so why should the line integrals become close to each other?

Notice that $\sum_i \oint_{C_i} \mathbf{F} \cdot d\gamma = \oint_{\tilde{C}} \mathbf{F} \cdot d\gamma$, because all interior edges cancel. In other words, each interior edge receives opposite orientations from the two rectangles that share it. Since Green's theorem holds on each rectangle, we have

$$\oint_C \mathbf{F} \cdot d\gamma \approx \oint_{\tilde{C}} \mathbf{F} \cdot d\gamma = \sum_i \oint_{C_i} \mathbf{F} \cdot d\gamma = \sum_i \iint_{R_i} (Q_x - P_y) dA \approx \iint_D (Q_x - P_y) dA.$$

This completes our proof-sketch of Green's theorem.

We end this section by discussing the alternative flux version of Green's theorem. We begin with the concept of *flux*. Let C be a simple closed plane curve parametrized as $\gamma : [a, b] \rightarrow \mathbb{R}^2$. Let D be the interior of C . Let \mathbf{F} be a vector field defined on an open set containing $D \cup C$. For each $t \in [a, b]$, let $\mathbf{n}(t)$ denote the outward-pointing unit-length vector orthogonal to $\mathbf{v}(t)$, which is purple in Fig. 2.29.

DEFINITION 2.34.

The **flux** of \mathbf{F} across C is defined as

$$\text{flux} = \int_a^b \langle \mathbf{F}(\gamma(t)), \mathbf{n}(t) \rangle |\mathbf{v}(t)| dt.$$

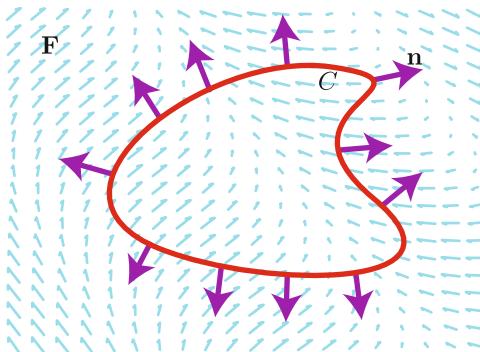


FIGURE 2.29. The flux measures the net outflow of \mathbf{F} across C

The flux of \mathbf{F} across C is independent of the parametrization of C . For a unit-speed parametrization, it is simply the integral of $\langle \mathbf{F}, \mathbf{n} \rangle$ along γ ; see Fig. 2.29. So the flux is positive if the vectors of \mathbf{F} encountered while traversing C mostly point outward. The flux is negative if they mostly point inward. For this reason, the flux is sometimes called the “net outflow” of \mathbf{F} across C .

Here is an imperfect but use-

ful metaphor: if C is a net, and \mathbf{F} represents the velocity vectors of a school of small fish, then the flux is roughly the net rate at which fish are escaping from the net. The flux would be negative if fish were mostly swimming into the net.

DEFINITION 2.35.

The **divergence** of the vector field $\mathbf{F} = (P, Q)$ is the real-valued function defined as $P_x + Q_y$ on the domain of \mathbf{F} .

The analogy with Definition 2.30 would be tighter if “divergence” were instead called “infinitesimal flux,” because of its geometric meaning:

COROLLARY 2.36.

If \mathbf{F} is a vector field defined in a neighborhood of $\mathbf{p} \in \mathbb{R}^2$, and C_r denotes the circle of radius r centered at \mathbf{p} , then

$$(P_x + Q_y)(\mathbf{p}) = \lim_{r \rightarrow 0} \frac{1}{\pi r^2} (\text{flux of } \mathbf{F} \text{ across } C_r).$$

The proof of Corollary 2.36 using the flux version of Green's theorem (which we are about to state) is essentially identical to our previous proof of Corollary 2.31 using the original version of Green's theorem.

THEOREM 2.37 (Flux Version of Green's Theorem).

Let C be a simple closed curve in the plane. Let D denote the interior of C . Let $\mathbf{F} = (P, Q)$ be a vector field defined on an open set containing $D \cup C$. Then,

$$(\text{flux of } \mathbf{F} \text{ across } C) = \iint_D (P_x + Q_y) dA.$$

Thus, the flux of \mathbf{F} across C equals the integral of the divergence over D . Returning to the fish metaphor, the divergence at a point roughly measures the extent to which fish are swimming away from that point (diverging from that point), so the theorem roughly says that net outflow of fish across a net equals the integral of the rate at which fish are diverging from all the points inside the net.

The following proof of Theorem 2.37 can be summarized as “rotate all of the vectors 90° and then apply Green's theorem.”

PROOF. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a unit-speed positively oriented parametrization of C . Denote the components of γ by $\gamma(t) = (x(t), y(t))$. Notice that

$$\mathbf{n}(t) = -R_{90}(\mathbf{v}(t)) = (y'(t), -x'(t)).$$

Let $\tilde{\mathbf{F}}$ denote the vector field obtained from \mathbf{F} by rotating all individual vectors 90° counterclockwise; that is, $\tilde{\mathbf{F}} = R_{90}(\mathbf{F}) = (-Q, P)$. We have

$$\begin{aligned} \text{flux} &= \int_a^b \langle \mathbf{F}(\gamma(t)), \mathbf{n}(t) \rangle dt = \int_a^b \langle (P(\gamma(t)), Q(\gamma(t))), (y'(t), -x'(t)) \rangle dt \\ &= \int_a^b \langle (-Q(\gamma(t)), P(\gamma(t))), (x'(t), y'(t)) \rangle dt \\ &= \int_a^b \langle \tilde{\mathbf{F}}(\gamma(t)), \mathbf{v}(t) \rangle dt \\ &= \underbrace{\oint_C \tilde{\mathbf{F}} \cdot d\gamma}_{\text{apply Green's theorem to } \tilde{\mathbf{F}}} = \iint_D (P_x + Q_y) dA. \end{aligned}$$

This completes the proof, but notice that the proof's heart can be concisely rewritten by expressing the integrand as

$$\langle \mathbf{F}, \mathbf{n} \rangle = \langle \mathbf{F}, -R_{90}\mathbf{v} \rangle = \langle R_{90}\mathbf{F}, -R_{90}^2\mathbf{v} \rangle = \langle R_{90}\mathbf{F}, \mathbf{v} \rangle = \left\langle \tilde{\mathbf{F}}, \mathbf{v} \right\rangle.$$

Here we're using the fact that the rotation map $R_{90} : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ is a rigid motion with the property that $R_{90} \circ R_{90} = -$ (the identity map). \square

EXERCISES

EXERCISE 2.17. Prove that line integrals are unchanged by orientation-preserving reparametrizations.

EXERCISE 2.18. Prove Proposition 2.26.

EXERCISE 2.19. What is the **escape velocity** of a rocket on the Earth's surface (the upward velocity needed to escape the planet's gravity, so that it will never be pulled back to the planet, neglecting air resistance and other forces)?

EXERCISE 2.20. For the vector field $\mathbf{F}(x, y) = (x, y + 2)$:

- (1) Calculate directly the line integral along the top half of the unit circle from $(1, 0)$ to $(-1, 0)$.
- (2) Calculate directly the line integral along the straight line from $(1, 0)$ to $(-1, 0)$.
- (3) Recalculate the above line integrals by finding a potential function for \mathbf{F} and applying Proposition 2.25.

EXERCISE 2.21. For the vector field $\mathbf{F}(x, y) = (2y + 3, x)$:

- (1) Calculate the line integral along the top half of the unit circle from $(1, 0)$ to $(-1, 0)$.
- (2) Calculate the line integral along the straight line from $(1, 0)$ to $(-1, 0)$.
- (3) Calculate the line integral around the loop that first traverses the top half of the unit circle from $(1, 0)$ to $(-1, 0)$ and then traverses the straight line from $(-1, 0)$ to $(1, 0)$. Solve this by subtracting the previous two answers, and also solve this using Green's theorem.

EXERCISE 2.22. Let $p_0 \in \mathbb{R}^n$ and let $f : \mathbb{R}^n \rightarrow \mathbb{R}$ denote the "distance to p_0 " function; that is, $f(p) = \text{dist}(p, p_0)$.

- (1) If $p \in \mathbb{R}^n$ with $p \neq p_0$, verify that $\nabla f(p) = \frac{p-p_0}{|p-p_0|}$.
- (2) Suppose that C is a simple closed curve in \mathbb{R}^n with $p_0 \notin C$. If $p \in C$ is the point of C that is closest to p_0 , prove that $p - p_0$ is orthogonal to the tangent line to C at p .

EXERCISE 2.23. With and without Green's theorem, show that a constant vector field has zero line integral around every circle in \mathbb{R}^2 .

EXERCISE 2.24. With and without Green's theorem, calculate the line integral of $\mathbf{F}(x, y) = (xy, x^2)$ around the triangle with vertices $(1, 1)$, $(1, 5)$, and $(3, 4)$.

EXERCISE 2.25. Prove the following partial converse to Lemma 2.29: If \mathbf{F} is a vector field whose domain is all of \mathbb{R}^2 with the property that $Q_x(p) = P_y(p)$ for all $p \in \mathbb{R}^2$, then \mathbf{F} is conservative.

EXERCISE 2.26. Consider the vector field \mathbf{F} with domain $\mathbb{R}^2 - \{(0, 0)\}$ defined as $\mathbf{F}(x, y) = \left(\frac{-y}{x^2+y^2}, \frac{x}{x^2+y^2}\right)$.

- (1) Verify that $Q_x - P_y = 0$ at every point of this domain.
- (2) Verify that the line integral of \mathbf{F} is *not* zero around a circle centered at the origin.
- (3) By Green's theorem, around what types of loops must the line integral of \mathbf{F} equal zero?

EXERCISE 2.27. Prove Green's theorem (Theorem 2.32) as a corollary of the flux version of Green's theorem (Theorem 2.37).

5. The Isoperimetric Inequality (*Optional*)

In this section, we prove the classic isoperimetric inequality in the plane as an application of Green's theorem.

We begin with a corollary of Green's theorem that is useful for computing area:

COROLLARY 2.38.

Let C be a positively oriented simple closed plane curve parametrized as $\gamma(t) = (x(t), y(t))$, $t \in [a, b]$. The area of the interior, D , of C equals

$$\text{Area}(D) = \int_a^b x(t)y'(t) dt = - \int_a^b y(t)x'(t) dt.$$

PROOF. To obtain these two formulas for area, apply Green's theorem separately to the two vector fields $\mathbf{F}_1(x, y) = (0, x)$ and $\mathbf{F}_2(x, y) = (-y, 0)$. Each of these vector fields has constant infinitesimal circulation of 1 and is illustrated in Fig. 2.26 on page 91. \square

The vector field $\mathbf{F}_1(x, y) = (0, x)$, which was used in the above proof, is also illustrated in Fig. 2.30. As you look at this image, think about how you could design a mechanical device that measures the area inside C when it is pushed around C . To measure the arc length of C would be much easier—a wheel on a stick with an odometer would suffice, like the one shown in Fig. 2.30 (left). But it should also be mechanically possible to measure the area by measuring $\oint_C \mathbf{F}_1 \cdot d\gamma = \int_a^b x(t)y'(t) dt$. The wheel and odometer would need to be modified to be sensitive only to the vertical component of motion, with sensitivity proportional to the distance to the y -axis. Exercise 2.33 discusses a **planimeter**, which is a mechanical device that uses Green's theorem to measure area (although not exactly in the manner suggested above).

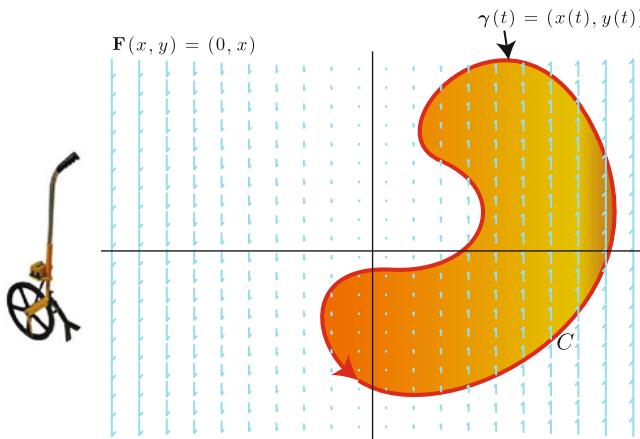


FIGURE 2.30. $\text{Area} = \oint_C \mathbf{F} \cdot d\gamma = \int_a^b x(t)y'(t) dt$

You can check that the length l and area A of a circle of any radius are related by the equation $l^2 = 4\pi A$ (here we're using “length” as an abbreviation for “arc length,” which is also often called “perimeter”). The remainder of this section is devoted to proving that every other simple closed curve has larger length than the circle that encloses the same area. More precisely, we have the following theorem.

THEOREM 2.39 (The Isoperimetric Inequality).

Let C be a simple closed plane curve. If l denotes the length of C and A denotes the area of the interior of C , then

$$l^2 \geq 4\pi A,$$

with equality if and only if C is a circle.

Thus, if C is not a circle, then its length is greater than the length of a circle with the same area as C . You could also read the theorem this way: if C is not a circle, then its area is less than the area of a circle with the same length as C . This second viewpoint justifies the term “isoperimetric” which means “same perimeter”—among all curves with the same perimeter, the circle has the largest area.

PROOF. Take two vertical lines that do not intersect C and move them together until they first touch C , so that C becomes tangent to both lines and lies between them. Let S^1 be a circle that is also tangent to both of these lines. Its radius, r , equals half the distance between the lines. Assume without loss of generality that the origin is the center of S^1 . Figure 2.31 shows C and S^1 as nonintersecting, but the vertical position of S^1 will be irrelevant for the proof.

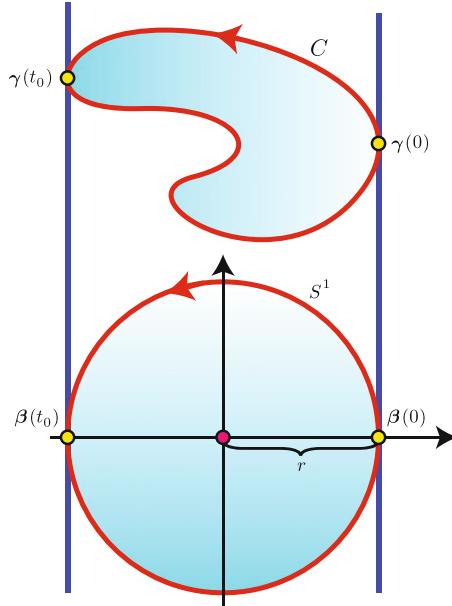


FIGURE 2.31. The proof of the isoperimetric inequality

Let $\gamma(t) = (x(t), y(t))$, $t \in [0, l]$, denote a parametrization of C by arc length such that $\gamma(0)$ is the intersection of C with one of the vertical lines, and $\gamma(t_0)$ is the intersection with the other for some $t_0 \in (0, l)$. We will choose a parametrization of S^1 , denoted by $\beta : [0, l] \rightarrow \mathbb{R}^2$, contrived so that for all $t \in [0, l]$, the x -coordinates of γ and β agree; that is, we will traverse S^1 so as to remain vertically aligned with someone who is traversing C at unit speed. This is achieved by defining $\beta(t) = (x(t), \tilde{y}(t))$, where

$$\tilde{y}(t) = \pm \sqrt{r^2 - x(t)^2},$$

with the sign depending on whether $t \in [0, t_0]$ or $t \in [t_0, l]$. Notice that β is not necessarily a regular parametrization, but this will not affect the calculations that follow.

By Corollary 2.38, the area of C is $A = \int_0^l x(t)y'(t) dt$, while the area of S^1 is $\pi r^2 = -\int_0^l \tilde{y}(t)x'(t) dt$ (check that this is valid even though the parametrization of C is not necessarily simple or regular). Adding these areas together (and suppressing the input variable) yields

$$\begin{aligned}
A + \pi r^2 &= \int_0^l (xy' - \tilde{y}x') dt \leq \int_0^l |xy' - \tilde{y}x'| dt = \int_0^l \sqrt{(xy' - \tilde{y}x')^2} dt \\
&= \int_0^l \sqrt{x^2 y'^2 - 2xy' \tilde{y}x' + \tilde{y}^2 x'^2} dt \\
&= \int_0^l \sqrt{(x^2 + \tilde{y}^2)(x'^2 + y'^2) - (xx' + \tilde{y}y')^2} dt \\
&\leq \int_0^l \sqrt{(x^2 + \tilde{y}^2)(x'^2 + y'^2)} dt \\
&= \int_0^l \sqrt{x^2 + \tilde{y}^2} dt \quad \text{because } \gamma \text{ is unit-speed} \\
&= \int_0^l |\beta(t)| dt = \int_0^l r dt = lr \quad \text{because } \beta(t) = (x(t), \pm \sqrt{r^2 - x(t)^2}).
\end{aligned}$$

In summary, $A + \pi r^2 \leq lr$. We now use the fact that the geometric mean of two positive numbers is bounded below by the arithmetic mean:

$$\sqrt{A}\sqrt{\pi r^2} \leq \frac{1}{2}(A + \pi r^2) \leq \frac{1}{2}lr.$$

This gives that $l^2 \geq 4\pi A$, as desired.

It remains to discuss the equality case in which $l^2 = 4\pi A$. In this case, all inequalities above become equalities. In particular, $(xx' + \tilde{y}y')^2 = 0$. Therefore,

$$0 = xx' + \tilde{y}y' = \langle (x, \tilde{y}), (x', y') \rangle = \langle \beta, \gamma' \rangle.$$

Since $|\beta|$ is constant, we also know that $\langle \beta, \beta' \rangle = 0$. Since both β' and γ' are orthogonal to β , they must be parallel to each other. But since they have identical first components (namely x'), this implies that they are equal to each other, at least when their common first component x' is nonzero.

In summary, $\beta'(t) = \gamma'(t)$ for all $t \in [0, l]$ at which $\gamma'(t)$ is not vertical. By continuity, the same must be true at isolated times when $\gamma'(t)$ is vertical. In fact, all such times must be isolated, for otherwise γ would be vertical on an interval, so β would have zero speed on that interval, yet would have unit speed everywhere it did not have zero speed, contradicting the continuity of its speed.

Thus, $\beta'(t) = \gamma'(t)$ for all $t \in [0, l]$. Since antiderivatives are unique up to an additive constant, $\gamma = \beta + \mathbf{w}$ for some constant vector $\mathbf{w} \in \mathbb{R}^2$. In other words, C is a translation of the circle. \square

EXERCISES

EXERCISE 2.28. If $l, A \in \mathbb{R}$ are positive numbers for which $l^2 \geq 4\pi A$, prove that there exists a simple closed plane curve with length l and area A .

EXERCISE 2.29. For $p > q > 0$, consider the ellipse $\gamma(t) = (p \cos t, q \sin t)$, exactly as in Exercise 2.4 on page 77. Let A denote its area and l its length.

- (1) Use Green's theorem to calculate A in terms of p and q .
- (2) Set $q = 1$, and use a computer algebra system to plot the graph of $\frac{4\pi A}{l^2}$ as a function of p . *COMMENT: a computer is necessary because the arc-length integral does not evaluate to an elementary closed-form expression for general p .*

EXERCISE 2.30. Use Green's theorem to find area of the region bounded by the x -axis and the trace of the curve

$$\gamma(t) = (t - \sin(t), 1 - \cos(t)), \quad t \in [0, 2\pi].$$

COMMENT: This curve is called a “cycloid” and is illustrated in the next section.

EXERCISE 2.31.

- (1) Show that the line integral of $\mathbf{F}(x, y) = (-y, x)$ along the line segment from (x_1, y_1) to (x_2, y_2) equals $x_1 y_2 - x_2 y_1$.
- (2) If C is a polygon with vertices denoted by $(x_1, y_1), \dots, (x_n, y_n)$ (ordered counterclockwise), prove that the area A of the polygon is given by

$$2A = (x_1 y_2 - x_2 y_1) + (x_2 y_3 - x_3 y_2) + \cdots + (x_{n-1} y_n - x_n y_{n-1}) + (x_n y_1 - x_1 y_n).$$

EXERCISE 2.32. Describe the history of isoperimetric problems including alternative proofs. An excellent reference is [3].

EXERCISE 2.33. A **planimeter** is a mechanical drafting instrument used to determine the area enclosed in a region; see the figure on page 61. Describe how these devices work and how their function is related to Green's theorem.

6. Huygens's Tautochrone Clock (*Optional*)

For the major countries of Europe, the seventeenth and eighteenth centuries represented an era of naval exploration. The main impediment to navigation at sea was the **longitude problem**—while it was relatively easy to establish one's latitude at sea, there was no effective known method to determine one's longitude. This limitation resulted in countless maritime disasters. As the lost life and lost gold piled up, several countries offered prizes and established observatories and scientific centers devoted to solving the problem. For example, in 1714, the British government established a *Board of Longitude* and offered a *longitude prize* of 20,000 pounds, writing:

The discovery of the longitude is of such consequence to Great Britain for the safety of the navy and merchant ships as well as for the improvement of trade that for want thereof many ships have been retarded in their voyages, and many lost...

Potential solution methods based on astronomical observations were championed by Galileo, Newton, Halley, and others. Work in this direction led to several key scientific discoveries including the speed of light. The main alternative approach involved attempting to build a better clock. If a clock could be constructed that maintained accurate time during long voyages at sea, then longitude could be determined each day from the exact time of sunrise or high noon.

Christiaan Huygens is credited with inventing pendulum clocks, which could keep accurate time on land but not at sea. The motion of waves rocking a ship would render a pendulum erratic—on some swings, the pendulum bob would traverse a wider circular arc than others. A wider swing takes slightly more time than a narrow one, eventually leading to inaccurate readings.

In fact, Huygens was the first to prove that a circular arc is not quite isochronous (wide swings take slightly more time than narrow swings). The best way to model this phenomenon is to forget about the string. Pretend that the circular arc traced by the pendulum bob is an actual physical track (colored blue in Fig. 2.32) and that the bob is a frictionless object sliding down the track. Huygens proved that the time to reach the bottom increases when the bob begins higher up the track. This might sound obvious, but it's not true of every track shape. He discovered a track shape with the property that the time to reach the bottom is independent of the object's starting position on the track. He called this shape a **tautochrone curve** (Greek for “same time”).

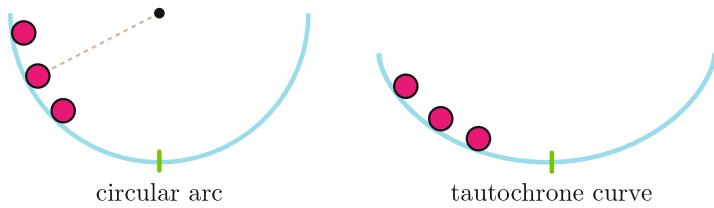


FIGURE 2.32. On the circular arc, the higher object takes the most time to slide to the bottom. On the tautochrone curve, all objects take the same amount of time

Huygens's next idea was to build bumpers against which the pendulum string would wrap, causing the pendulum bob to traverse a tautochrone curve rather than a circular arc; see Fig. 2.33. He constructed a **tautochrone clock** based on this design, with hopes of solving the longitude problem. He believed that his clock would keep accurate time at sea because all pendulum swings would take the same amount of time, even though the waves would cause some swings to be wider than others. Unfortunately, his clock did not function as accurately as he hoped on its trial voyage. Perhaps the added friction of the string against the bumpers offset the theoretical advantages, or perhaps the storms were severe enough to cause the bob to jerk about.

The longitude prize was eventually awarded to John Harrison for designing a seaworthy clock based on springs and balances.²

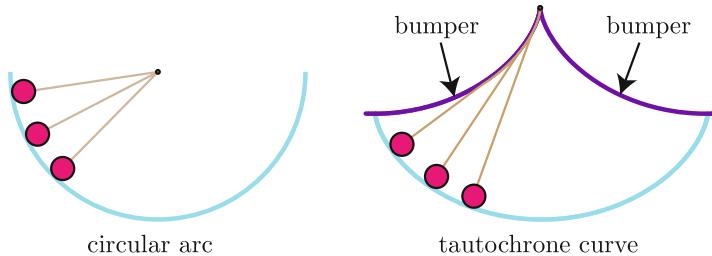


FIGURE 2.33. The pendulum string wraps against the bumpers, causing the bob to traverse a tautochrone curve

Although Huygens failed to win the longitude prize, the mathematics he invented to design his tautochrone clock survived and later found other interesting applications. The remainder of this section is devoted to discussing this mathematics and some of these applications.

The mathematical story begins with the *cycloid*. Imagine a wheel of radius 1 initially centered at $(0, 1)$, so it is tangent to the x -axis at the origin. Imagine marking this point of tangency on the wheel with a chalk mark and then letting the wheel roll without slipping along the x -axis. The curve traced by the chalk mark is called the **cycloid**; see Fig. 2.34. Parametrizing the cycloid is a simple matter of adding together the two purple vectors:

$$(2.5) \quad \gamma(t) = (t, 1) + (-\sin(t), -\cos(t)) = (t - \sin(t), 1 - \cos(t)).$$

The cycloid is regular on the domain $(0, 2\pi)$, which corresponds to the portion shown in Fig. 2.34.

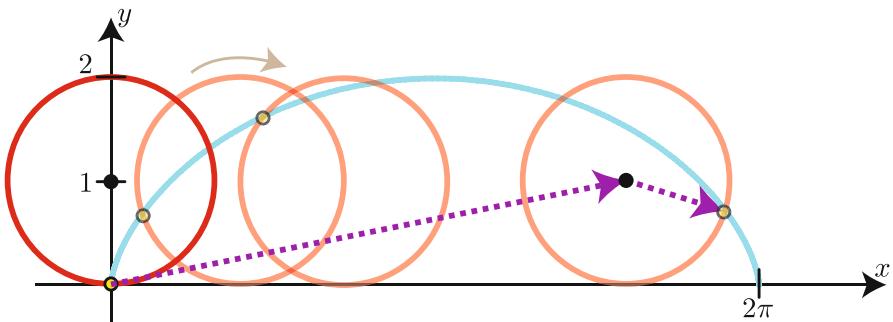


FIGURE 2.34. A cycloid is the path traced by a chalk mark on the edge of a wheel that rolls along a straight line

²We recommend [7] for the history of the longitude problem and John Harrison's story.

The cycloid is turned upside down by the rigid motion $(x, y) \mapsto (x, -y)$. The resultant **inverted cycloid** is parametrized as $\hat{\gamma}(t) = (\underbrace{t - \sin(t)}_x, \underbrace{\cos(t) - 1}_y)$, $t \in (0, 2\pi)$, is a tautochrone curve.

THEOREM 2.40.

The inverted cycloid, $\hat{\gamma}(t) = (\underbrace{t - \sin(t)}_x, \underbrace{\cos(t) - 1}_y)$, $t \in (0, 2\pi)$, is a tautochrone curve.

PROOF. Consider an object beginning at $\hat{\gamma}(t_0) = (x_0, y_0)$ and sliding under the influence only of gravity (without friction) to the bottom of the curve $\hat{\gamma}(\pi)$. We must show that the time required is independent of t_0 . Since the right half of the curve is the mirror image of the left, it suffices to assume that $t_0 \in (0, \pi)$.

Our first job is to relate the following variables that change as the curve is traversed: t, x, y, s, T . Here, s is the arc-length parameter defined as $s(t) = \int_{t_0}^t |\hat{\gamma}'(t)| dt$, while T is the time parameter, so that $T(t)$ denotes the time required to slide from position $\hat{\gamma}(t_0)$ to position $\hat{\gamma}(t)$. Notice that

$$\left(\frac{ds}{dt}\right)^2 = \left|\frac{d\hat{\gamma}}{dt}\right|^2 = \left(\frac{dx}{dt}\right)^2 + \left(\frac{dy}{dt}\right)^2 = (1 - \cos(t))^2 + (-\sin(t))^2 = 2 - 2\cos(t).$$

Kinetic energy is traded for potential energy such that the following is always true:

$$(2.6) \quad \frac{1}{2}mv^2 = mg(y_0 - y),$$

where g is the Earth's gravitational constant (Exercise 2.35). Here v is the object's speed, measured as

$$\frac{ds}{dT} = v = \sqrt{2g(y_0 - y)}.$$

The chain rules says that $\frac{ds}{dt} = \frac{ds}{dT} \cdot \frac{dT}{dt}$. Solving for $\frac{dT}{dt}$ gives

$$\frac{dT}{dt} = \frac{\left(\frac{ds}{dt}\right)}{\left(\frac{ds}{dT}\right)} = \frac{\sqrt{2 - 2\cos(t)}}{\sqrt{2g(y_0 - y)}} = \frac{\sqrt{2 - 2\cos(t)}}{\sqrt{2g(\cos(t_0) - \cos(t))}}.$$

So the time required to reach the bottom is

$$T(\pi) = \int_{t_0}^{\pi} \left(\frac{dT}{dt}\right) dt = \int_{t_0}^{\pi} \frac{\sqrt{2 - 2\cos(t)} dt}{\sqrt{2g(\cos(t_0) - \cos(t))}} = \frac{1}{\sqrt{g}} \int_{t_0}^{\pi} \sqrt{\frac{1 - \cos(t)}{\cos(t_0) - \cos(t)}} dt.$$

Using integration tricks or computer assistance, this final integral can be shown to be independent of t_0 , and it evaluates to π , so we have $T(\pi) = \frac{\pi}{\sqrt{g}}$. \square

Bernoulli and Euler later proved the converse: every tautochrone curve is a segment of the inverted cycloid (possibly resized or translated).

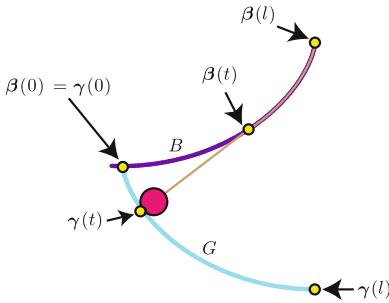


FIGURE 2.35. G is an involute of B with $(\lambda_0 = 0)$

arc length. At every $t \in [0, l]$, the length of unwrapped string equals t , so the bob's position has distance t from $\beta(t)$ in the direction of $-\beta'(t)$; see Fig. 2.35. That is, $\boxed{\gamma(t) = \beta(t) - t\beta'(t)}$.

The following definition slightly generalizes the above discussion by allowing an initial tail of unwrapped string with length λ_0 .

DEFINITION 2.41.

Let B be an oriented plane curve with nonzero curvature. An **involute** of B is a plane curve G that has a parametrization $\gamma : (0, l) \rightarrow \mathbb{R}^2$ of the form

$$(2.7) \quad \gamma(t) = \beta(t) - (t + \lambda_0)\beta'(t),$$

where $\beta : (0, l) \rightarrow \mathbb{R}^2$ is a unit-speed parametrization of B , and $\lambda_0 \in \mathbb{R}$ is a constant with $\lambda_0 \notin (-l, 0)$.

Notice that $\gamma'(t) = -(t + \lambda_0)\beta''(t)$, which is nonzero because $\lambda_0 \notin (-l, 0)$ and because B has nonzero curvature, so γ is regular. For example, the nonzero curvature hypothesis does not allow B to be a straight line, because then G would be a single point. The case $\lambda_0 \geq 0$ corresponds to an “unwrapping string,” as was previously illustrated in Fig. 2.35. The case $\lambda_0 \leq -l$ corresponds to a string that is wrapped along B .

If $\beta : (0, l) \rightarrow \mathbb{R}^2$ is a regular (not necessarily unit-speed) parametrization of B , one can avoid reparametrizing it by arc length by replacing Eq. 2.7 with

$$(2.8) \quad \gamma(t) = \beta(t) - (s(t) + \lambda_0)\frac{\beta'(t)}{|\beta'(t)|},$$

where $s(t) = \int_0^t |\beta'(u)| du$ is the arc-length function.

Huygens's second mathematical problem was to determine the shape of the bumpers. For this, let $\beta : [0, l] \rightarrow \mathbb{R}^2$ be a parametrized plane curve (the bumper curve) and let $\gamma : [0, l] \rightarrow \mathbb{R}^2$ be the induced parametrization of the path that the pendulum bob will follow. It is simplest to imagine that initially at $t = 0$, the string is completely wrapped along the bumper curve, ending at $\beta(0)$, and is about to begin unwrapping. So the initial position of the pendulum bob is $\gamma(0) = \beta(0)$. Since the trace of γ will depend only on the trace of β , we can assume that β is parametrized by

Huygens really needed to solve the inverse problem: find the bumper curve B whose involute is the tautochrone curve G . He solved this by understanding the inverse relationship between involutes and evolutes. Here is a more precise formulation of Eq. 1.8 on page 30:

DEFINITION 2.42.

Let G be an oriented plane curve parametrized as $\gamma : (0, l) \rightarrow \mathbb{R}^2$. Assume for all $t \in (0, l)$ that $\kappa(t) \neq 0$ and $\kappa'(t) \neq 0$. Let \mathbf{n} denote the unit normal to γ . The **evolute**, B , of G is the plane curve with parametrization $\beta : (0, l) \rightarrow \mathbb{R}^2$ given by

$$\beta(t) = \gamma(t) + \frac{1}{\kappa(t)} \mathbf{n}(t) = \text{the center of the osculating circle of } G \text{ at } \gamma(t).$$

You can check that $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$, so the hypothesis that κ' is nonvanishing ensures that β is regular. For example, this hypothesis does not allow G to be a circle, because then B would be a single point—its center. The variable choices have already hinted at the following inverse relationship between involutes and evolutes:

PROPOSITION 2.43.

- (1) With the assumptions of Definition 2.41, B is the evolute of G .
- (2) With the assumptions of Definition 2.42, G is an involute of B .

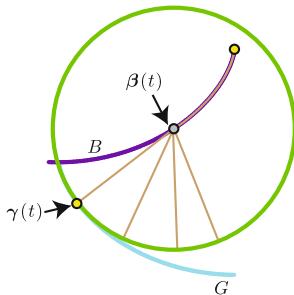


FIGURE 2.36. Pinning the string to $\beta(t)$ at time t makes the pendulum bob begin traversing the osculating circle of G at $\gamma(t)$

Before beginning the proof, we mention a visual reason to believe claim (1) that B is the evolute of G in Fig. 2.35. If the string were pinned to $\beta(t)$ at time t , then the pendulum bob would begin traversing the circle of radius $t + \lambda_0$ centered at $\beta(t)$, colored green in Fig. 2.36. This circle approximates G well at $\gamma(t)$ (in fact, it is the osculating circle) because second derivatives don't detect the difference between pinning and not pinning. This suggests that $\beta(t)$ is the center of the osculating circle of G at $\gamma(t)$.

PROOF OF PROPOSITION 2.43(2). To prove part (2), let G be an oriented plane curve parametrized as $\gamma : (0, l) \rightarrow \mathbb{R}^2$, and assume that its curvature function satisfies $\kappa(t) \neq 0$ and $\kappa'(t) \neq 0$ for all $t \in (0, l)$. We will consider the case that κ' is strictly negative (the other case, that κ' is strictly

positive, is handled similarly). The evolute of G is the curve B parametrized as

$$(2.9) \quad \beta(t) = \gamma(t) + \frac{1}{\kappa(t)} \mathbf{n}(t).$$

As previously mentioned, $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$. In particular, $\frac{\beta'(t)}{|\beta'(t)|} = \mathbf{n}(t)$; see Fig. 2.37. Thus, Eq. 2.9 can be rewritten as

$$\gamma(t) = \beta(t) - \frac{1}{\kappa(t)} \frac{\beta'(t)}{|\beta'(t)|}.$$

Comparing to Eq. 2.8, to prove that G is an involute of B , it will suffice to verify that

$$\frac{1}{\kappa(t)} = s(t) + \lambda_0$$

for some $\lambda_0 \notin (-l, 0)$, where $s(t) = \int_0^t |\beta'(u)| du$ is the arc-length function of β . For this, observe that

$$\frac{d}{dt} \left(\frac{1}{\kappa(t)} - s(t) \right) = -\frac{\kappa'(t)}{\kappa(t)^2} - |\beta'(t)| = -\frac{\kappa'(t)}{\kappa(t)^2} - \left| \frac{\kappa'(t)}{\kappa(t)^2} \right| = 0.$$

Thus, $\frac{1}{\kappa(t)} - s(t)$ is a constant, and by considering the time $t = 0$, one confirms that this constant is ≥ 0 .

The proof of part (1) is left to the reader in Exercise 2.44. □

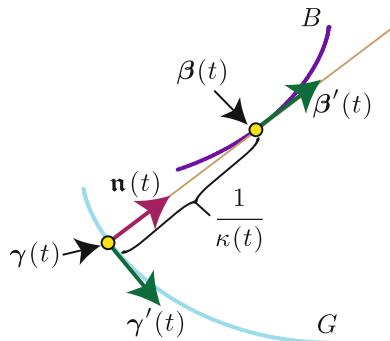


FIGURE 2.37. B is the evolute of G ; therefore, G is an involute of B

In light of Proposition 2.43, Huygens could choose his bumper curve to be the evolute of the tautochrone curve (the inverted cycloid). Figure 2.38 zooms out to show more of the curves G and B from Fig. 2.33, leading one to guess the following:

PROPOSITION 2.44.

The evolute of the inverted cycloid is a translation of the inverted cycloid.

PROOF. Exercise 2.46. □

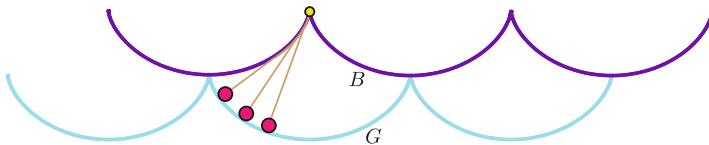


FIGURE 2.38. The evolute, B , of G is a translation of G

The intended interpretation of Proposition 2.44 should be clear from Fig. 2.38. But to be pedantic about satisfying the hypotheses of Definition 2.42, we should restrict the inverted cycloid to the domain $(0, \pi)$; in other words, we should consider only the left half of one period of the inverted cycloid, in which case its evolute is a translation of the right half.

Huygens's tautochrone clock did not win the longitude prize, but the underlying mathematics has found other interesting applications through the centuries. For example, evolutes are important in the field of optics. To understand the relevance, consider an alternative method of illustrating the evolute of the inverted cycloid G . Figure 2.39 (left) shows G together with its *normal line* (the line in the direction of \mathbf{n}) at 40 points along G . These 40 purple lines intersect in a pattern that tricks your eye into seeing the curved shape of the evolute, B , of G . If the purple lines are light rays, then B is a curve of focused brightness called a *caustic* (Latin for “burn,” because focused sun rays can burn). Intuitively, the normal lines of G should focus along B because G is well approximated at each point by its osculating circle, whose normal lines all focus on its center.

To describe this focusing more precisely, let \mathcal{F} denote the family of all normal lines (not just at 40 points, but at all points of G). Then B is called the **envelope** of \mathcal{F} , which means the unique curve whose tangent lines are all members of \mathcal{F} . In computer graphics, modern ray-tracing software is capable of rendering envelopes as bright caustics. In Fig. 2.39 (right), the bright caustic is the envelope of a certain family, \mathcal{F} , of lines emanating from the curved mirror. In contrast to our cycloid example, these lines are not normal to the mirror, but rather point in the directions that appropriately model how light rays from a single source would bounce off the mirror.

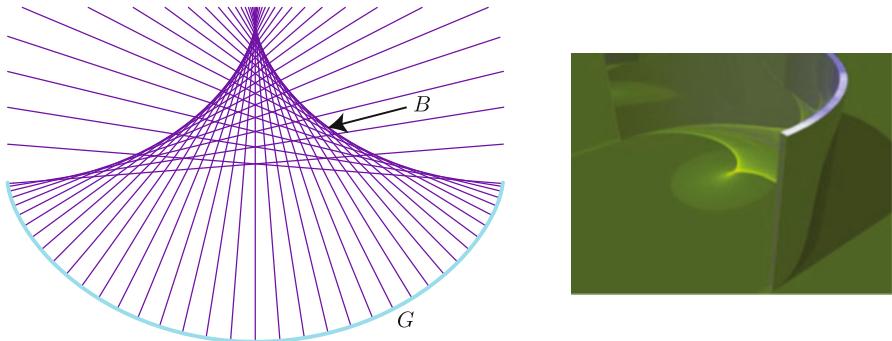


FIGURE 2.39. *Left:* the evolute, B , of G is the envelope of its normal lines. *Right:* the bright caustic is the envelope of the light rays reflected off the curved mirror

Another application involves the shape of gears. Old-fashioned gears with triangular or rectangular teeth clanged against each other with excessive vibration and noise. To solve this problem, Euler invented “involute gears” whose edges have the shape of a segment of an involute of a circle. This shape ensures that a pair of interlocking teeth meet each other at a single contact point that varies along a straight line as the gears turn, greatly reducing vibration and wear. Most gears in use today are involute gears.

We hope that the illustrations in this section were helpful, but don’t settle for still images. It is easy to find online animated illustrations of these concepts. We particularly recommend the animations found on the Wikipedia pages for “cycloid,” “tautochrone curve,” “involute,” “evolute,” and “involute gear.” Also check out the animated tautochrone clock on the Wolfram MathWorld “Tautochrone Problem” page.

EXERCISES

EXERCISE 2.34. The cycloid in Eq. 2.5 can be generalized as

$$\gamma(t) = (at - b \sin(t), a - b \cos(t))$$

for arbitrary constants $a, b > 0$.

- (1) Interpret in terms of the path of a chalk mark on (not necessarily the edge of) a rolling wheel.
- (2) Use a computer graphing application to plot the graph for several choices with $a < b$ and several choices with $a > b$.

EXERCISE 2.35. How can Eq. 2.6 be derived from the conservation of energy law in Example 2.28 on page 89?

EXERCISE 2.36. Verify the claim after Definition 2.42 that $\beta'(t) = -\frac{\kappa'(t)}{\kappa(t)^2} \mathbf{n}(t)$.

EXERCISE 2.37. Definition 2.42 defines only the evolute of a *plane* curve. The evolute of a *space* curve $\gamma : (0, l) \rightarrow \mathbb{R}^3$ can be defined by the same formula: $\beta(t) = \gamma(t) + \frac{1}{\kappa(t)}\mathbf{n}(t)$. Modify the formula for $\beta'(t)$ from the previous exercise to make it valid for space curves. Describe the general condition under which β is regular for space curves. Determine the evolute of the helix from Example 1.3 on page 2.

EXERCISE 2.38. Let $\beta : (0, l) \rightarrow \mathbb{R}^2$ be a unit-speed plane curve with nonzero curvature. Let $\gamma : (0, l) \rightarrow \mathbb{R}^2$ be a regular plane curve such that for all $t \in (0, l)$, $\gamma(t)$ meets the tangent line to $\beta(t)$ at a right angle (that is, the tangent line to the trace of β at $\beta(t)$ contains $\gamma(t)$ and is orthogonal to $\gamma'(t)$). Prove that γ is an involute of β .

EXERCISE 2.39. Prove that the evolute of the parabola $y = x^2$ is the graph of $y = \frac{1}{2} + 3|\frac{x}{4}|^{2/3}$, illustrated in Fig. 2.40.

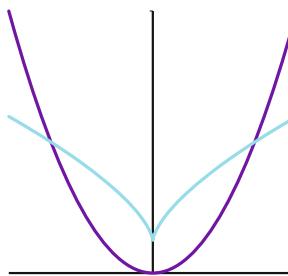


FIGURE 2.40. The evolute of the parabola

EXERCISE 2.40. Let $p > q > 0$ and consider the ellipse $\gamma(t) = (p \cos(t), q \sin(t))$, $t \in [0, 2\pi]$. prove that its evolute is an astroid (defined in Exercise 1.12 on page 8.)

EXERCISE 2.41. For the parabola and ellipse from the previous two exercises, construct a graph similar to Fig. 2.39 showing a family of normal lines intersecting to form the shape of the evolute.

EXERCISE 2.42. Describe an infinite family of plane curves that all have the same evolute.

EXERCISE 2.43. Use a computer graphing application to plot several involutes of a circle of radius 1.

EXERCISE 2.44. Prove Proposition 2.43(1).

EXERCISE 2.45. If the string of Huygens's tautochrone clock is lengthened, will the bob still traverse a tautochrone curve?

EXERCISE 2.46. Prove Proposition 2.44.

EXERCISE 2.47. An **epicycloid** (respectively **hypocycloid**) is the path followed by a chalk mark on the rim of a circle of radius b that rolls without slipping outside (respectively inside) a circle of radius a , as illustrated

in Fig. 2.41. Find formulas for these curves and use a computer graphing application to plot them for several choices of the constants a, b with $b < a$.

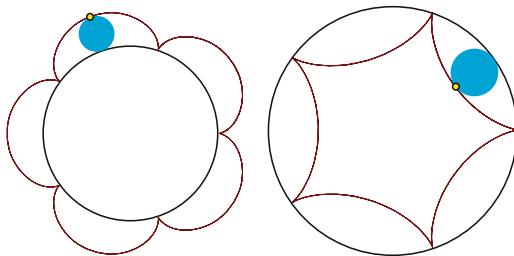


FIGURE 2.41. An epicycloid (*left*) and a hypocycloid (*right*)





This 1632 illustration by Rubens represents stereographic projection—the shadowing of the curved sphere onto the flat plane from a single light source. In this chapter, surfaces will be studied using the natural coordinates arising from projections like this one.

Surfaces

For the remainder of the book, we turn our attention from curves to surfaces. The graphs of functions of two variables are familiar examples of surfaces from multivariable calculus. Whereas a curve locally looks like its tangent line, \mathbb{R}^1 , a surface locally looks like its tangent plane, \mathbb{R}^2 . Thus, we are moving from intrinsically one-dimensional to intrinsically two-dimensional objects.

1. The Derivative of a Function from \mathbb{R}^m to \mathbb{R}^n

In this section, we define and study derivatives of functions with multiple input *and* multiple output variables. The exposition is intended as an overview of the key definitions and visual ideas for the benefit of readers who may previously have considered only derivatives of functions with multiple input *or* multiple output variables. Proofs will be omitted when their techniques are not necessary for understanding the remainder of this book. Omitted proofs can be found in most real analysis textbooks.

We will henceforth write $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ to mean that f is a function from the open set U of \mathbb{R}^m into \mathbb{R}^n . In this case, f can be thought of as n separate **component functions**; we write $f = (f_1, \dots, f_n)$, where for each

i , we have $f_i : U \rightarrow \mathbb{R}$. For example, the function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as $f(x, y) = (\sin(xy), e^{x+5y}, x^2y^3)$ has component functions $f_1(x, y) = \sin(xy)$, $f_2(x, y) = e^{x+5y}$, and $f_3(x, y) = x^2y^3$.

The partial derivative of f with respect to the input variable x_i (denoted by f_{x_i} or by $\frac{\partial f}{\partial x_i}$) is defined such that for all $p \in U$,

$$(3.1) \quad f_{x_i}(p) = \frac{\partial f}{\partial x_i}(p) = \lim_{t \rightarrow 0} \frac{f(p + te_i) - f(p)}{t}$$

(if this limit exists), where $e_i \in \mathbb{R}^m$ denotes the i th member of the standard orthonormal basis.¹ The visual meaning will be explained shortly, but for now, notice that partial derivatives can be computed componentwise:

$$(3.2) \quad \frac{\partial f}{\partial x_i}(p) = \left(\frac{\partial f_1}{\partial x_i}(p), \dots, \frac{\partial f_n}{\partial x_i}(p) \right).$$

Each component has the familiar meaning from multivariable calculus (the rate of change of that component at p as x_i increases) and is computed in the familiar manner (treat all input variables other than x_i as constants).

For fixed i , if $\frac{\partial f}{\partial x_i}(p)$ exists at each $p \in U$, then $p \mapsto \frac{\partial f}{\partial x_i}(p)$ is another function from U to \mathbb{R}^n ; its partial derivatives (if they exist) are called second-order partial derivatives of f , and so on. For example, the partial derivative of $f_{x_i} = \frac{\partial f}{\partial x_i}$ with respect to x_j is denoted by $f_{x_i x_j}$ or by $\frac{\partial^2 f}{\partial x_j \partial x_i}$ (if it exists). A crucial result is that the order does not matter here; that is, **mixed partials commute**:

$$(3.3) \quad \boxed{f_{x_i x_j} = f_{x_j x_i}}$$

(provided both sides of this equation exist and are continuous on U). We'll use Eq. 3.3 many times throughout the remainder of the book—differential geometry wouldn't work without it!

The function f is called “ **C^r on U** ” if all r th-order partial derivatives exist and are continuous on U , and f is called “**smooth on U** ” if f is C^∞ on U for all positive integers r .

EXAMPLE 3.1. *The smooth function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as*

$$f(x, y) = (\sin(xy), e^{x+5y}, x^2y^3)$$

has the following first-order partial derivatives,

$$f_x = (y \cos(xy), e^{x+5y}, 2xy^3), \quad f_y = (x \cos(xy), 5e^{x+5y}, 3x^2y^2),$$

¹For the remainder of the book, we will discontinue the use of boldface typesetting for the names of vector-valued quantities. This new convention is more common and more natural-looking for discussing surfaces.

and has the following second-order partial derivatives,

$$\begin{aligned} f_{xx} &= (-y^2 \sin(xy), e^{x+5y}, 2y^3), & f_{yy} &= (-x^2 \sin(xy), 25e^{x+5y}, 6x^2y), \\ f_{xy} &= f_{yx} = (\cos(xy) - xy \sin(xy), 5e^{x+5y}, 6xy^2). \end{aligned}$$

Notice that a function is smooth on its domain if and only if its component functions are smooth on this domain. In the above example, the component functions are smooth because the familiar classes of functions from multivariable calculus are known to be smooth on their domains (polynomial, rational, exponential, logarithmic, trigonometric, and inverse-trigonometric; also power functions except possibly at zero). The sum, difference, or product of two smooth functions on the same domain is a smooth function on this domain (the same is true for quotients, excluding points where the denominator vanishes). Furthermore, every composition of smooth functions is smooth. We state this more precisely:

PROPOSITION 3.2.

If $g : U \subset \mathbb{R}^l \rightarrow \mathbb{R}^m$ is smooth on U , $f : V \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ is smooth on V , and $g(U) \subset V$, then the composition $f \circ g : U \rightarrow \mathbb{R}^n$ is smooth on U .

For example, the above function $f(x, y) = (\sin(xy), e^{x+5y}, x^2y^3)$ could be considered the composition of the smooth functions $(x, y) \mapsto (xy, x+5y, x^2y^3)$ and $(u, v, w) \mapsto (\sin(u), e^v, w)$.

The definition of a partial derivative in Eq. 3.1 can be generalized by replacing the coordinate vector e_i with an arbitrary vector $v \in \mathbb{R}^m$; the result is called a *directional derivative*.²

DEFINITION 3.3.

Let $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$, $p \in U$, and $v \in \mathbb{R}^m$. The **directional derivative** of f in the direction v at p is

$$df_p(v) = \lim_{t \rightarrow 0} \frac{f(p + tv) - f(p)}{t} = (f \circ \gamma)'(0)$$

(if this limit exists), where $\gamma(t) = p + tv$.

In other words, $df_p(v)$ is the initial velocity vector of the composition with f of the straight line γ in \mathbb{R}^m that at time $t = 0$ passes through p with velocity v ; see Fig. 3.1. In the special case $v = e_i$, this provides a good visual interpretation of the partial derivative $\frac{\partial f}{\partial x_i}(p) = df_p(e_i)$.

²Some books reserve the name “directional derivative” for the special case $|v| = 1$, but we will use it generally.

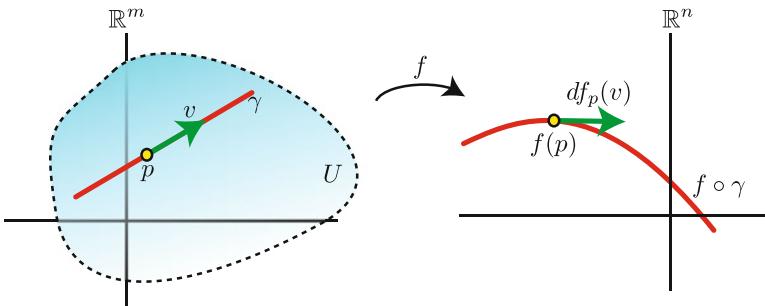


FIGURE 3.1. $df_p(v)$ is the initial velocity vector of the composition with f of the straight line γ that at time $t = 0$ passes through p with velocity v

If the directional derivative $df_p(v) = \lim_{t \rightarrow 0} \frac{f(p+tv) - f(p)}{t}$ exists, then this limit can be reexpressed as the claim that

$$(3.4) \quad \lim_{t \rightarrow 0} \frac{|f(p+tv) - f(p) - t df_p(v)|}{t} = 0,$$

which really just says that

$$(3.5) \quad f(p+tv) - f(p) \approx t df_p(v)$$

is a good approximation of f near p .

If f is smooth on a neighborhood of p , then an important result from analysis says not only that all directional derivatives at p exist, but that they fit together into a linear transformation that uniformly approximates f well near p :

PROPOSITION 3.4.

If $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ is smooth (or even just C^1) on U , then for all $p \in U$:

- (1) *The directional derivative $df_p(v)$ exists for all $v \in \mathbb{R}^m$.*
- (2) *$v \mapsto df_p(v)$ is a linear transformation from \mathbb{R}^m to \mathbb{R}^n (which we call the **derivative of f at p**).*
- (3) *$f(q) \approx f(p) + df_p(q-p)$ is a good approximation of f near p in the following sense: for every infinite sequence $\{q_1, q_2, \dots\}$ of points in U converging to p , we have*

$$\lim_{t \rightarrow \infty} \frac{|f(q_t) - f(p) - df_p(q_t - p)|}{|q_t - p|} = 0.$$

Notice that condition (3) would become equivalent to Eq. 3.4 if all of the points q_t were in the span of the single unit-length vector $v \in \mathbb{R}^m$. Thus, condition (3) essentially says that Eq. 3.5 approximates f near p uniformly well over all unit-length choices of v .

Since the derivative df_p is a linear transformation, we must pause to review the manner in which every linear transformation can be represented as left multiplication by some matrix. If A is an $n \times m$ matrix, recall that $L_A : \mathbb{R}^m \rightarrow \mathbb{R}^n$ denotes the “left-multiplication by A ” linear transformation. It is defined such that $L_A(v) = A \cdot v$ for all $v \in \mathbb{R}^m$, where v is considered to be an $m \times 1$ column matrix, so that the matrix multiplication $A \cdot v$ makes sense. Notice that the i th column of A equals $L_A(e_i)$. In particular, the linear transformation $df_p : \mathbb{R}^m \rightarrow \mathbb{R}^n$ is represented by the matrix whose i th column equals $df_p(e_i) = \frac{\partial f}{\partial x_i}(p) = \left(\frac{\partial f_1}{\partial x_i}(p), \dots, \frac{\partial f_n}{\partial x_i}(p) \right)$. In summary:

PROPOSITION AND DEFINITION 3.5.

If $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ is smooth and $p \in U$, then $df_p = L_A$, where A is the $n \times m$ matrix of all first-order partial derivatives of f at p :

$$A = \begin{pmatrix} \frac{\partial f_1}{\partial x_1}(p) & \cdots & \frac{\partial f_1}{\partial x_m}(p) \\ \vdots & \ddots & \vdots \\ \frac{\partial f_n}{\partial x_1}(p) & \cdots & \frac{\partial f_n}{\partial x_m}(p) \end{pmatrix}.$$

This matrix A is called the **Jacobian matrix** of f at p . The fact that $df_p = L_A$ is often abbreviated as “ $df_p = A$.”

Notice that when $n = 1$, $A = \left(\frac{\partial f}{\partial x_1}(p) \quad \cdots \quad \frac{\partial f}{\partial x_m}(p) \right)$ is the gradient of f at p (expressed as a $1 \times m$ matrix rather than a vector, which is essentially the same thing). In this case, we have for all $v = (v_1, \dots, v_m) \in \mathbb{R}^m$,

$$\begin{aligned} df_p(v) &= L_A(v) = \left(\frac{\partial f}{\partial x_1}(p) \quad \cdots \quad \frac{\partial f}{\partial x_m}(p) \right) \cdot \begin{pmatrix} v_1 \\ \vdots \\ v_m \end{pmatrix} \\ &= \frac{\partial f}{\partial x_1}(p)v_1 + \cdots + \frac{\partial f}{\partial x_m}(p)v_m = \langle \nabla f(p), v \rangle. \end{aligned}$$

This recovers the familiar fact (previously mentioned in Eq. 2.3 on page 86) that directional derivatives are computed by taking the inner product with the gradient. If $n > 1$, notice that the rows of A are the gradients of the components of f , so each component of $df_p(v)$ equals the inner product of v with the gradient of that component of f . In particular, directional derivatives are computed componentwise.

EXAMPLE 3.6. The conversion from polar to rectangular coordinates is encoded in the smooth function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined as $f(r, \theta) = (\underbrace{r \cos \theta}_x, \underbrace{r \sin \theta}_y)$.

When $r = 2$ and $\theta = \pi/4$, the Jacobian of f at $p = (r, \theta)$ equals

$$df_p = \begin{pmatrix} x_r(p) & x_\theta(p) \\ y_r(p) & y_\theta(p) \end{pmatrix} = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix} = \begin{pmatrix} \sqrt{2}/2 & -\sqrt{2}/2 \\ \sqrt{2}/2 & \sqrt{2}/2 \end{pmatrix}.$$

The columns of this matrix are $\mathbf{f}_r(p) = df_p(1, 0)$ and $\mathbf{f}_\theta(p) = df_p(0, 1)$, which are the initial velocity vectors of the images under f of the coordinate lines through p in the $r\theta$ -plane, as illustrated in Fig. 3.2.

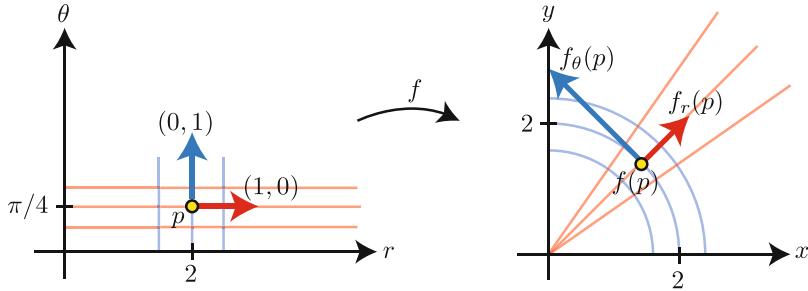


FIGURE 3.2. Polar coordinates

Assuming that f is C^2 on U , we now wish to improve Eq. 3.5 into a *second*-order Taylor approximation, but only in the case $m = 2$, since this is the only case required for the remainder of the book. If we denote the components of v in this case by $v = (a, b)$, the second-order Taylor approximation of f at p in the direction of v is

(3.6)

$$f(p+tv) - f(p) \approx t \underbrace{(af_x(p) + bf_y(p))}_{df_p(v)} + \frac{t^2}{2} (a^2 f_{xx}(p) + 2ab f_{xy}(p) + b^2 f_{yy}(p)).$$

If f is C^2 on U , it can be shown that the left and right sides of Eq. 3.6 differ in norm by an error term, $E(t)$, with $\lim_{t \rightarrow 0} \frac{E(t)}{t^2} = 0$.

For a differentiable function, the straight line γ in Definition 3.3 (illustrated in Fig. 3.1) can be replaced by *any* (possibly curved) path with the same initial position and initial velocity:

PROPOSITION 3.7.

If $f : U \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ is smooth, $v \in \mathbb{R}^m$, and γ is any regular curve in U with $\gamma(0) = p$ and $\gamma'(0) = v$, then

$$(3.7) \quad df_p(v) = (f \circ \gamma)'(0).$$

This proposition is illustrated in Fig. 3.3. Throughout the remainder of the book, it will repeatedly serve as a more flexible and more useful characterization of $df_p(v)$.

EXAMPLE 3.8. The case $\{m = 2, n = 3\}$ in Proposition 3.7 will be important throughout the book, typically with the following variable names. Let $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ be smooth, with component functions denoted by $\sigma(u, v) = (x(u, v), y(u, v), z(u, v))$. Let $\gamma : I \rightarrow \mathbb{R}^2$ be a plane curve in the uv -plane, with component functions denoted by $\gamma(t) = (u(t), v(t))$. In this case, Proposition 3.7 says that

$$(\sigma \circ \gamma)'(t) = d\sigma_{\gamma(t)}(\gamma'(t))$$

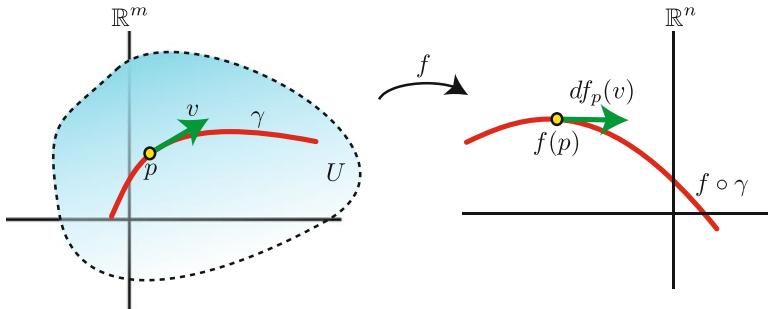


FIGURE 3.3. $df_p(v)$ is the initial velocity vector of the composition with f of *any* regular curve γ with $\gamma(0) = p$ and $\gamma'(0) = v$

for all $t \in I$ (since there is nothing special about $t = 0$). This simplifies as follows, with everything assumed to be evaluated at an arbitrary $t \in I$ or at the corresponding point $(u(t), v(t))$ as appropriate:

$$(\sigma \circ \gamma)' = \begin{pmatrix} \color{red}{x_u} & \color{blue}{x_v} \\ \color{red}{y_u} & \color{blue}{y_v} \\ \color{red}{z_u} & \color{blue}{z_v} \end{pmatrix} \begin{pmatrix} u' \\ v' \end{pmatrix} = u' \sigma_u + v' \sigma_v.$$

Proposition 3.7 is essentially the $l = 1$ special case of the following fundamental rule, which says that the derivative of a composition of two functions equals the composition of their derivatives:

PROPOSITION 3.9 (Chain Rule).

If $g : U \subset \mathbb{R}^l \rightarrow \mathbb{R}^m$ is smooth on U , $f : V \subset \mathbb{R}^m \rightarrow \mathbb{R}^n$ is smooth on V , and $g(U) \subset V$ (exactly as in Proposition 3.2), then for all $q \in U$,

$$d(f \circ g)_q = df_{g(q)} \circ dg_q,$$

where $p = g(q)$.

The chain rule is an important tool, and several comments about it are in order. First, notice that each side of the chain rule is a linear transformation from \mathbb{R}^l to \mathbb{R}^n . The rule says that they are the same linear transformation, which means they act the same on every vector; that is:

$$d(f \circ g)_q(w) = df_{g(q)}(v) \quad \text{for all } w \in \mathbb{R}^l, \text{ where } v = dg_q(w),$$

as illustrated in Fig. 3.4. This figure hints that the general chain rule essentially follows from the $l = 1$ special case of the chain rule stated in Proposition 3.7.

Second, the right side of the chain rule is a composition of two linear transformations. Such a composition is represented by the product of the matrices representing the two linear transformations. So you could really

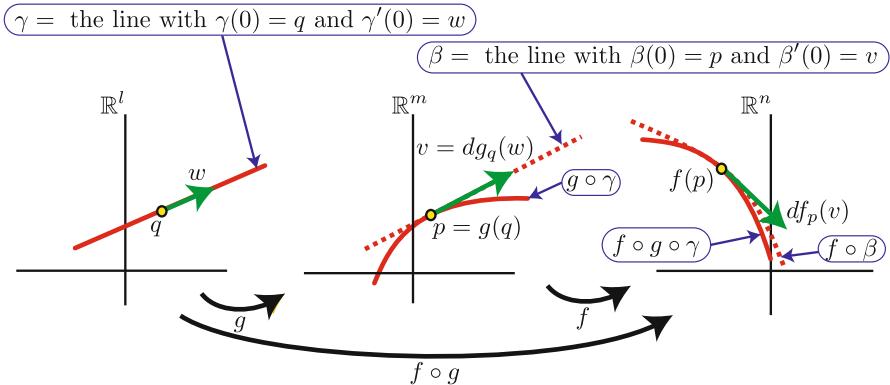


FIGURE 3.4. The chain rule says that $d(f \circ g)_q(w) = df_p(v)$; that is, the solid and the dashed red curves in \mathbb{R}^n have the same initial velocity vector

think of the right side of the chain rule as a matrix product (the product of two Jacobian matrices).

Third, the chain rule implies that *the derivative of an invertible function is an invertible matrix*. More precisely, suppose that $f : U \subset \mathbb{R}^n \rightarrow \mathbb{R}^n$ (same dimensions) is an invertible function from U to its image $f(U)$. Suppose that f is smooth on a neighborhood of $q \in U$ and f^{-1} is smooth on a neighborhood of $f(q)$. The chain rule says that

$$(3.8) \quad d(f^{-1} \circ f)_q = d(f^{-1})_{f(q)} \circ df_q.$$

On the other hand, $f^{-1} \circ f$ is the identity function, whose derivative at every point is the identity map. So $d(f^{-1})_{f(q)} \circ df_q$ is the identity linear transformation, which means that df_q is an invertible linear transformation (the corresponding matrix must be an invertible matrix).

A crucial result from analysis is the following converse:

THEOREM 3.10 (Inverse Function Theorem).

If $f : \mathbb{R}^n \rightarrow \mathbb{R}^n$ is smooth on a neighborhood of $p \in \mathbb{R}^n$ and df_p is an invertible linear transformation, then there exists a (possibly smaller) neighborhood U of p such that $V = f(U)$ is a neighborhood of $f(p)$, and $f : U \rightarrow V$ is invertible with smooth inverse.

The inverse function theorem is quite remarkable. It reduces the seemingly difficult problem of deciding whether f is locally invertible near p to the computationally simple task of checking whether the determinant of the linear transformation df_p is nonzero! The proof is nontrivial, but the theorem should be believable, since $f(q) \approx f(p) + df_q(q - p)$ is a first-order

approximation of f near p . The theorem says that if this first-order approximation is invertible (between neighborhoods), then f is invertible (between neighborhoods).

For example, the polar coordinate function $f(r, \theta) = (r \cos \theta, r \sin \theta)$ is guaranteed to be invertible with smooth inverse in a neighborhood of $p = (2, \pi/4)$, because the matrix for df_p (computed in Example 3.6) has nonzero determinant. In this example, it is straightforward to find an explicit formula for $f^{-1}(x, y)$ (converting from rectangular back to polar coordinates), but in many other examples it is not.

EXAMPLE 3.11. The function $f : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined as $f(x, y) = (\underbrace{x^2 + y}_{f_1}, \underbrace{xy}_{f_2})$

is smooth. At $p = (x, y) \in \mathbb{R}^2$, we have

$$df_p = \begin{pmatrix} \frac{\partial f_1}{\partial x}(p) & \frac{\partial f_1}{\partial y}(p) \\ \frac{\partial f_2}{\partial x}(p) & \frac{\partial f_2}{\partial y}(p) \end{pmatrix} = \begin{pmatrix} 2x & 1 \\ y & x \end{pmatrix}.$$

For example, $df_{(1,1)} = \begin{pmatrix} 2 & 1 \\ 1 & 1 \end{pmatrix}$ has nonzero determinant. The inverse function theorem implies that the restriction of f to a sufficiently small neighborhood of $(1, 1)$ is an invertible function with smooth inverse between this neighborhood and its image, which is a neighborhood of $f(1, 1) = (2, 1)$. To verify this explicitly without the inverse function theorem would involve the messy work of solving the system $\{a = x^2 + y, b = xy\}$ for x and y uniquely in terms of a and b near $x = 1, y = 1$.

Also notice that $df_{(0,0)} = \begin{pmatrix} 0 & 1 \\ 0 & 0 \end{pmatrix}$ is not an invertible matrix, so the discussion surrounding Eq. 3.8 implies that f is not an invertible function with smooth inverse between neighborhoods $(0, 0)$ and $f(0, 0)$. In particular, the above-mentioned neighborhood of $(1, 1)$ does not include $(0, 0)$.

We end this section with a characterization of smoothness for functions whose domains are not necessarily open. Suppose that $X \subset \mathbb{R}^m$ is a (not necessarily open) set and $f : X \rightarrow \mathbb{R}^n$ is a function. The partial derivatives of f at $p \in X$ might not make sense, because f is defined only on X , but you might immediately leave X if you move away from p in a coordinate direction of \mathbb{R}^m . To solve this problem, we will call f smooth if it locally extends to a smooth function on \mathbb{R}^m . More precisely:

DEFINITION 3.12.

If $X \subset \mathbb{R}^m$ is a (not necessarily open) set, then $f : X \rightarrow \mathbb{R}^n$ is called **smooth** if for all $p \in X$, there exist a neighborhood U of p in \mathbb{R}^m and a smooth function $\tilde{f} : U \rightarrow \mathbb{R}^n$ that agrees with f on $X \cap U$.

Notice that when $m = 1$, this definition agrees with the convention discussed in Exercise 1.2 for what smoothness of a curve $\gamma : [a, b] \rightarrow \mathbb{R}^n$ should mean at the endpoints a, b of the domain. This general extended notion of smoothness allows us to define an important type of equivalence for subsets of Euclidean space:

DEFINITION 3.13.

$X \subset \mathbb{R}^{m_1}$ and $Y \subset \mathbb{R}^{m_2}$ are called **diffeomorphic** if there exists a smooth bijective function $f : X \rightarrow Y$ whose inverse is also smooth. In this case, f is called a **diffeomorphism**.

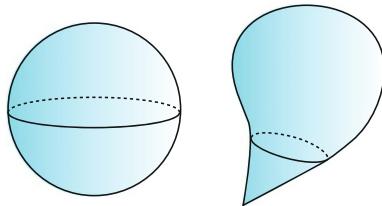


FIGURE 3.5. Homeomorphic but not diffeomorphic

A diffeomorphism is a homeomorphism (Definition A.15 on page 351 of the appendix) that is smooth and has a smooth inverse. Figure 3.5 shows two sets that are homeomorphic, but are not diffeomorphic because no homeomorphism between them could be smooth at the cone point.

EXAMPLE 3.14. *The sphere and the ellipsoid:*

$$S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\},$$

$$E = \{(x, y, z) \in \mathbb{R}^3 \mid (x/a)^2 + (y/b)^2 + (z/c)^2 = 1\},$$

are diffeomorphic for every choice of $a, b, c > 0$. A diffeomorphism $f : S^2 \rightarrow E$ is given by $f(x, y, z) = (ax, by, cz)$. It is smooth because it extends to the smooth function on \mathbb{R}^3 with the same formula: $(x, y, z) \mapsto (ax, by, cz)$. The inverse is given by $f^{-1}(x, y, z) = (x/a, y/b, z/c)$, which is smooth for the same reason.

EXAMPLE 3.15. *The plane \mathbb{R}^2 is diffeomorphic to the paraboloid*

$$P = \{(x, y, z) \in \mathbb{R}^3 \mid z = x^2 + y^2\}$$

via the diffeomorphism $\sigma : \mathbb{R}^2 \rightarrow P$ defined as $\sigma(x, y) = (x, y, x^2 + y^2)$. The inverse $\sigma^{-1} : P \rightarrow \mathbb{R}^2$ is given by $\sigma^{-1}(x, y, z) = (x, y)$. It is smooth because it extends to the smooth function from \mathbb{R}^3 to \mathbb{R}^2 with the same formula: $(x, y, z) \mapsto (x, y)$.

EXAMPLE 3.16. *The upper hemisphere and the unit disk:*

$$S_+^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1 \text{ and } z > 0\},$$

$$D^2 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\},$$

are diffeomorphic via the diffeomorphism $\sigma : D^2 \rightarrow S_+^2$ defined as

$$\sigma(x, y) = \left(x, y, \sqrt{1 - x^2 - y^2} \right).$$

The inverse $\sigma^{-1} : S_+^2 \rightarrow D$ is given by $\sigma^{-1}(x, y, z) = (x, y)$. It is smooth because it extends to the smooth function from \mathbb{R}^3 to \mathbb{R}^2 with the same formula: $(x, y, z) \mapsto (x, y)$.

The last two examples seem repetitive because they are both special cases of the following general rule, whose proof should now be clear:

LEMMA 3.17.

If $U \subset \mathbb{R}^2$ is open and $f: U \rightarrow \mathbb{R}$ is smooth, then its graph $G = \{(x, y, f(x, y)) \mid (x, y) \in U\}$ is diffeomorphic to U via the diffeomorphism $\sigma: U \rightarrow G$ defined as $\sigma(x, y) = (x, y, f(x, y))$.

Thus, the graph of a smooth function of two variables “looks like” (is diffeomorphic to) its domain, which is an open set in \mathbb{R}^2 . Inhabitants of the graph might believe that they live on the plane, at least if they are sufficiently small and nearsighted. We’ll see in the next section that this is exactly what qualifies such a graph to be called a *regular surface*.

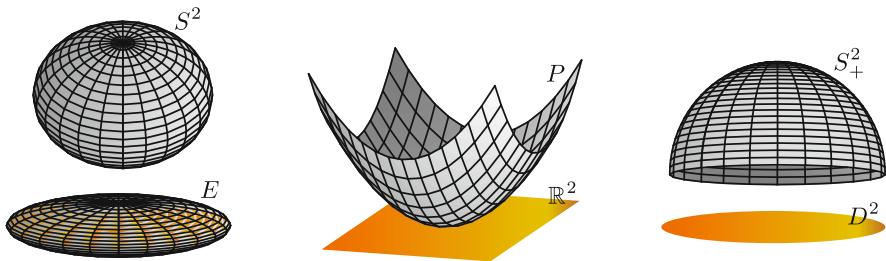


FIGURE 3.6. Diffeomorphic pairs from Examples 3.14, 3.15, and 3.16

EXERCISES

EXERCISE 3.1 (The Derivative of a Linear Transformation Is Itself). Let $f: \mathbb{R}^m \rightarrow \mathbb{R}^n$ be a linear transformation. For every $p \in \mathbb{R}^m$, prove that $df_p = f$.

EXERCISE 3.2. If $A \in O(3)$ is an orthogonal matrix, $v \in \mathbb{R}^3$, and f is the rigid motion $f = T_v \circ L_A: \mathbb{R}^3 \rightarrow \mathbb{R}^3$ (as in Proposition 1.58 on page 52), prove for all $p \in \mathbb{R}^3$ that $df_p = L_A$.

EXERCISE 3.3. Define $f: \mathbb{R}^2 \rightarrow \mathbb{R}^3$ as $f(x, y) = (5x^2y^3, 2x + y^2, x^2 - y^2)$. Compute the Jacobian matrix for f at $p_0 = (1, -1)$. What is the rank of df_{p_0} ?

EXERCISE 3.4. Is the function $f: \mathbb{R}^2 \rightarrow \mathbb{R}^2$ defined as $f(x, y) = (x^3, y)$ a diffeomorphism?

EXERCISE 3.5. Let $U \subset \mathbb{R}^2$ be open and $f: U \rightarrow \mathbb{R}$ smooth. Denote the coordinates of \mathbb{R}^2 by $\{u, v\}$. Assume that the point $q \in U$ is a **critical point** of f , which means that $df_q(w) = 0$ for all $w \in \mathbb{R}^2$, or equivalently, $f_u(q) = f_v(q) = 0$. Let $\gamma: I \rightarrow U$ be a regular curve with $\gamma(0) = q$ and $\gamma'(0) = w = (a, b)$. Prove that

$$(f \circ \gamma)''(0) = a^2 f_{uu}(q) + 2ab f_{uv}(q) + b^2 f_{vv}(q).$$

In particular, the value $(f \circ \gamma)''(0)$ does not depend on the second or higher derivative of γ at 0. In other words, the expression $\text{Hess}(f)_q(w) = (f \circ \gamma)''(0)$ is well defined (independent of the choice of γ with $\gamma(0) = q$ and $\gamma'(0) = w$). This expression is called the **Hessian** of f at q in the direction w .

EXERCISE 3.6. In Example 3.6, describe a largest-possible neighborhood of $p = (2, \pi/4)$ such that f restricts to a diffeomorphism between this neighborhood and its image.

EXERCISE 3.7. Let E be the ellipsoid from Example 3.14. Characterize all of the linear transformations from \mathbb{R}^3 to \mathbb{R}^3 that restrict to a diffeomorphism from E to E , assuming that a, b, c are all distinct. What if two of them are equal? What if all of them are equal?

EXERCISE 3.8. Let $\gamma : [0, 2\pi] \rightarrow \mathbb{R}^n$ be a closed curve. Recall that

$$S^1 = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 = 1\}.$$

Prove that there exists a *smooth* function $f : S^1 \rightarrow \mathbb{R}^n$ such that $\gamma(t) = f(\cos t, \sin t)$ for all $t \in [0, 2\pi]$ (*smooth* as in Definition 3.12). Modify this conclusion to apply more generally when the domain of γ is $[a, b]$.

EXERCISE 3.9. Prove that a composition of two diffeomorphisms is a diffeomorphism.

EXERCISE 3.10. Are the cone $\{(x, y, z) \in \mathbb{R}^3 \mid z^2 = x^2 + y^2, z > 0\}$ and the cylinder $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ diffeomorphic?

EXERCISE 3.11 (Tubular Neighborhood of a Plane Curve). In this exercise, you will prove Proposition 2.5, illustrated in Fig. 2.3 on page 65. Let $\gamma : [a, b] \rightarrow \mathbb{R}^2$ be a unit-speed simple closed plane curve. For $t \in [a, b]$, set $N(t) = R_{90}(\gamma'(t))$. Define $\varphi : (-\epsilon, \epsilon) \times [a, b] \rightarrow \mathbb{R}^2$ as $\varphi(s, t) = \gamma(t) + s \cdot N(t)$. Prove that there exists $\epsilon > 0$ such that:

- (1) φ is a diffeomorphism from $(-\epsilon, \epsilon) \times (a, b)$ onto its image.
- (2) φ is injective on $(-\epsilon, \epsilon) \times [a, b]$.

HINT: Extend γ to a periodic function on \mathbb{R} . For $t_0 \in [a, b]$, use the inverse function theorem to find $\epsilon_0 > 0$ such that φ is a diffeomorphism on $(-\epsilon_0, \epsilon_0) \times U_0$, where U_0 is a neighborhood of t_0 in \mathbb{R} . Since $[a, b]$ is compact, there are finitely many such times $t_1, \dots, t_n \in [a, b]$ (with corresponding values $\epsilon_1, \dots, \epsilon_n$ and corresponding neighborhoods U_1, \dots, U_n) such that $\{U_1, \dots, U_n\}$ is an open cover of $[a, b]$. Let δ denote the Lebesgue number of this open cover, as defined in Proposition A.23 on page 355 of the appendix. Define $\epsilon = \min\{\epsilon_1, \dots, \epsilon_n, \frac{\delta}{2}\}$.

EXERCISE 3.12 (Tubular Neighborhood of a Space Curve). Let $\gamma : [a, b] \rightarrow \mathbb{R}^3$ be a unit-speed simple closed space curve with nowhere-vanishing curvature. For $t \in [a, b]$, let $\{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}$ denote the Frenet frame vectors at $\gamma(t)$. Set $B_\epsilon = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < \epsilon\}$, and define

$$\begin{aligned} \varphi : [a, b] \times B_\epsilon &\rightarrow \mathbb{R}^3 & \text{as} & \quad \varphi(t, x, y) = \gamma(t) + x \cdot \mathbf{n}(t) + y \cdot \mathbf{b}(t), \\ \phi : [a, b] \times [0, 2\pi] &\rightarrow \mathbb{R}^3 & \text{as} & \quad \phi(t, \theta) = \gamma(t) + (\epsilon \cdot \cos \theta) \cdot \mathbf{n}(t) + (\epsilon \cdot \sin \theta) \cdot \mathbf{b}(t). \end{aligned}$$

Let S_ϵ denote the image of ϕ (equivalently, the boundary of the image of φ). Prove there exists $\epsilon > 0$ such that:

- (1) φ is a diffeomorphism from $(a, b) \times B_\epsilon$ onto its image, and is injective on $[a, b] \times B_\epsilon$.
- (2) ϕ is a diffeomorphism from $(a, b) \times (0, 2\pi)$ onto its image, and is injective on $[a, b] \times [0, 2\pi]$.
- (3) $S_\epsilon = \{p \in \mathbb{R}^3 \mid \text{dist}(p, C) = \epsilon\}$, where C is the trace of γ and “ $\text{dist}(p, C)$ ” denotes the distance from p to the point of C closest to p .

Use a computer graphics application to plot S_ϵ for the space curves that you plotted in Exercise 1.13 on page 8 (for suitable choices of ϵ).

2. Regular Surfaces

The remainder of this book is devoted to exploring the geometry of a *regular surface*, which roughly means a subset in \mathbb{R}^3 that “locally looks like” the plane \mathbb{R}^2 . Sufficiently small and nearsighted inhabitants of a regular surface would think they live on the plane, just as early humans believed the Earth to be flat.

Our definition of a regular surface will be based on the following *relative* notion of “open in” and “neighborhood in,” which is found in the book’s appendix:

DEFINITION 3.18.

*Let $S \subset \mathbb{R}^n$ be a subset. A set $V \subset S$ is called **open in S** if V is the intersection with S of an open set in \mathbb{R}^n . If $p \in S$, then a **neighborhood of p in S** means a subset of S that is open in S and that contains p .*

This is a good time to read the book’s appendix. It contains useful ways to think about and work with the above definition.

DEFINITION 3.19.

*A set $S \subset \mathbb{R}^3$ is called a **regular surface** if each of its points has a neighborhood in S that is diffeomorphic to an open set in \mathbb{R}^2 . That is, for every $p \in S$, there exist a neighborhood, V , of p in S , an open set $U \subset \mathbb{R}^2$, and a diffeomorphism $\sigma : U \rightarrow V$. Such a diffeomorphism σ is called a **surface patch** or a **coordinate chart**. A collection of surface patches that together cover all of the points of S is called an **atlas** for S .*

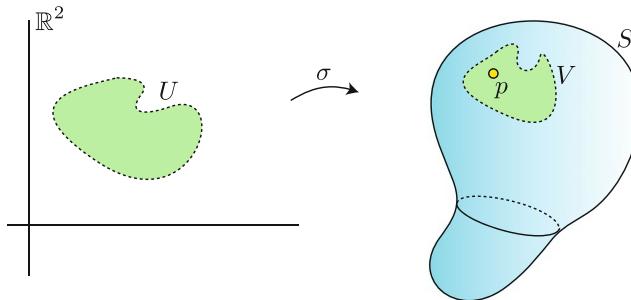


FIGURE 3.7. A surface patch $\sigma : U \rightarrow V$ that covers the point $p \in S$

This definition is illustrated in Fig. 3.7. The definition of a **surface** is obtained by replacing “diffeomorphism” with “homeomorphism” in Definition 3.19. Thus, the adjective “regular” means that the surface patches can be chosen to be smooth and have smooth inverses, so that they are diffeomorphisms rather than only homeomorphisms.

EXAMPLE 3.20. The cone $S = \{(x, y, z) \in \mathbb{R}^3 \mid z = \sqrt{x^2 + y^2}\}$ is a surface that can be covered by the single surface patch $\sigma : \mathbb{R}^2 \rightarrow S$ defined as $\sigma(u, v) = (u, v, \sqrt{u^2 + v^2})$; see Fig. 3.8. Notice that σ is a homeomorphism but not a diffeomorphism, because it is not differentiable at the origin. You will verify in Exercise 3.49 that S is not a regular surface.

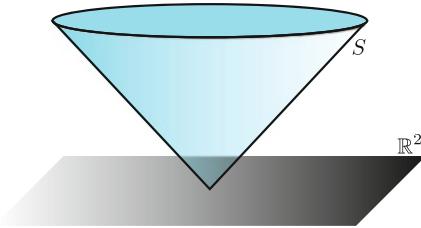


FIGURE 3.8. A surface that is not regular

Differential geometry involves doing calculus on surfaces, so we'll need our surfaces to be regular. Here is a useful consequence of regularity:

PROPOSITION 3.21.

If S is a regular surface, and $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ is a surface patch, then for all $q \in U$, the linear transformation $d\sigma_q : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ has rank 2.

We will henceforth restrict our study to *regular* surfaces. In the previous example, inhabitants of the cone living at the origin, no matter how small and near-sighted, could detect that their world differs from the plane, and if they knew a bit of calculus, they could even explain the difference. Dif-

PROOF. Set $p = \sigma(q)$. Consider σ^{-1} to be extended to a smooth function on a neighborhood of p in \mathbb{R}^3 . We'll abuse notation by also denoting this extension by σ^{-1} . The chain rule says that

$$d(\sigma^{-1} \circ \sigma)_q = d(\sigma^{-1})_p \circ d\sigma_q.$$

On the other hand, $\sigma^{-1} \circ \sigma$ is the identity function, whose derivative at every point is the identity map. So $d(\sigma^{-1})_p \circ d\sigma_q$ is the identity linear transformation on \mathbb{R}^2 , which has rank 2. Since the rank of a composition of two linear transformations is less than or equal to the rank of the first, we learn that $d\sigma_q$ has rank 2. \square

The conclusion that $d\sigma_q$ has rank 2 means that its image is two-dimensional. This image is spanned by $d\sigma_q(e_1)$ and $d\sigma_q(e_2)$ (where $e_1 = (1, 0)$ and $e_2 = (0, 1)$), so the rank-2 conclusion means that $d\sigma_q(e_1)$ and $d\sigma_q(e_2)$ are linearly independent. To be more explicit, let's name the coordinate variables of \mathbb{R}^2 as $\{u, v\}$, so that our surface patch $\sigma : U \rightarrow V \subset S$ is given in components as

$$\sigma(u, v) = (x(u, v), y(u, v), z(u, v)).$$

Its Jacobian matrix at $q = (u_0, v_0) \in U$ is

$$d\sigma_q = \begin{pmatrix} \frac{\partial x}{\partial u}(q) & \frac{\partial x}{\partial v}(q) \\ \frac{\partial y}{\partial u}(q) & \frac{\partial y}{\partial v}(q) \\ \frac{\partial z}{\partial u}(q) & \frac{\partial z}{\partial v}(q) \end{pmatrix}.$$

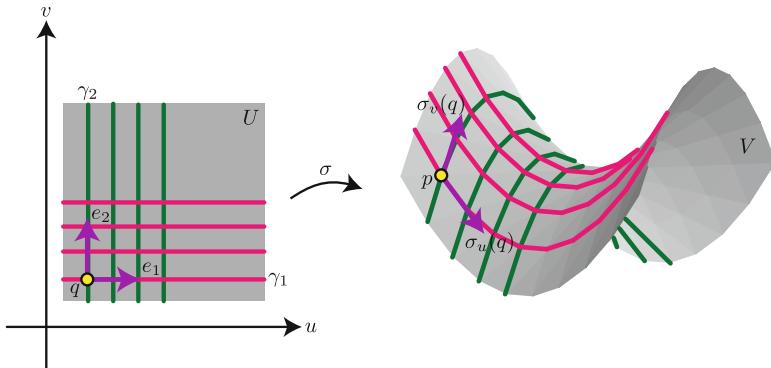
The two columns of this Jacobian matrix are the (componentwise) partial derivatives of σ with respect to u and v at q ; as in the previous section, these are denoted by $\sigma_u(q)$ and $\sigma_v(q)$. That is,

$$\sigma_u(q) = \left(\frac{\partial x}{\partial u}(q), \frac{\partial y}{\partial u}(q), \frac{\partial z}{\partial u}(q) \right) \quad \text{and} \quad \sigma_v(q) = \left(\frac{\partial x}{\partial v}(q), \frac{\partial y}{\partial v}(q), \frac{\partial z}{\partial v}(q) \right).$$

Recall from the previous section that $\sigma_u(q) = d\sigma_q(e_1)$ and $\sigma_v(q) = d\sigma_q(e_2)$ are the initial tangent vectors to the images under σ of the coordinate lines in \mathbb{R}^2 through q , as shown in Fig. 3.9. In summary,

$$\begin{aligned} \sigma_u(q) &= d\sigma_q(e_1) = (\sigma \circ \gamma_1)'(0) \text{ where } \gamma_1(t) = q + te_1 = (u_0 + t, v_0), \\ \sigma_v(q) &= d\sigma_q(e_2) = (\sigma \circ \gamma_2)'(0) \text{ where } \gamma_2(t) = q + te_2 = (u_0, v_0 + t). \end{aligned}$$

The variables u and v are called **local coordinates** on V . Inhabitant of V could uniquely identify their location from their u and v values (just as locations on Earth are identified by their latitude and longitude values). An inhabitant living at $\sigma(q)$ could move in the direction of $\sigma_u(q)$ along the pink coordinate curve or in the direction of $\sigma_v(q)$ along the green coordinate curve. The rank-2 conclusion ensures that these two possible directions are really independent. In other words, it ensures that the local coordinate system does not degenerate anywhere on V (the way the latitude/longitude coordinate system for the earth degenerates at the North and South Poles).

FIGURE 3.9. $\sigma_u(q) = d\sigma_q(e_1)$ and $\sigma_v(q) = d\sigma_q(e_2)$

EXAMPLE 3.22 (The Plane). Every open set $U \subset \mathbb{R}^2$ is a regular surface covered by a single surface patch, namely, the identity function from U to U . Since technically only subsets of \mathbb{R}^3 can be called surfaces, we are implicitly regarding \mathbb{R}^2 as the xy -plane of \mathbb{R}^3 .

In each of the remaining examples, we will prove that the given set is a surface by constructing explicit surface patches. We will then independently verify the rank-2 conclusion for these surface patches (even though it is guaranteed by Proposition 3.21). This unnecessary work is worthwhile, because as we will later see, the rank-2 verification has independent uses, and it could replace some of the steps of verifying that the purported surface patch is really a diffeomorphism.

EXAMPLE 3.23 (The Cylinder). We will show that the cylinder

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$$

is a regular surface. Every point $p = (x, y, z) \in C$ with $x \neq -1$ lies in the image of the smooth function $\sigma : \underbrace{(-\pi, \pi)}_U \times \mathbb{R} \rightarrow C$ defined as $\sigma(u, v) =$

$$\left(\underbrace{\cos(u)}_{x(u,v)}, \underbrace{\sin(u)}_{y(u,v)}, \underbrace{v}_{z(u,v)} \right); \text{ see Fig. 3.10.}$$

The image $V = \sigma(U)$ is open in C , because it is the intersection with C of the open set $\{(x, y, z) \in \mathbb{R}^3 \mid x \neq -1\}$. It is straightforward to verify that σ is injective. To complete the verification that σ is a diffeomorphism (and is thus a valid surface patch), it remains to show that the inverse $\sigma^{-1} : V \rightarrow U$ is smooth. At points of $V_+ = \{(x, y, z) \in V \mid x > 0\}$, the map σ^{-1} is given by the formula $(x, y, z) \mapsto (\tan^{-1}(y/x), z)$, which is smooth because it extends to the smooth function with the same formula on the open set $\{(x, y, z) \in \mathbb{R}^3 \mid x > 0\}$. The smoothness of σ^{-1} at all other points of V can be verified in a similar manner. Thus, σ is a surface patch.

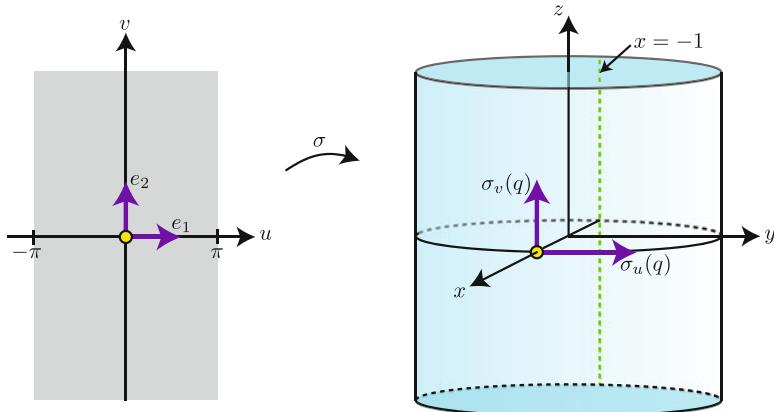


FIGURE 3.10. The surface patch σ covers all but the green line

The points of C at which $x = -1$ can be covered by the second function $\mu : (-\pi, \pi) \times \mathbb{R} \rightarrow C$ defined as $\mu(u, v) = (\cos(u + \pi), \sin(u + \pi), v)$, which is a valid surface patch by similar arguments. Notice that μ is the composition of σ with the rigid motion of \mathbb{R}^3 that rotates 180° about the z -axis.

This completes the verification that C is a regular surface. For later reference, we will now independently verify the rank-2 condition guaranteed by Proposition 3.21 for the surface patch σ . The Jacobian matrix at $q = (u, v) \in U$ is

$$d\sigma_q = \begin{pmatrix} \frac{\partial x}{\partial u}(q) & \frac{\partial x}{\partial v}(q) \\ \frac{\partial y}{\partial u}(q) & \frac{\partial y}{\partial v}(q) \\ \frac{\partial z}{\partial u}(q) & \frac{\partial z}{\partial v}(q) \end{pmatrix} = \begin{pmatrix} -\sin(u) & 0 \\ \cos(u) & 0 \\ 0 & 1 \end{pmatrix}.$$

It has rank 2 for all choices of $q \in U$, because the two columns

$$\sigma_u(q) = (-\sin(u), \cos(u), 0), \quad \sigma_v(q) = (0, 0, 1),$$

are linearly independent. Notice that $\sigma_u(q)$ points “around,” while $\sigma_v(q)$ points “up.”

We saw in Example 3.16 on page 122 that the upper hemisphere ($z > 0$) of S^2 is diffeomorphic to an open disk in \mathbb{R}^2 . The other five hemispheres ($z < 0$, $x > 0$, $x < 0$, $y > 0$, and $y < 0$) are also diffeomorphic to disks, by similar arguments. For example, the hemisphere $y < 0$ is covered by the surface patch $\sigma(x, z) = (x, -\sqrt{1 - x^2 - z^2}, z)$, whose domain is an open disk in the xz -plane, $\{(x, z) \in \mathbb{R}^2 \mid x^2 + z^2 < 1\}$. Thus, S^2 is a regular surface covered by an atlas of six surface patches. But we can do better—in the next example, we will cover S^2 with an atlas of two surface patches.

EXAMPLE 3.24 (The Sphere). Recall that a point of S^2 can be identified by its **spherical coordinates** (θ, ϕ) , illustrated in Fig. 3.11 (left). The conver-

sion from spherical to rectangular coordinates is encoded in the surface patch $\sigma : \underbrace{(0, 2\pi)}_U \times (0, \pi) \rightarrow S^2$, defined as

 U

$$\sigma(\theta, \phi) = (\underbrace{\sin \phi \cos \theta}_{x(\theta, \phi)}, \underbrace{\sin \phi \sin \theta}_{y(\theta, \phi)}, \underbrace{\cos \phi}_{z(\theta, \phi)}).$$

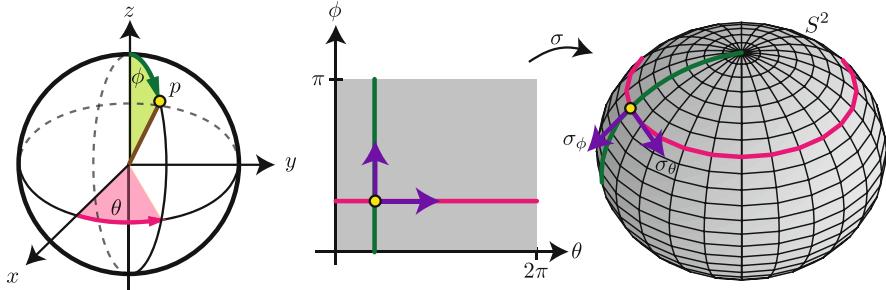


FIGURE 3.11. Spherical coordinates determine a surface patch

The image $V = \sigma(U)$ is open in S^2 because it is the intersection with S^2 of the following open set: $\mathbb{R}^3 - \{(x, 0, z) \in \mathbb{R}^3 \mid x \geq 0\}$. To show that $\sigma : U \rightarrow V$ is a diffeomorphism (and is thus a valid surface patch), it remains to show that σ is one-to-one, and that $\sigma^{-1} : V \rightarrow U$ is smooth. This requires explicit formulas for translating from rectangular back to spherical coordinates. We leave the details to the reader.

This surface patch covers all points of S^2 except the north and south poles and the arc between them along which $\theta = 0$. Let's call this excluded set C ; in Fig. 3.12 it is the red arc of a great circle connecting the poles. Choose an orthogonal matrix A such that $L_A(C)$ is disjoint from C . For example, if $A = \begin{pmatrix} -1 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$, then $L_A(C)$ will be the purple arc in Fig. 3.12. The composition $L_A \circ \sigma$ is a second surface patch that covers all of S^2 except $L_A(C)$, so together the two patches cover all of S^2 .

For later reference, we will now independently verify the rank-2 condition guaranteed by Proposition 3.21 for the surface patch σ . The Jacobian at a point $q = (\theta, \phi) \in U$ is

$$d\sigma_q = \begin{pmatrix} \frac{\partial x}{\partial \theta}(q) & \frac{\partial x}{\partial \phi}(q) \\ \frac{\partial y}{\partial \theta}(q) & \frac{\partial y}{\partial \phi}(q) \\ \frac{\partial z}{\partial \theta}(q) & \frac{\partial z}{\partial \phi}(q) \end{pmatrix} = \begin{pmatrix} -\sin \phi \sin \theta & \cos \phi \cos \theta \\ \sin \phi \cos \theta & \cos \phi \sin \theta \\ 0 & -\sin \phi \end{pmatrix}.$$

The columns, σ_θ and σ_ϕ , are linearly independent because their cross product is nonzero. In fact, it is straightforward to compute

$$|\sigma_\theta \times \sigma_\phi| = |(-\sin^2 \phi \cos \theta, -\sin^2 \phi \sin \theta, -\sin \phi \cos \phi)| = \sin \phi \neq 0.$$

Thus, $d\sigma_q$ has rank 2 for all $q \in U$.

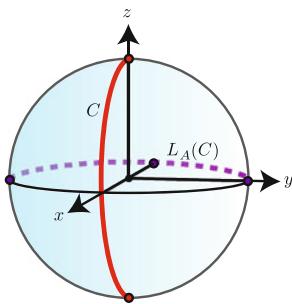


FIGURE 3.12. The arcs C and $L_A(C)$ do not intersect

Our next example is not a surface but a class of surfaces. This class is narrow enough to allow for explicit calculations, but wide enough to exemplify a number of important concepts for general surfaces.

EXAMPLE 3.25 (Surfaces of Revolution).

Let $\gamma(t) = (x(t), 0, z(t))$, $t \in (a, b)$, be a regular curve in the xz -plane. Let C denote its trace. For simplicity, we make the following two assumptions (which can later be weakened):

- (1) $x(t) > 0$ for all $t \in (a, b)$ (in particular, C does not intersect the z -axis).
- (2) $z'(t) > 0$ for all $t \in (a, b)$ (in particular, C has no self-intersections).

Let S denote the surface of revolution resulting when C is revolved about the z -axis; see Fig. 3.13. We claim that S is a regular surface.

A natural surface patch $\sigma : \underbrace{(-\pi, \pi) \times (a, b)}_U \rightarrow S$ is

$$\boxed{\sigma(\theta, t) = (x(t) \cos \theta, x(t) \sin \theta, z(t)).}$$

Notice that $\sigma(\theta, t) = R_\theta(\gamma(t))$, where R_θ is the linear rigid motion that rotates an angle θ about the z -axis, which is represented by the following orthogonal matrix:

$$R_\theta = \begin{pmatrix} \cos \theta & -\sin \theta & 0 \\ \sin \theta & \cos \theta & 0 \\ 0 & 0 & 1 \end{pmatrix}.$$

The image $V = \sigma(U)$ is open in S because it is the intersection with S of the following open set: $\mathbb{R}^3 - \{(x, 0, z) \in \mathbb{R}^3 \mid x \leq 0\}$. It is straightforward to verify that σ is injective. To complete the verification that σ is a diffeomorphism (and is thus a valid surface patch), it remains to show that the inverse $\sigma^{-1} : V \rightarrow U$ is smooth. At points of $V_+ = \{(x, y, z) \in V \mid x > 0\}$, the map σ^{-1} is given by the formula $(x, y, z) \mapsto (\tan^{-1}(y/x), t(z))$, where $t(z)$ is the inverse of the function $z(t)$. This is smooth because it extends to the smooth function with the same formula on the open set $\{(x, y, z) \in \mathbb{R}^3 \mid x > 0, z \in (z(a), z(b))\}$. The smoothness of σ^{-1} at all other points of V can be verified with similar arguments. Thus, σ is a surface patch.

This surface patch covers all of S except the curve along which it intersects that portion of the xz -plane where $x < 0$. As in the cylinder example, this omitted curve is covered by a second surface patch obtained by composing σ with a 180° rotation about the z -axis.

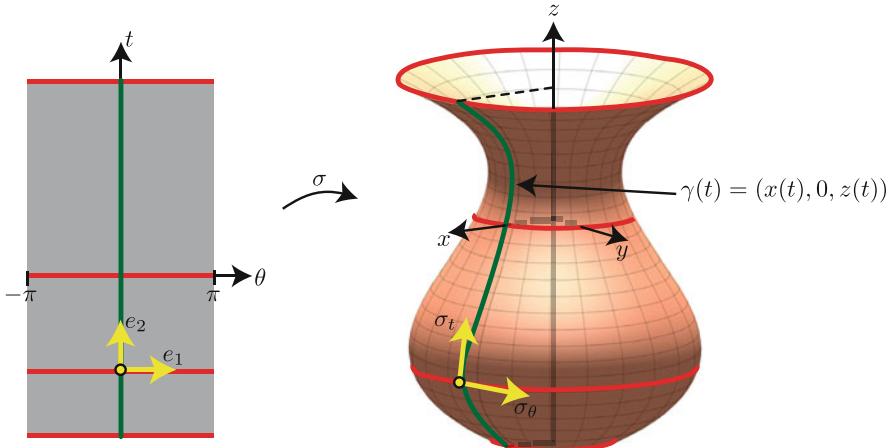


FIGURE 3.13. A surface of revolution

This completes the verification that S is a regular surface. For later reference, we will now independently verify the rank-2 condition guaranteed by Proposition 3.21 for the surface patch σ . The Jacobian matrix at $q = (\theta, t)$ is

$$d\sigma_q = \begin{pmatrix} -x(t) \sin \theta & x'(t) \cos \theta \\ x(t) \cos \theta & x'(t) \sin \theta \\ 0 & z'(t) \end{pmatrix}.$$

The columns of this matrix,

$$\sigma_\theta(q) = (-x(t) \sin \theta, x(t) \cos \theta, 0), \quad \sigma_t(q) = (x'(t) \cos \theta, x'(t) \sin \theta, z'(t)),$$

are linearly independent, because their cross product is nonzero:

$$(3.9) \quad |\sigma_\theta \times \sigma_t| = |(x(t)z'(t) \cos \theta, x(t)z'(t) \sin \theta, -x(t)x'(t))| = x(t)|\gamma'(t)| \neq 0.$$

The θ -parameter curves on S (shown in red) are called **latitudes**, while the t -parameter curves (shown in green) are called **longitudes**. Notice that σ_θ is tangent to the latitudes, while σ_t is tangent to the longitudes.

The sphere and the cylinder are both examples of surfaces of revolution. In fact, the surface patches previously constructed for the sphere and the cylinder essentially come from the above general construction. For the cylinder, choose $\{x(t) = 1, z(t) = t\}$, $t \in \mathbb{R}$. For the sphere (minus the north and south poles), choose $\{x(t) = \sin t, z(t) = \cos t\}$, $t \in (0, \pi)$.

We next present a theorem that makes it easy to verify that certain subsets of \mathbb{R}^3 are regular surfaces. We first require some vocabulary:

DEFINITION 3.26.

Let $U \subset \mathbb{R}^3$ be open, $f : U \rightarrow \mathbb{R}$ be smooth, and $\lambda \in \mathbb{R}$ be in the image of f .

- (1) The set $f^{-1}(\lambda) = \{p \in U \mid f(p) = \lambda\}$ is called the **preimage** of λ .
- (2) λ is called a **regular value** of f if for all $p \in f^{-1}(\lambda)$, the Jacobian matrix for f at p is not the zero matrix, that is, if

$$df_p = \left(\frac{\partial f}{\partial x}(p), \frac{\partial f}{\partial y}(p), \frac{\partial f}{\partial z}(p) \right) \neq (0, 0, 0).$$

As mentioned previously, the Jacobian matrix here is just the gradient of f , so the condition says that f has a nonzero gradient vector at every point of the preimage $f^{-1}(\lambda)$.

THEOREM 3.27.

Under the hypotheses of Definition 3.26, if λ is a regular value of f , then the preimage $f^{-1}(\lambda)$ is a regular surface.

PROOF. Assume that λ is a regular value of f , and denote the preimage by $S = f^{-1}(\lambda)$. Let $p \in S$. By hypothesis, one of the three partial derivative of f at p is nonzero. Assume without loss of generality that $\frac{\partial f}{\partial z}(p) \neq 0$. Define the function $\psi : U \rightarrow \mathbb{R}^3$ as $\psi(x, y, z) = (x, y, f(x, y, z))$. If you imagine that f represents temperature, then ψ replaces the z -coordinate of every point with the temperature at that point. The Jacobian matrix for ψ at p is

$$d\psi_p = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ \frac{\partial f}{\partial x}(p) & \frac{\partial f}{\partial y}(p) & \frac{\partial f}{\partial z}(p) \end{pmatrix}.$$

Its determinant is $\det(d\psi_p) = \frac{\partial f}{\partial z}(p) \neq 0$, so it is an invertible matrix. The inverse function theorem (page 120) now guarantees the existence of a neighborhood \tilde{U} of p and a neighborhood W of $\psi(p)$ such that the restriction $\psi : \tilde{U} \rightarrow W$ is invertible with smooth inverse $\psi^{-1} : W \rightarrow \tilde{U}$.

Only the third component of ψ^{-1} is interesting. If we denote this component by ψ_z^{-1} , and we denote the “temperature” output of f by T , we have

$$\psi^{-1}(x, y, T) = (x, y, \psi_z^{-1}(x, y, T))$$

for all $(x, y, T) \in W$. That is, $\psi_z^{-1}(x, y, T)$ is the unique value of z such that $(x, y, z) \in \tilde{U}$ and $f(x, y, z) = T$. Returning to the temperature metaphor, $\psi_z^{-1}(x, y, T)$ is the unique height of a point in \tilde{U} above (x, y) where the temperature is T .

Define $\mathcal{U} = \{(x, y) \mid (x, y, z) \in \tilde{U}\}$, which is the projection of \tilde{U} onto the xy -plane. Notice that \mathcal{U} is an open set in the xy -plane. Define $h : \mathcal{U} \rightarrow \mathbb{R}$

as $h(x, y) = \psi_z^{-1}(x, y, \lambda) =$ the height of the unique point in \tilde{U} above (x, y) where the temperature equals λ . Notice that

$$S \cap \tilde{U} = \{(x, y, h(x, y)) \mid (x, y) \in \mathcal{U}\}.$$

This says that $S \cap \tilde{U}$ is the graph of h , so it is diffeomorphic to the domain, \mathcal{U} , of h by Lemma 3.17 (on page 123). In summary, $S \cap \tilde{U}$ is a neighborhood of p in S that is diffeomorphic to the open set \mathcal{U} in \mathbb{R}^2 . \square

The preimage $f^{-1}(\lambda)$ is called a **level set** of f . This title deserves to be promoted to **level surface** when λ is a regular value.

This theorem makes quick work of proving that certain sets are regular surfaces. For example, consider the quadratic function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ defined as

$$f(x, y, z) = x^2 + y^2 - z^2.$$

The Jacobian matrix of f at $p = (x, y, z)$ is $df_p = (2x, 2y, -2z)$, which vanishes only at the origin. The origin belongs to the cone-shaped level set $f = 0$, which is not a regular surface. The theorem guarantees that every other level set of this function is a regular surface. The level surface $f = 1$ is a **hyperboloid of one sheet**, while the level surface $f = -1$ is a **hyperboloid of two sheets**; see Fig. 3.14.

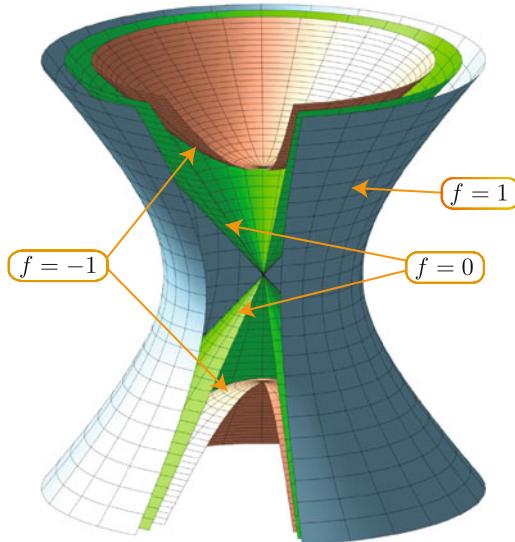


FIGURE 3.14. Some level sets of $f(x, y, z) = x^2 + y^2 - z^2$, shown with a cutaway sector for clarity

We end this section by mentioning that some of the literature about surfaces is devoted not to regular surfaces but to parametrized surfaces, defined as follows:

DEFINITION 3.28.

A **parametrized surface** is a smooth function $\sigma : U \rightarrow \mathbb{R}^3$ (where U is an open set in \mathbb{R}^2) such for all $q \in U$, $d\sigma_q$ has rank 2.

Notice that a regular surface is a set, while a parametrized surface is a function. Every surface patch for a regular surface is a parametrized surface. A parametrized surface need not be one-to-one. For example, the function $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as $\sigma(u, v) = (\cos u, \sin u, v)$ is a parametrized surface that wraps the plane infinitely many times around the cylinder; the restriction of this function to a smaller domain was used in Example 3.23 as a surface patch for the cylinder. Exercise 3.27 in this section (illustrated in Fig. 3.17) shows a parametrized surface that fails to be one-to-one because of more complicated kinds of self-intersections. In studying only local properties, it doesn't matter whether one works with regular surfaces or parametrized surfaces, because of the following proposition:

PROPOSITION 3.29.

If $\sigma : U \rightarrow \mathbb{R}^3$ is a parametrized surface, then for each $q_0 \in U$, there exists an open set $\tilde{U} \subset U$ containing q_0 such that the image $S = \sigma(\tilde{U})$ is a regular surface covered by a single surface patch (namely, the restriction of σ to \tilde{U}).

PROOF. Let $q_0 \in U$. Choose any $N \in \mathbb{R}^3$ with $N \notin \text{span}\{\sigma_u(q_0), \sigma_v(q_0)\}$. Consider the smooth function $f : U \times \mathbb{R} \rightarrow \mathbb{R}^3$ defined as follows:

$$f(q, t) = \sigma(q) + tN.$$

Set $p_0 = \sigma(q_0) = f(q_0, 0)$. The derivative $df_{(q_0, 0)} : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ is invertible, because it sends the basis $\{e_1, e_2, e_3\}$ to the basis $\{\sigma_u(q_0), \sigma_v(q_0), N\}$. By the inverse function theorem (on page 120), f restricts to a diffeomorphism from a neighborhood, \mathcal{A} , of $(q_0, 0)$ in $U \times \mathbb{R} \subset \mathbb{R}^2 \times \mathbb{R} = \mathbb{R}^3$ to a neighborhood, \mathcal{B} , of p_0 in \mathbb{R}^3 . We can assume (after possibly shrinking the neighborhoods) that \mathcal{A} has the form $\mathcal{A} = \tilde{U} \times (-\epsilon, \epsilon)$, where $\epsilon > 0$ and $\tilde{U} \subset U$ is a neighborhood of q_0 .

Define $S = \sigma(\tilde{U})$. Since f is injective on \mathcal{A} , it follows that σ is injective on \tilde{U} , so it has an inverse $\sigma^{-1} : S \rightarrow \tilde{U}$. It remains to prove that σ^{-1} is smooth. For this, let $P : \mathcal{A} \rightarrow U$ denote the natural projection that maps $(q, t) \mapsto q$. The function $P \circ f^{-1} : \mathcal{B} \rightarrow \tilde{U}$ is a smooth function that agrees with σ^{-1} on $\mathcal{B} \cap S$, which verifies that σ^{-1} is smooth. \square

In light of Proposition 3.29, you might guess that the image of every injective parametrized surface is a regular surface, but counterexamples to this guess will be given in Sect. 10.

EXERCISES

EXERCISE 3.13. Prove that a set $S \subset \mathbb{R}^3$ is a regular surface if and only if each of its points has a neighborhood in S that is diffeomorphic to the disk $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$.

EXERCISE 3.14. Show that the cylinder $\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ can be entirely covered by a single surface patch with domain $U = \mathbb{R}^2 - \{(0, 0)\}$.

EXERCISE 3.15. Prove that a connected regular surface is path-connected. This is a converse (for surfaces) to Proposition A.18 on page 352 of the appendix.

EXERCISE 3.16. Let S be a regular surface. If $X \subset S$ is open in S , prove that X is itself a regular surface.

EXERCISE 3.17. Give an example of a smooth function $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ that is a bijection onto its image V such that the inverse $\sigma^{-1} : V \rightarrow \mathbb{R}^2$ is continuous but not smooth.

EXERCISE 3.18 (Generalized Cylinders).

- (1) Let $\gamma : [a, b] \rightarrow \mathbb{R}^3$ be a simple closed space curve whose trace lies in a plane $P \subset \mathbb{R}^3$. Let $n \in \mathbb{R}^3$ be a normal vector to P . Prove that

$$C = \{\gamma(t) + sn \mid t \in [a, b], s \in \mathbb{R}\}$$

is a regular surface that is diffeomorphic to the standard cylinder from Example 3.23. The set C is called a **generalized cylinder**. Notice that C is unaffected by reparametrizing γ or scaling n , so we can assume without loss of generality that γ is of unit speed and that n is of unit length. *HINT: After applying a rigid motion, you can assume that P is the xy -plane. Use Exercise 3.11 (on page 124) to establish the smoothness of the inverse of the natural surface patch.*

- (2) In an attempt to further generalize the definition of a cylinder, let $\tilde{\gamma} : [a, b] \rightarrow \mathbb{R}^3$ be a simple closed space curve (not necessarily contained in any plane). Let $n \in \mathbb{R}^3$ be a unit vector that is not parallel to $\tilde{\gamma}'(t)$ for any $t \in [a, b]$, and define

$$\tilde{C} = \{\tilde{\gamma}(t) + sn \mid t \in [a, b], s \in \mathbb{R}\}.$$

Under what conditions on $\tilde{\gamma}$ and n will \tilde{C} be a regular surface diffeomorphic to the standard cylinder?

- (3) Let $\tilde{\gamma}$, n , and \tilde{C} be as in part (2). Define $\gamma : [a, b] \rightarrow \mathbb{R}^3$ as

$$\gamma(t) = \tilde{\gamma}(t) - \langle \tilde{\gamma}(t), n \rangle \cdot n,$$

and define $C = \{\gamma(t) + sn \mid t \in [a, b], s \in \mathbb{R}\}$, as in part (1); see Fig. 3.15 (left). Show that $\tilde{C} = C$. Show that the trace of γ lies in the plane P containing $(0, 0, 0)$ with normal vector n . By part (1), if γ is regular and simple, then $\tilde{C} = C$ is a regular surface. In summary, \tilde{C} is not really a more general type of cylinder than the type considered in part (1).

EXERCISE 3.19 (Generalized Cones).

- (1) Let $\gamma : [a, b] \rightarrow \mathbb{R}^3$ be a simple closed space curve whose trace lies in the sphere S^2 . Define $C = \{s \cdot \gamma(t) \mid t \in [a, b], s > 0\}$. Prove that S is a regular surface that is diffeomorphic to the standard cone from Example 3.20 (with the origin of the standard cone removed to make it a regular surface). The set C is called a **generalized cone**. Notice that the set C is unaffected by reparametrizing γ , so we can assume without loss of generality that γ is of unit speed.
HINT: Use Exercise 3.12 (on page 124) to establish the smoothness of the inverse of the natural surface patch.
- (2) In an attempt to further generalize the definition of a cone, let $\tilde{\gamma} : [a, b] \rightarrow \mathbb{R}^3$ be a simple closed space curve (not necessarily contained in S^2). Define $\tilde{C} = \{s \cdot \tilde{\gamma}(t) \mid t \in [a, b], s > 0\}$. Under what condition on $\tilde{\gamma}$ will \tilde{C} be a regular surface diffeomorphic to the standard cone from Example 3.20?
- (3) Let $\tilde{\gamma}$ and \tilde{C} be as in part (2). Assume that the trace of $\tilde{\gamma}$ does not contain the origin. Define $\gamma : [a, b] \rightarrow \mathbb{R}^3$ as

$$\gamma(t) = \frac{\tilde{\gamma}(t)}{|\tilde{\gamma}(t)|},$$

and define $C = \{s \cdot \gamma(t) \mid t \in [a, b], s > 0\}$, as in part (1); see Fig. 3.15 (right). Show that $\tilde{C} = C$. Show that the trace of $\tilde{\gamma}$ lies in S^2 . By part (1), if γ is regular and simple, then $\tilde{C} = C$ is a regular surface. In summary, \tilde{C} is not really a more general type of cone than the type considered in part (1).

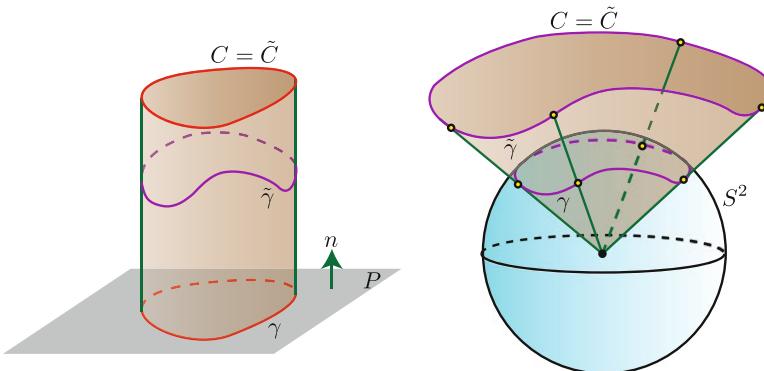


FIGURE 3.15. *Left:* the generalized cylinder from Exercise 3.18. *Right:* the generalized cone from Exercise 3.19

EXERCISE 3.20. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ be smooth. Let $\lambda \in \mathbb{R}$ be such that the preimage $S = f^{-1}(\lambda)$ is nonempty. *Prove or disprove:* λ is a regular value if and only if S is a regular surface.

EXERCISE 3.21. Let $U \subset \mathbb{R}^2$ be open, $f : U \rightarrow \mathbb{R}$ a smooth function, $p \in U$, and $\lambda = f(p)$. If $df_p = \left(\frac{\partial f}{\partial x}(p), \frac{\partial f}{\partial y}(p) \right) \neq (0, 0)$, prove that there exists a neighborhood of p in the preimage $C = f^{-1}(\lambda)$ that is the trace of a regular plane curve.

EXERCISE 3.22. Let $f : \mathbb{R}^3 \rightarrow \mathbb{R}^2$ be smooth. Let $\lambda \in \mathbb{R}^2$ be such that the preimage $S = f^{-1}(\lambda)$ is nonempty. Let $p \in S$. Assume that df_p has rank 2. Prove that there exists a neighborhood of p in S that is the trace of a parametrized space curve.

EXERCISE 3.23 (A Torus). Let S denote the torus obtained by revolving about the z -axis the circle in the xz -plane of radius 1 centered at $(2, 0, 0)$, which can be parametrized as $x(t) = 2 + \cos t$, $z(t) = \sin t$; see Fig. 3.16. Verify that the surface patch constructed in Example 3.25 becomes

$$\sigma(\theta, t) = ((2 + \cos t) \cos \theta, (2 + \cos t) \sin \theta, \sin t).$$

Verify that S is the $F = 0$ level surface of the function $F : \mathbb{R}^3 \rightarrow \mathbb{R}$ defined as

$$F(x, y, z) = (x^2 + y^2 + z^2 + 3)^2 - 16(x^2 + y^2).$$

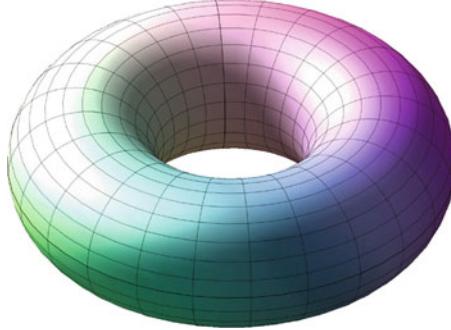


FIGURE 3.16. The torus of revolution from Exercise 3.23

EXERCISE 3.24 (A Tangent Developable). Let $\gamma : (a, b) \rightarrow \mathbb{R}^3$ be a regular space curve with nowhere-vanishing curvature. Let C denote its trace. Define $U = \{(s, t) \in \mathbb{R}^2 \mid s > 0, t \in (a, b)\}$. The function $\sigma : U \rightarrow \mathbb{R}^3$ defined as

$$\sigma(s, t) = \gamma(t) + s\gamma'(t)$$

is called a *tangent developable*.

- (1) Verify that σ is a parametrized surface.
- (2) Use a computer graphing application to plot a tangent developable of the helix $\gamma(t) = (\cos t, \sin t, t)$.

EXERCISE 3.25. Define $\sigma : \{(x, y) \in \mathbb{R}^2 \mid x > 0\} \rightarrow \mathbb{R}^3$ as

$$\sigma(x, y) = (x \cos y, x \sin y, x).$$

Verify that σ is a parametrized surface. Describe the image of σ .

EXERCISE 3.26. Which of the following functions $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ are parametrized surfaces?

- (1) $\sigma(u, v) = (u^2, v^2, u^2 + v^2 + u + v)$,
- (2) $\sigma(u, v) = (u, u^2, v^3)$,
- (3) $\sigma(u, v) = (\cos u, \sin v, \sin(u + v))$.

EXERCISE 3.27. The function $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as

$$\sigma(u, v) = \left(u - \frac{u^3}{3} + uv^2, v - \frac{v^3}{3} + vu^2, u^2 - v^2 \right)$$

is called **Enneper's surface**. It is illustrated in Fig. 3.17. Verify that σ is a parametrized surface.

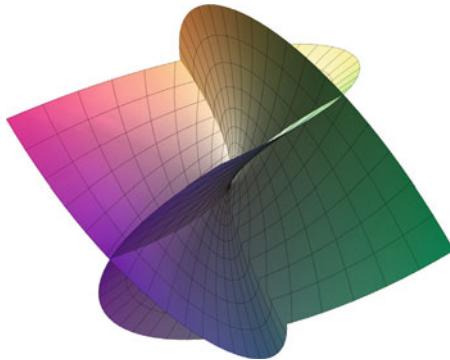


FIGURE 3.17. Enneper's surface

EXERCISE 3.28. Let $\gamma : (a, b) \rightarrow \mathbb{R}^3$ be a unit-speed space curve whose curvature function satisfies $0 < \kappa(t) < \frac{1}{\epsilon}$ for all $t \in (a, b)$. Define $\phi : (a, b) \times (0, 2\pi) \rightarrow \mathbb{R}^3$ exactly as in Exercise 3.12 on page 124:

$$\phi(t, \theta) = \gamma(t) + (\epsilon \cdot \cos \theta) \cdot \mathbf{n}(t) + (\epsilon \cdot \sin \theta) \cdot \mathbf{b}(t).$$

Prove that ϕ is a parametrized surface.

EXERCISE 3.29 (Ruled Surfaces).

- (1) The trace of the space curve $\gamma(t) = (1, 0, 0) + t(0, 1, 1)$ is a straight line. Define $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ as $\sigma(\theta, t) = R_\theta(\gamma(t))$, where R_θ denotes the rotation through the angle θ about the z -axis (the matrix representing R_θ is mentioned in Example 3.25). Demonstrate that σ is a parametrized surface. Show that the hyperboloid of one sheet (the graph, G , of $x^2 + y^2 - z^2 = 1$) equals the image of σ . Conclude that the hyperboloid of one sheet equals the union of a collection of nonintersecting straight lines; see Fig. 3.18 (left). Identify a second collection of nonintersecting straight lines whose union also equals G , such that every line in the second collection intersects all but one of the lines in the first collection and vice versa.

- (2) Define $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ as $\sigma(s, t) = (t, 0, 0) + s(0, 1, t)$. Demonstrate that σ is a parametrized surface. Show that the image of σ equals the graph, S , of the equation $z = xy$. Conclude that S also equals the union of a collection of nonintersecting straight lines; see Fig. 3.18 (right). Observe that the parameter curves of the form $t \mapsto \sigma(s_0, t)$ and of the form $s \mapsto \sigma(s, t_0)$ are straight lines.

COMMENT: A **ruled parametrized surface** is a parametrized surface $\sigma : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ that has the form $\sigma(s, t) = \gamma(s) + tw(s)$, where γ and w are space curves. For example, every tangent developable (Exercise 3.24 on page 138) is a ruled parametrized surface. A **ruled surface** is a regular surface that can be covered by an atlas of ruled parametrized surfaces. In addition to the hyperboloid of one sheet and the saddle surface described above, examples include generalized cylinders (Exercise 3.18) and generalized cones (Exercise 3.19). The straight lines that constitute a ruled surface are called **rulings**.

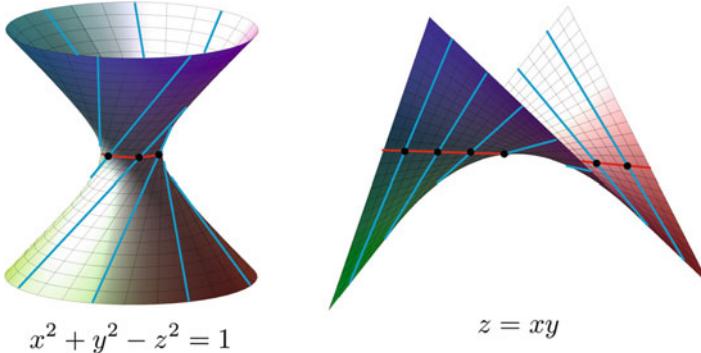


FIGURE 3.18. Ruled surfaces

EXERCISE 3.30. Define $f : \mathbb{R}^3 \rightarrow \mathbb{R}$ as $f(x, y, z) = x^2yz^3$. For which values of $\lambda \in \mathbb{R}$ is $f^{-1}(\lambda)$ a regular surface?

EXERCISE 3.31. Define $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ as $\sigma(u, v) = (u \cosh v, u \sinh v, u^2)$. Verify that the image of σ is contained in the graph of $z = x^2 - y^2$. Is this image all of the graph? Is σ a parametrized surface?

EXERCISE 3.32. For an arbitrary integer $m \geq 2$, define

$$S_m = \{(x, y, z) \in \mathbb{R}^3 \mid x^m + y^m + z^m = 1\}.$$

- (1) Prove that S_m is a regular surface.
- (2) Use a computer graphing application to plot S_m for several choices of m . What does S_m look like for large m ?

EXERCISE 3.33. Use a computer graphing application to plot the images of the following functions. Which point(s) of the domain of each function must be removed (if any) to make it a parametrized surface?

- (1) $\sigma(u, v) = (u^3, v^3, uv)$ with $u, v \in (-1, 1)$.
 - (2) $\sigma(u, v) = (u, v, u^3 - 3uv^2)$ with $u, v \in (-10, 10)$.
 - (3) $\sigma(u, v) = (\cos u \cos v \sin v, \sin u, \cos v \sin v, \sin v)$ with $u, v \in (-2\pi, 2\pi)$.
 - (4) $\sigma(u, v) = (uv, u, v^2)$ with $u, v \in (-10, 10)$.
 - (5) $\sigma(u, v) = (\sin u, \sin v, \sin(u + v))$ with $u, v \in (-2\pi, 2\pi)$.
 - (6) $\sigma(u, v) = ((1 - u^2) \sin v, (1 - u^2) \sin(2v), u)$ with $u, v \in (-\pi, \pi)$.
-

□

3. Tangent Planes

In this section, we define the tangent plane to a surface at a point, and we modify the previous definition of *derivative* to apply to functions whose domains are surfaces. It all begins with the following definition:

DEFINITION 3.30.

Let S be a regular surface. A **regular curve in S** means a regular curve in \mathbb{R}^3 whose trace is contained in S . The **tangent plane** to S at the point $p \in S$ is the set of all initial velocity vectors of regular curves in S with initial position p . That is,

$$T_p S = \{\gamma'(0) \mid \gamma \text{ is a regular curve in } S \text{ with } \gamma(0) = p\}.$$

You can consider the domain of such a curve γ to be $(-\epsilon, \epsilon)$ for some small $\epsilon > 0$. If γ is defined on a larger domain, this is irrelevant, because only $\gamma'(0)$ matters here. A vector will be called “**tangent to S at p** ” when it lies in $T_p S$.

The tangent plane to S at p can be characterized in terms of a coordinate chart $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ with $p \in V$. Setting $q = \sigma^{-1}(p)$, and $\{u, v\}$ the coordinate variables of U , we have the following:

LEMMA 3.31.

$T_p S = \text{span}\{\sigma_u(q), \sigma_v(q)\}$. In particular, $T_p S$ is a two dimensional subspace of \mathbb{R}^3 .

PROOF. Definition 3.30 is local, so it would be unchanged if “regular curve in S ” were replaced with “regular curve in V . Since $\sigma : U \rightarrow V$ is a diffeomorphism, the regular curves in V with initial position p are exactly the compositions with σ of the regular curves in U with initial position q . Proposition 3.7 (on page 118) gives

$$\begin{aligned} \text{span}\{\sigma_u(q), \sigma_v(q)\} &= \text{image}(d\sigma_q) \\ &= \{(\sigma \circ \beta)'(0) \mid \beta \text{ is a regular curve in } U \text{ with } \beta(0) = q\} \\ &= \{\gamma'(0) \mid \gamma \text{ is a regular curve in } V \text{ with } \gamma(0) = p\} \\ &= T_p S. \end{aligned}$$

□

Notice that $T_p S \subset \mathbb{R}^3$ is a subspace, so it passes through the origin. Its translation to p , denoted by $p + T_p S = \{p + v \mid v \in T_p S\}$, passes through p and is tangent to the surface, as in Fig. 3.19. Small nearsighted inhabitants of S living at p believe that they live on a plane, and $p + T_p S$ is exactly the plane they believe they live on. If you zoom in enough near p , then S and $p + T_p S$ become visually indistinguishable (see Exercise 3.37 for a more precise formulation of this assertion). For more illustrations of tangent planes, look back at the figures in the previous section illustrating σ_u and σ_v ; in each such figure, it is easy to picture their span.

We might instead have defined $T_p S$ as $\text{span}\{\sigma_u(q), \sigma_v(q)\}$, but that would have obliged us to prove that the definition is independent of the choice of surface patch. To avoid this calculation, we constructed an initial geometric definition of $T_p S$ that didn't refer to any surface patch, and then showed how our definition is related to any particular surface patch. We will employ this general strategy as often as possible when defining new geometric objects and measurements on surfaces throughout this book.

EXAMPLE 3.32. An open set $U \subset \mathbb{R}^2$ can be regarded as a regular surface, as mentioned in Example 3.22 on page 128. Notice that $T_p U = \mathbb{R}^2$ for every $p \in U$.

The derivative of a function whose domain is a surface can be defined by turning the conclusion of Proposition 3.7 (on page 118) into a definition:

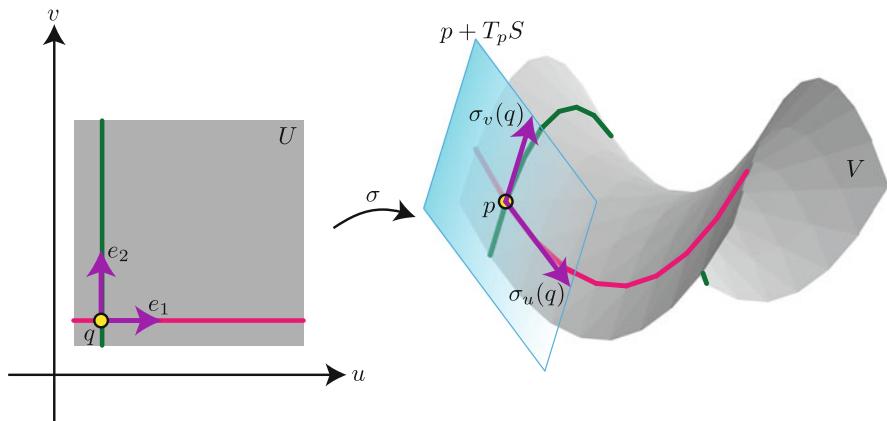


FIGURE 3.19. The translation to p of the subspace $T_p S$

DEFINITION 3.33.

Let S be a regular surface and $f : S \rightarrow \mathbb{R}^n$ a smooth function (as in Def. 3.12). The **derivative** of f at $p \in S$ is the linear transformation $df_p : T_p S \rightarrow \mathbb{R}^n$ defined so that for all $v \in T_p S$,

$$df_p(v) = (f \circ \gamma)'(0),$$

where γ is any regular curve in S with $\gamma(0) = p$ and $\gamma'(0) = v$.

Why is df_p linear, and why is the definition well defined (independent of the choice of γ)? Because the smoothness of f means that there is a neighborhood, \mathcal{O} , of p in \mathbb{R}^3 and a smooth function $\tilde{f} : \mathcal{O} \rightarrow \mathbb{R}^n$ that agrees with f on $\mathcal{O} \cap S$. For every $v \in \mathbb{R}^3$, Proposition 3.7 says that $d\tilde{f}_p(v) = (\tilde{f} \circ \gamma)'(0)$, where γ is any regular curve in \mathbb{R}^3 with $\gamma(0) = p$ and $\gamma'(0) = v$. If $v \in T_p S$, then regular curves in S exist that satisfy these initial conditions. Since $\tilde{f} = f$ along such a regular curve γ in S , $d\tilde{f}_p(v) = (\tilde{f} \circ \gamma)'(0) = (f \circ \gamma)'(0) = df_p(v)$. In summary, df_p equals the restriction to the domain $T_p S$ of the well-defined linear transformation $d\tilde{f}_p : \mathbb{R}^3 \rightarrow \mathbb{R}^n$.

If $f : S_1 \rightarrow S_2$ is a smooth function between a pair of regular surfaces, and $p \in S_1$, then $df_p : T_p S_1 \rightarrow \mathbb{R}^3$ makes sense using the above definition, by ignoring S_2 and regarding f as a function from S_1 to the ambient \mathbb{R}^3 . But don't ignore S_2 for too long, or you'll miss the useful fact that the image of df_p lies in $T_{f(p)} S_2$, simply because f sends curves in S_1 to curves in S_2 . This observation is illustrated in Fig. 3.20 and summarized in the following proposition:

PROPOSITION 3.34.

If $f : S_1 \rightarrow S_2$ is smooth, then for every $p \in S_1$,

$$df_p : T_p S_1 \rightarrow T_{f(p)} S_2.$$

The inverse function theorem is one of the most important results about the derivatives of functions between open sets in Euclidean spaces. We end this section with a surface version of this fundamental theorem.

THEOREM 3.35 (The Inverse Function Theorem for Surfaces).

Let S_1 and S_2 be regular surfaces and $f : S_1 \rightarrow S_2$ a smooth function.

For $p \in S_1$, if $df_p : T_p S_1 \rightarrow T_{f(p)} S_2$ is invertible, then f restricted to a sufficiently small neighborhood of p in S_1 is a diffeomorphism onto its image, and this image is a neighborhood of $f(p)$ in S_2 .

The small nearsighted inhabitants of S_1 at p and of S_2 at $f(p)$ believe they live on \mathbb{R}^2 . From their point of view, this is the original inverse function theorem (on page 120). So for the proof, we should look at things from their point of view, which is achieved with coordinate charts.

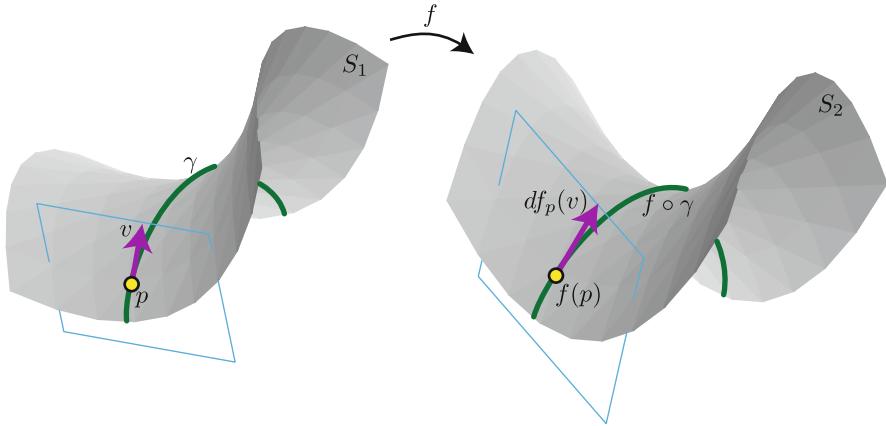


FIGURE 3.20. $df_p(v) \in T_{f(p)}S_2$ is the initial velocity vector of the image under f of *any* regular curve γ in S_1 with $\gamma(0) = p$ and $\gamma'(0) = v$

PROOF. Consider coordinate charts $\sigma_1 : U_1 \rightarrow V_1$ and $\sigma_2 : U_2 \rightarrow V_2$, where V_1 is a neighborhood of p in S_1 , while V_2 is a neighborhood of $f(p)$ in S_2 . After shrinking U_1 and V_1 if necessary, we can assume that $f(V_1) \subset V_2$ (more precisely, redefine V_1 as $f^{-1}(V_2)$ and then redefine U_1 as $\sigma_1^{-1}(V_1)$). Define $\psi : U_1 \rightarrow U_2$ as the following composition:

$$\psi = \sigma_2^{-1} \circ f \circ \sigma_1.$$

Think of ψ as the function f from the point of view of inhabitants who are nearsighted enough to identify $U_1 \cong V_1$ and $U_2 \cong V_2$. Set $q = \sigma_1^{-1}(p) \in U_1$. By the chain rule (applied twice),

$$d\psi_q = d(\sigma_2^{-1})_{f(p)} \circ df_p \circ d(\sigma_1)_q.$$

All of this is summarized in the following diagram, in which the left side shows the functions and the right side shows their derivatives at the relevant points:

$$\begin{array}{ccc}
 V_1 \subset S_1 & \xrightarrow{f} & V_2 \subset S_2 \\
 \uparrow \sigma_1 & & \uparrow \sigma_2 \\
 U_1 & \xrightarrow{\psi} & U_2
 \end{array}
 \quad
 \begin{array}{ccc}
 T_p S_1 & \xrightarrow{df_p} & T_{f(p)} S_2 \\
 \uparrow d(\sigma_1)_q & & \uparrow d(\sigma_2)_{\psi(q)} \\
 \mathbb{R}^2 & \xrightarrow{d\psi_q} & \mathbb{R}^2
 \end{array}$$

For the chain rule (Proposition 3.9 on page 119) to be valid here, we need the domains of all of the functions to be open sets of Euclidean spaces, so we must consider f and (σ_2^{-1}) to be extended to smooth functions on open neighborhoods in \mathbb{R}^3 of p and $f(p)$ respectively. This is possible because (according to Definition 3.12 on page 121) that's exactly what it means that the functions are smooth. We are abusing notation slightly by referring to the extensions by the same names as the original functions.

Thus, $d\psi_q$ is the composition of three linear transformations, and we claim that all three are invertible. First, $d(\sigma_1)_q : \mathbb{R}^2 \rightarrow T_p S_1$ is invertible by Proposition 3.21 (on page 126). Second, $df_p : T_p S_1 \rightarrow T_{f(p)} S_2$ is invertible by hypothesis. Third, $d(\sigma_2^{-1})_{f(p)} : T_{f(p)} S_2 \rightarrow \mathbb{R}^2$ is invertible because it equals $(d(\sigma_2)_{\psi(q)})^{-1}$, so it is the inverse of an invertible linear transformation.

In summary, $d\psi_q$ is the composition of three invertible linear transformations between two-dimensional vector spaces, so it must be invertible. The inverse function theorem (Theorem 3.10 on page 120) now implies that after possibly shrinking U_1 and U_2 to smaller neighborhoods of q and $\psi(q)$ respectively, ψ is a diffeomorphism. After redefining U_1, U_2, V_1, V_2 as the corresponding smaller neighborhoods,

$$f = \sigma_2 \circ \psi \circ \sigma_1^{-1} : V_1 \rightarrow V_2$$

is the composition of three diffeomorphisms, so it is a diffeomorphism. \square

EXERCISES

EXERCISE 3.34. Definition 3.30 can be used to define $T_p S$ when $S \subset \mathbb{R}^3$ is an *arbitrary* set and $p \in S$. Give a variety of examples in which $T_p S$ is not a subspace.

EXERCISE 3.35. Let S be a regular surface and $p \in S$. Suppose $P \subset \mathbb{R}^3$ is a plane such that $p \in P$ and S lies on one (closed) side of P . Prove that $T_p S = P$.

EXERCISE 3.36. Let S be the torus from Exercise 3.23 (on page 138), which was parametrized as

$$\sigma(\theta, t) = ((2 + \cos t) \cos \theta, (2 + \cos t) \sin \theta, \sin t).$$

Describe the tangent plane at an arbitrary point of the $t = 0$ parameter curve and at an arbitrary point of the $\theta = 0$ parameter curve of S . Draw an illustration.

EXERCISE 3.37. Let S be a regular surface and $p \in S$. Let $N \in \mathbb{R}^3$ denote a normal vector to $T_p S$. If p_1, p_2, \dots is an infinite sequence of points of S converging to p , prove that

$$\lim_{n \rightarrow \infty} \left\langle N, \frac{p_n - p}{|p_n - p|} \right\rangle = 0.$$

EXERCISE 3.38. Let $U \subset \mathbb{R}^2$ be open, $f : U \rightarrow \mathbb{R}$ smooth, G its graph, and $q = (x, y) \in U$. Prove that the tangent plane to G at $(x, y, f(x, y))$ is the graph of df_q .

EXERCISE 3.39. It is useful to characterize smoothness in local coordinates. For example, smoothness of a real-valued function on S^2 should mean that it is a smooth function of the latitude and longitude parameters (at least at points other than the north and south poles). Prove the following:

- (1) If $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ is a surface patch, then a function $f : V \rightarrow \mathbb{R}$ is smooth (in the sense of Definition 3.12 on page 121) if and only if $f \circ \sigma : U \rightarrow \mathbb{R}$ is smooth.
- (2) If $f : S_1 \rightarrow S_2$ is a function between regular surfaces, and for $i \in \{1, 2\}$, $\sigma_i : U_i \subset \mathbb{R}^2 \rightarrow V_i \subset S_i$ is a surface patch, and $f(V_1) \subset V_2$, then the restriction of f to V_1 is smooth if and only if $\sigma_2^{-1} \circ f \circ \sigma_1 : U_1 \rightarrow U_2$ is smooth.

EXERCISE 3.40. Let S be a path-connected regular surface, and $f : S \rightarrow \mathbb{R}$ a smooth function. Prove that f is a constant function if and only if $df_p(v) = 0$ for all $p \in S$ and all $v \in T_p S$.

EXERCISE 3.41. Is the converse to Theorem 3.35 true?

EXERCISE 3.42. Prove the following converse to Exercise 3.16 (on page 136): Let S be a regular surface. If a subset $X \subset S$ is itself a regular surface, then X is open in S . *HINT: Apply Theorem 3.35 to a surface patch for X to learn that after possibly shrinking its domain, it becomes a surface patch for S .*

EXERCISE 3.43. If S is a regular surface and $\sigma : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is an injective parametrized surface whose image lies in S , prove that σ is a surface patch for S .

EXERCISE 3.44. Let S be a regular surface and $f : S \rightarrow \mathbb{R}$ a smooth function. Assume that the point $p \in S$ is a **critical point** of f , which means that $df_p(v) = 0$ for all $v \in T_p S$. Define the **Hessian** of f at p in the direction v as

$$\text{Hess}(f)_p(v) = (f \circ \gamma)''(0),$$

where γ is a regular curve in S with $\gamma(0) = p$ and $\gamma'(0) = v$. Prove that the Hessian is well defined in the sense that it is independent of the choice of γ . *HINT: Work in local coordinates and use Exercise 3.5 on page 123.*

EXERCISE 3.45. Let S be a regular surface and let $q \in \mathbb{R}^3$ with $q \notin S$. Define $f : S \rightarrow \mathbb{R}$ as $f(p) = \text{dist}(p, q)$. If $p \in S$ is a global minimum or maximum of f , prove that $q - p$ is orthogonal to every vector in $T_p S$.

EXERCISE 3.46 (The Chain Rule for Surfaces). Let $S_1 \xrightarrow{g} S_2 \xrightarrow{f} S_3$ denote smooth functions between regular surfaces. Prove that for all $q \in S_1$,

$$d(f \circ g)_q = df_p \circ dg_q,$$

where $p = g(q)$.

EXERCISE 3.47. Let S be a regular surface, and let $p \in S$. Assume that $T_p S = \text{span}\{e_1, e_2\}$, where $e_1 = (1, 0, 0)$ and $e_2 = (0, 1, 0)$. Prove there exist an open set $U \subset \mathbb{R}^2$ and a neighborhood of p in S that is the graph of a smooth function from U to \mathbb{R} . *HINT: Apply Theorem 3.35 to the function from S to \mathbb{R}^2 defined as $f(x, y, z) = (x, y)$.*

EXERCISE 3.48. Let S be a regular surface, and let $p \in S$. Prove that there exists a neighborhood of p in S that is the graph of a smooth function of one of the following forms: $z = f(x, y)$, $y = f(x, z)$, $x = f(y, z)$.

EXERCISE 3.49. Prove that the cone in Example 3.20 (on page 126) is not a regular surface.

EXERCISE 3.50. In Exercise 3.12 (on page 124), prove that S_ϵ is a regular surface for sufficiently small $\epsilon > 0$. *HINT: Use Exercise 3.28 and modify the hint from Exercise 3.11.*



4. Area Distortion and Orientation (*Linear Algebra Background*)

As the name *differential* geometry suggests, to study a smooth function between surfaces, $f : S \rightarrow \tilde{S}$, one should study its derivative $df_p : T_p S \rightarrow T_{f(p)}\tilde{S}$ at points $p \in S$. This idea will be a recurring theme in the next several sections. In preparation, we will devote this section to an abstract discussion of the type of object that df_p is in most situations of interest to us, namely, a linear isomorphism between a pair of two-dimensional subspaces of \mathbb{R}^3 .

So throughout this section, \mathcal{V} and $\tilde{\mathcal{V}}$ will denote a pair of two-dimensional subspaces of \mathbb{R}^3 , and $g : \mathcal{V} \rightarrow \tilde{\mathcal{V}}$ will denote a linear isomorphism.

DEFINITION 3.36.

We say that the matrix $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ represents g with respect to the ordered bases $B = \{v_1, v_2\}$ of \mathcal{V} and $\tilde{B} = \{\tilde{v}_1, \tilde{v}_2\}$ of $\tilde{\mathcal{V}}$ if

$$g(v_1) = a\tilde{v}_1 + b\tilde{v}_2 \quad \text{and} \quad g(v_2) = c\tilde{v}_1 + d\tilde{v}_2.$$

In other words, if we identify $\mathcal{V} \cong \mathbb{R}^2$ via $a_1v_1 + a_2v_2 \leftrightarrow (a_1, a_2)$, and we identify $\tilde{\mathcal{V}} \cong \mathbb{R}^2$ via $a_1\tilde{v}_1 + a_2\tilde{v}_2 \leftrightarrow (a_1, a_2)$, then g becomes identified with the linear transformation $L_A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$, as depicted in the following diagram:

$$\begin{array}{ccc} \mathcal{V} & \xrightarrow{g} & \tilde{\mathcal{V}} \\ \uparrow & & \uparrow \\ \mathbb{R}^2 & \xrightarrow{L_A} & \mathbb{R}^2 \end{array}$$

We next formulate a measurement of the factor by which g distorts area. To interpret the following, recall that the norm of the cross product of two vectors equals the area of the parallelogram that they span:

PROPOSITION AND DEFINITION 3.37.

The following expression, which is called the **area distortion** of g , and is denoted by $\|g\|$, does not depend on the choice of basis $B = \{v_1, v_2\}$ of \mathcal{V} :

$$\|g\| = \frac{|g(v_1) \times g(v_2)|}{|v_1 \times v_2|}.$$

PROOF. An arbitrary other basis has the form $\{av_1 + bv_2, cv_1 + dv_2\}$ for some scalars $a, b, c, d \in \mathbb{R}$ with $ad - bc \neq 0$. Using algebraic properties of the cross product (Sect. 7 of Chap. 1), we see that the area distortion computed with respect to this arbitrary basis is the same as with respect to $\{v_1, v_2\}$, because

$$\frac{|g(av_1 + bv_2) \times g(cv_1 + dv_2)|}{|(av_1 + bv_2) \times (cv_1 + dv_2)|} = \frac{|(ad - bc)(g(v_1) \times g(v_2))|}{|(ad - bc)(v_1 \times v_2)|} = \frac{|g(v_1) \times g(v_2)|}{|v_1 \times v_2|}.$$

□

The term “area distortion” is appropriate because every pair of vectors in \mathcal{V} spanning a parallelogram with area A will be sent by g to a pair of vectors in $\tilde{\mathcal{V}}$ spanning a parallelogram with area $\|g\| \cdot A$. The notation $\|g\|$ is nonstandard. The boldface is intended to visually distinguish it, even though this book contains no other uses of a double vertical bar. The notation is natural because one can think of the outer vertical bars as the absolute value, and the inner vertical bars as the bars that are sometimes used to denote determinants:

PROPOSITION 3.38.

$\|g\|$ equals the absolute value of the determinant of the matrix representing g with respect to any choice of orthonormal bases for \mathcal{V} and $\tilde{\mathcal{V}}$.

PROOF. If $B = \{v_1, v_2\}$ and $\tilde{B} = \{\tilde{v}_1, \tilde{v}_2\}$ are arbitrary bases for \mathcal{V} and $\tilde{\mathcal{V}}$ respectively, and $A = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ represents g with respect to these bases, then

$$(3.10) \quad g(v_1) \times g(v_2) = (a\tilde{v}_1 + b\tilde{v}_2) \times (c\tilde{v}_1 + d\tilde{v}_2) = (ad - bc)(\tilde{v}_1 \times \tilde{v}_2) = \det(A)(\tilde{v}_1 \times \tilde{v}_2).$$

Therefore,

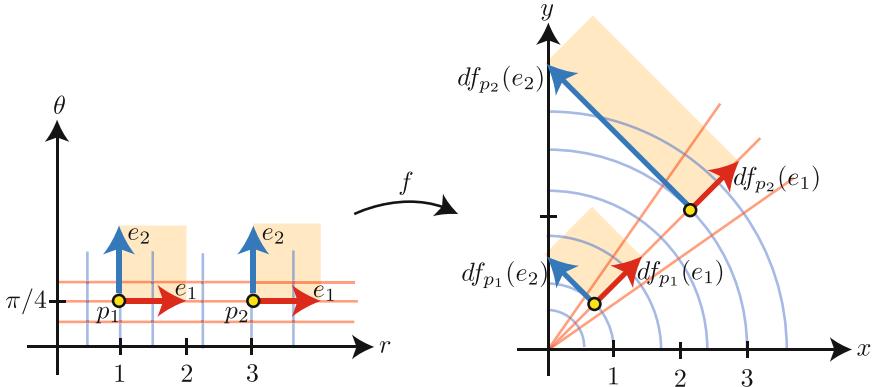
$$\|g\| = \frac{|g(v_1) \times g(v_2)|}{|v_1 \times v_2|} = |\det(A)| \frac{|\tilde{v}_1 \times \tilde{v}_2|}{|v_1 \times v_2|},$$

which equals $|\det(A)|$ if the two bases are orthonormal. □

EXAMPLE 3.39. For the polar coordinate function $f(r, \theta) = (r \cos \theta, r \sin \theta)$, we computed in Example 3.6 (on page 117) that at $p = (r, \theta)$,

$$df_p = \begin{pmatrix} \cos \theta & -r \sin \theta \\ \sin \theta & r \cos \theta \end{pmatrix}$$

(with respect to the standard orthonormal basis). Therefore, $\|df_p\| = \det(df_p) = r$. Figure 3.21 illustrates the images under df_p of $e_1 = (1, 0)$ and $e_2 = (0, 1)$ at the two points $p = p_1 = (1, \pi/4)$ and $p = p_2 = (3, \pi/4)$. We'll soon discuss something that is apparent in this figure—the area distortion of df_p reflects the amount by which f itself distorts area in a small neighborhood of p .

FIGURE 3.21. $\|df_{p_1}\| = 1$ and $\|df_{p_2}\| = 3$

Recall that the determinant of the matrix representing a linear transformation from a two-dimensional vector space *to itself* (with respect to a single basis of the vector space) does not depend on the choice of basis, and its absolute value represents area distortion. But in our situation, where g is a linear transformation between *different* vector spaces, we must insist that the bases be *orthonormal* in order for the absolute value of the determinant to represent area distortion, and to be basis-independent.

But what about the sign of the determinant? If we reorder only one of the bases, for example as $B = \{v_1, v_2\}$ and $\tilde{B} = \{\tilde{v}_2, \tilde{v}_1\}$, then the determinant of the matrix representing g will clearly change sign. To ensure that the sign of the determinant is basis-independent, we must insist that our bases not only be orthonormal but also *positively oriented*, defined as follows:

DEFINITION 3.40.

An **orientation** for \mathcal{V} means a choice of a unit-length normal vector, N , to \mathcal{V} . With respect to a given orientation, N , for \mathcal{V} , an ordered basis $\{v_1, v_2\}$ of \mathcal{V} is called **positively oriented** if $\frac{v_1 \times v_2}{|v_1 \times v_2|} = N$. The only other possibility is that $\frac{v_1 \times v_2}{|v_1 \times v_2|} = -N$, in which case it is called **negatively oriented**.

Notice that there are exactly two possible orientations for \mathcal{V} , and each is the negative of the other. Changing to the other orientation would reverse the issue of which ordered bases are positively oriented and which are negatively oriented.

An orientation for \mathcal{V} allows us to define a “counterclockwise rotation by 90° ” map $R_{90} : \mathcal{V} \rightarrow \mathcal{V}$ (as done in Sect. 6 of Chap. 1 when $\mathcal{V} = \mathbb{R}^2$). It is because of the orientation that “counterclockwise” makes sense here. Specifically, defining R_{90} such that for all $a, b \in \mathbb{R}$,

$$(3.11) \quad R_{90}(av_1 + bv_2) = -bv_1 + av_2,$$

can be shown not to depend on the choice of ordered oriented orthonormal basis $\{v_1, v_2\}$ of \mathcal{V} (Exercise 3.51). In fact, this transformation R_{90} has the following basis-free description:

$$(3.12) \quad R_{90}(v) = N \times v.$$

DEFINITION 3.41.

The map g is called **orientation-preserving** (with respect to given orientations of \mathcal{V} and $\tilde{\mathcal{V}}$) if the following equivalent conditions are satisfied:

- (1) If $B = \{v_1, v_2\}$ is a positively oriented ordered basis of \mathcal{V} , then $\{g(v_1), g(v_2)\}$ is a positively oriented ordered basis of $\tilde{\mathcal{V}}$.
- (2) The matrix representing g with respect to any positively oriented ordered bases for \mathcal{V} and $\tilde{\mathcal{V}}$ has positive determinant.

PROOF OF EQUIVALENCE. Suppose that $\{v_1, v_2\}$ and $\{\tilde{v}_1, \tilde{v}_2\}$ are positively oriented ordered bases for \mathcal{V} and $\tilde{\mathcal{V}}$ respectively. Let $A = (\begin{smallmatrix} a & c \\ b & d \end{smallmatrix})$ be the matrix that represents g with respect to these bases. Equation 3.10 says that

$$g(v_1) \times g(v_2) = \det(A) (\tilde{v}_1 \times \tilde{v}_2).$$

Therefore, $\{g(v_1), g(v_2)\}$ is positively oriented if and only if $\det(A) > 0$. \square

In summary, the determinant of the matrix representing g does not depend on the choice of positively oriented orthonormal ordered bases for \mathcal{V} and $\tilde{\mathcal{V}}$. Its sign (positive or negative) signifies whether g is orientation-preserving or reversing, while its absolute value equals the area distortion of g .

EXERCISES

EXERCISE 3.51. Prove that the definition of R_{90} in Equation 3.11 is independent of the choice of ordered oriented orthonormal basis $\{v_1, v_2\}$ of \mathcal{V} .

EXERCISE 3.52. How is the definition of a positively oriented basis that is found in this section related to Definition 1.62 on page 53?

EXERCISE 3.53. Give an example of $g : \mathcal{V} \rightarrow \tilde{\mathcal{V}}$ (as in this section) demonstrating that:

- (1) The absolute value of the determinant of the matrix representing g depends on the choice of (not necessarily orthonormal) bases of \mathcal{V} and $\tilde{\mathcal{V}}$.
- (2) The sign of the determinant of the matrix representing g depends on the choice of bases of \mathcal{V} and $\tilde{\mathcal{V}}$.

EXERCISE 3.54. Let \mathcal{V} denote the xy -plane in \mathbb{R}^3 , oriented by the vector $N = (0, 0, 1)$. Define $f : \mathcal{V} \rightarrow \mathcal{V}$ as $f(x, y, 0) = (2x - 3y, -5x + y, 0)$. Compute $\|f\|$ and decide whether f is orientation-preserving or orientation-reversing.

EXERCISE 3.55. Let \mathcal{V} denote the xy -plane in \mathbb{R}^3 , oriented by $N = (0, 0, 1)$. Let $\tilde{\mathcal{V}}$ denote the yz -plane, oriented by $N = (-1, 0, 0)$. Is the function $f : \mathcal{V} \rightarrow \tilde{\mathcal{V}}$ defined as $f(x, y, 0) = (0, y, x)$ orientation-preserving or orientation-reversing?

EXERCISE 3.56. Let $\mathcal{V}_1, \mathcal{V}_2, \mathcal{V}_3 \subset \mathbb{R}^3$ be two-dimensional subspaces. Let $g : \mathcal{V}_1 \rightarrow \mathcal{V}_2$ and $f : \mathcal{V}_2 \rightarrow \mathcal{V}_3$ be linear isomorphisms.

- (1) Prove that $\|f \circ g\| = \|f\| \cdot \|g\|$.
- (2) Prove that $\|g^{-1}\| = \frac{1}{\|g\|}$.

EXERCISE 3.57. Let $\lambda \in \mathbb{R}$ and define $\mathcal{V} = \text{span}\{(1, 0, 0), (0, 1, \lambda)\}$ and $\tilde{\mathcal{V}} = \text{span}\{(1, 0, 0), (0, 1, 0)\}$. Compute the area distortion of $g : \mathcal{V} \rightarrow \tilde{\mathcal{V}}$ defined such that $g(x, y, z) = (x, y, 0)$ for all $(x, y, z) \in \mathcal{V}$.

EXERCISE 3.58. Let $\mathcal{V}, \tilde{\mathcal{V}} \subset \mathbb{R}^3$ be a pair of two-dimensional subspaces with orientations. Let $g : \mathcal{V} \rightarrow \tilde{\mathcal{V}}$ be an orientation-preserving linear isomorphism. Let $\{v_1, v_2\}$ be a positively oriented orthonormal basis of \mathcal{V} . Define $\tilde{v}_1 = \frac{g(v_1)}{\|g(v_1)\|}$ and $\tilde{v}_2 = R_{90}(\tilde{v}_1)$, so that $\{\tilde{v}_1, \tilde{v}_2\}$ is a positively oriented orthonormal basis of $\tilde{\mathcal{V}}$. If $v = av_1 + bv_2$ and $g(v) = \tilde{a}\tilde{v}_1 + \tilde{b}\tilde{v}_2$, prove that (\tilde{a}, \tilde{b}) is not a negative scalar multiple of (a, b) .

□

5. Orientable Surfaces

In this section, we define and explore *orientable* surfaces. At a point p of a regular surface S , we previously visualized $T_p S$ as the set of directions in which an inhabitant of S could travel away from p while remaining on S . But what about directions that leave S and move directly away from it into the ambient \mathbb{R}^3 ?

DEFINITION 3.42.

A **normal vector** to S at p means a vector in \mathbb{R}^3 that is orthogonal to $T_p S$. That is, $N \in \mathbb{R}^3$ is called a normal vector to S at p if $\langle N, v \rangle = 0$ for all $v \in T_p S$. A **unit normal vector** to S at p means a unit-length normal vector, or equivalently, an orientation for $T_p S$.

Let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a coordinate chart with $\sigma(q) = p$ as before. Since $\sigma_u(q)$ and $\sigma_v(q)$ span $T_p S$, their cross product is orthogonal to $T_p S$, so the following is a unit normal vector at p :

$$(3.13) \quad N(p) = \frac{\sigma_u(q) \times \sigma_v(q)}{|\sigma_u(q) \times \sigma_v(q)|}.$$

In fact, there are only two unit normal vectors at p , namely $N(p)$ and $-N(p)$. The first choice might seem best, but it's not special, because it

depends on σ . With a different coordinate chart, Eq. 3.13 might yield the other choice. In fact, it is always possible to construct a different coordinate chart that yields the other choice. For this, define the domain $\hat{U} = \{(u, v) \in \mathbb{R}^2 \mid (u, -v) \in U\}$ and consider $\hat{\sigma} : \hat{U} \rightarrow S$ defined as $\hat{\sigma}(u, v) = \sigma(u, -v)$. Using this coordinate chart instead would reverse the sign of σ_v and hence also of N in Eq. 3.13. In summary, there are two possible unit normal vectors at each $p \in S$, but it's not yet clear whether there is any coordinate-free method to distinguish one of the choices as preferred.

Equation 3.13 defines a normal vector not just at one point, but at every $p \in V$. It's a whole field of vectors on V .

DEFINITION 3.43.

A **vector field** on a regular surface S means a smooth map $v : S \rightarrow \mathbb{R}^3$. It is called a **tangent field** if $v(p) \in T_p S$ for all $p \in S$. It is called a **normal field** if for all $p \in S$, $v(p)$ is a normal vector to S at p .

A vector field on S is a function from S to \mathbb{R}^3 , and you already know from Definition 3.12 (on page 121) what smoothness means in this context. But vector fields are visualized differently from other types of functions. Always picture $v(p)$ as a vector drawn with its tail at p .

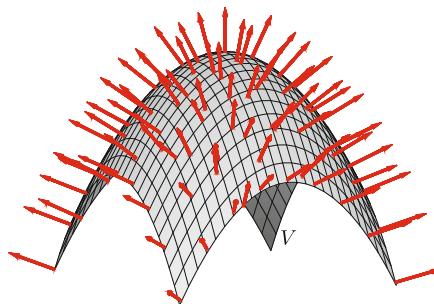


FIGURE 3.22. A unit normal field on V means a smoothly varying choice of a unit normal vector at each point of V

The vector field on V defined by Eq. 3.13 with respect to the coordinate chart σ is a **unit normal field** on V (which means a normal field whose vectors are all of unit length); see Fig. 3.22. But notice that V is only the image of one coordinate chart, which might not be all of S . Here is an important question: is it possible to find a *global* unit normal field on *all* of the surface S ? Surfaces for which this is possible are called *orientable*.

DEFINITION 3.44.

An **orientation** for a regular surface means a unit normal field on it. A regular surface S is called **orientable** if there exists an orientation for S . An orientable surface together with a given choice of an orientation is called an **oriented surface**. A surface patch σ for an oriented surface S is called **compatible with the orientation** if Eq. 3.13 agrees with the given orientation for S .

The following examples demonstrate that many familiar types of surfaces are orientable.

EXAMPLE 3.45 (A Plane Is Orientable). Let $x, y \in \mathbb{R}^3$ be linearly independent, and let S denote the two-dimensional subspace of \mathbb{R}^3 that they span. Notice that S is a regular surface covered by the single surface patch $\sigma : \mathbb{R}^2 \rightarrow S$ defined as $\sigma(u, v) = ux + vy$. Since $\sigma_u = x$ and $\sigma_v = y$ are constant, $T_p S = \text{span}\{x, y\} = S$ for all $p \in S$. Equation 3.13 determines the constant vector field $N(p) = \frac{x \times y}{|x \times y|}$, which is a global unit normal field on S .

EXAMPLE 3.46 (A Level Surface Is Orientable). If $\lambda \in \mathbb{R}$ is a regular value of the smooth function $f : \mathbb{R}^3 \rightarrow \mathbb{R}$, Theorem 3.27 (on page 133) implies that $S = f^{-1}(\lambda)$ is a regular surface. Our assumption that λ is a regular value means that the restriction to S of the gradient, $\nabla f(p) = \left(\frac{\partial f}{\partial x}(p), \frac{\partial f}{\partial y}(p), \frac{\partial f}{\partial z}(p) \right)$, is a nonvanishing vector field on S . Therefore, $N(p) = \frac{\nabla f(p)}{|\nabla f(p)|}$ is a global unit vector field on S . To verify that it is a normal field, we must check for every $p \in S$ that $\nabla f(p)$ is orthogonal to every vector $v \in T_p S$. For this, consider a regular curve γ in S with $\gamma(0) = p$ and $\gamma'(0) = v$. Since f is constant on S , we have $0 = (f \circ \gamma)'(0) = df_p(v) = \langle v, \nabla f(p) \rangle$.

EXAMPLE 3.47 (The Sphere Is Orientable). The sphere S^2 is the $f = 1$ level surface of the function $f(x, y, z) = x^2 + y^2 + z^2$. The gradient is $\nabla f(x, y, z) = (2x, 2y, 2z)$. In other words, $\nabla f(p) = 2p$. The strategy of the previous example yields the following global unit normal field on S^2 : $N(p) = \frac{\nabla f(p)}{|\nabla f(p)|} = p$. This field is outward-pointing. Alternatively, $N(p) = -p$ is an inward-pointing global unit normal field.

EXAMPLE 3.48 (A Graph Is Orientable). According to Lemma 3.17 (on page 123), if $U \subset \mathbb{R}^2$ is open and $f : U \rightarrow \mathbb{R}$ is smooth, then its graph $G = \{(x, y, f(x, y)) \mid (x, y) \in U\}$ is diffeomorphic to U via the diffeomorphism $\sigma : U \rightarrow G$ defined as $\sigma(x, y) = (x, y, f(x, y))$. In particular, G is a regular surface. Since G is covered by this single coordinate chart, the unit normal field on G determined by Eq. 3.13 is global. To find an explicit formula for this field at $p = (x, y, f(x, y))$, we compute as follows (with everything evaluated at $q = (x, y)$):

$$\sigma_x = (1, 0, f_x), \quad \sigma_y = (0, 1, f_y), \quad \sigma_x \times \sigma_y = (-f_x, -f_y, 1).$$

Therefore,

$$N = \frac{\sigma_x \times \sigma_y}{|\sigma_x \times \sigma_y|} = \frac{(-f_x, -f_y, 1)}{\sqrt{f_x^2 + f_y^2 + 1}}.$$

Every *compact* regular surface is orientable. This nontrivial fact follows from the following three-dimensional analogue of the Jordan curve theorem (Theorem 2.1 on page 62), whose proof is beyond the scope of this book:

THEOREM 3.49.

*If S is a compact regular surface, then $\mathbb{R}^3 - S = \{p \in \mathbb{R}^3 \mid p \notin S\}$ has exactly two path-connected components. Their common boundary is S . One component (which we call the **interior**) is bounded, while the other (which we call the **exterior**) is unbounded.*

Thus, a compact regular surface can be oriented by uniformly choosing either interior- or exterior-pointing unit normal vectors.

In general, choosing an orientation for an orientable surface isn't a very big choice—there are only two options in the following general situation:

LEMMA 3.50.

If S is a connected orientable regular surface, then there are exactly two orientations for S . Each is the negative of the other.

PROOF. Since S is orientable, there is at least one orientation, which we'll call $N : S \rightarrow \mathbb{R}^3$. Notice that $-N$ is another orientation. Let M be an arbitrary orientation on S . We must prove that M equals N or $-N$. For this, define $f : S \rightarrow \mathbb{R}$ such that for all $p \in S$,

$$f(p) = \langle N(p), M(p) \rangle.$$

At any particular point $p \in S$, either $M(p) = N(p)$ or $M(p) = -N(p)$, so either $f(p) = 1$ or $f(p) = -1$. Since S is connected, it follows from Proposition A.19 (on page 353 of the appendix) that f is constant on all of S , so M equals N or $-N$. \square

If you wish to paint the top of a surface, an orientation roughly means a choice of which side is considered the top. Think of each unit normal vector as a table umbrella shading the “top” side; see Fig. 3.23. With the other orientation, the umbrella would have shaded the other side, making us call this other side the “top.” There is no gravity or any other coordinate-free method for identifying a side that more deserves the title of “top”—it just depends on the choice of orientation.

The notion of an orientation of a surface can be related to the other contexts in which the term “orientation” has been used in this book. Specifically, an orientation for a surface means a smoothly varying choice of an orientation for each of its tangent planes (as in Definition 3.40 on page 149), which in turn determines whether a given basis of a given tangent plane $T_p S$ is positively oriented. With this distinction, one can specify a preferred orientation



FIGURE 3.23. An orientation of S roughly distinguishes its top side from its bottom side

for an arbitrary simple closed curve in $T_p S$; for this, just use a positively oriented basis to identify $T_p S$ with \mathbb{R}^2 , and select the “positive orientation” for the resulting plane curve (as defined in Sect. 1 of Chap. 2). This distinction also allows us to characterize the type of diffeomorphism that respects orientation:

DEFINITION 3.51.

A diffeomorphism $f : S \rightarrow \tilde{S}$ between oriented surfaces is called **orientation-preserving** if for each $p \in S$, $df_p : T_p S \rightarrow T_{f(p)} \tilde{S}$ is orientation-preserving (as in Definition 3.41 on page 150).

When the surfaces are connected, it is sufficient to check the test condition at a *single* point:

PROPOSITION 3.52.

A diffeomorphism $f : S \rightarrow \tilde{S}$ between connected oriented surfaces is orientation-preserving if and only if there exists a point $p \in S$ such that df_p is orientation-preserving.

PROOF. Exercise 3.60. □

EXAMPLE 3.53 (Rigid Motions Restricted to S^2). Let $A \in O(3)$ be an orthogonal matrix. Let $f : S^2 \rightarrow S^2$ denote the restriction to S^2 of the rigid motion $L_A : \mathbb{R}^3 \rightarrow \mathbb{R}^3$. Consider S^2 to have the “exterior-pointing” orientation, whose formula is $N(p) = p$. We claim that the diffeomorphism f is orientation-preserving if and only if $\det(A) = 1$ (equivalently, L_A is proper).

To see this, let $p \in S^2$, and let $\{v_1, v_2\}$ be a positively oriented ordered basis for $T_p S^2$. Notice that $\{v_1, v_2, N(p)\}$ is a positively oriented ordered basis for \mathbb{R}^3 (as in Definition 1.62 on page 53). According to Lemma 1.63 (on page 54), $\det(A) = 1$ if and only if the image under L_A of this ordered basis is itself a positively oriented ordered basis of \mathbb{R}^3 . This image is

$$\{L_A(v_1), L_A(v_2), L_A(N(p))\} = \{df_p(v_1), df_p(v_2), N(f(p))\}.$$

The above claim that $L_A(v_1) = df_p(v_1)$ is justified by noticing that df_p is the restriction to $T_p S^2$ of $d(L_A)_p$, and that $d(L_A)_p = L_A$ because the derivative of a linear transformation is itself (Exercise 3.1 on page 123). Now notice that $\{df_p(v_1), df_p(v_2), N(f(p))\}$ is a positively oriented ordered basis of \mathbb{R}^3 if

and only if $\{df_p(v_1), df_p(v_2)\}$ is a positively oriented ordered basis of $T_{f(p)}S^2$. In summary, $\det(A) = 1$ if and only if $df_p : T_p S^2 \rightarrow T_{f(p)}S^2$ is orientation-preserving.

The most infamous example of a nonorientable surface is the **Möbius strip** illustrated in Fig. 3.24. A model can be built with a thin strip of paper by gluing the ends together with twist. For it to be a surface, its “edge” (shown yellow in the right figure) must be excluded, but before you erase it, trace around it to confirm the interesting fact that there’s only one yellow edge, not two.



FIGURE 3.24. A Möbius strip is a nonorientable surface

If you try painting one side of a paper model, you will discover why it is nonorientable—it has only one side. Conveyor belts in engines, factories and supermarkets are sometimes twisted into the shape of a Möbius strip. Compared to the traditional two-sided belt, this design has a single side that is twice as long, which effectively doubles its lifespan.

To be more precise, let’s parametrize a particular Möbius strip S . For this, first let B denote the vertical line segment parametrized as $b(v) = (2, 0, v)$, $v \in (-1/2, 1/2)$, colored blue in Fig. 3.25. The surface S is formed when B spins around the z -axis (so its center traverses the red circle) while it simultaneously rotates so that it returns twisted. To return with exactly a half-twist, spinning an angle u must correspond to rotating an angle $u/2$. Imagine that B first rotates in place through an angle $u/2$ in the xz -plane about its center $(2, 0, 0)$; its new parametrization becomes $b_{u/2}(v) = (2, 0, 0) + v(\sin(u/2), 0, \cos(u/2))$ (shown in dashed-blue). It then gets left-multiplied by $A_u = \begin{pmatrix} \cos u & -\sin u & 0 \\ \sin u & \cos u & 0 \\ 0 & 0 & 1 \end{pmatrix}$ to achieve its spin by an angle u about the z -axis (as discussed in Example 3.25 on page 131). So S is parametrized as

$$(3.14) \quad \begin{aligned} \sigma(u, v) &= L_{A_u}(b_{u/2}(v)) \\ &= (\cos(u)(2 + v \sin(u/2)), \sin(u)(2 + v \sin(u/2)), v \cos(u/2)). \end{aligned}$$

Notice that the above formula for σ has period 4π with respect to the variable u . This might seem strange, because restricting σ to the domain where $u \in [0, 4\pi]$ and $v \in (-1/2, 1/2)$ covers S twice (just like when you painted around the paper model and ended up going around twice to paint

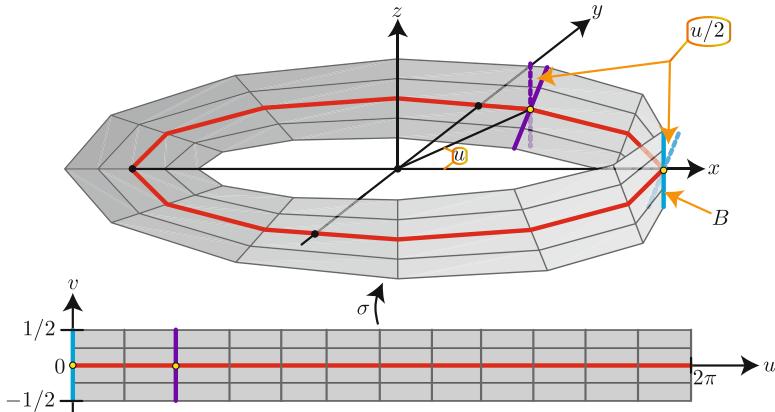


FIGURE 3.25. A parametrization of a Möbius strip

“both sides,” thereby discovering that there is really only one side). The restriction of σ to this domain is not one-to-one, so it is not a surface patch; however, σ does become a surface patch when it’s restricted to a domain where $v \in (-1/2, 1/2)$ and u is in any open interval of length less than 2π . This “double covering” feature is what allows us to prove the following:

PROPOSITION 3.54.

The Möbius strip S is not orientable.

PROOF. For $i \in \{1, 2\}$, let $\sigma_i : U_i \rightarrow V_i \subset S$ denote the restrictions of σ (from Eq. 3.14) to the following domains:

$$U_1 = \left\{ (u, v) \in \mathbb{R}^2 \mid \begin{array}{l} \pi/3 < u < 5\pi/3 \\ -1/2 < v < 1/2 \end{array} \right\}, \quad U_2 = \left\{ (u, v) \in \mathbb{R}^2 \mid \begin{array}{l} 4\pi/3 < u < 8\pi/3 \\ -1/2 < v < 1/2 \end{array} \right\}.$$

On these domains, the variable u is constrained to open intervals of length $4\pi/3$ centered at π and 2π respectively.

These two surface patches together cover S . Let N_1 and N_2 denote the unit normal fields induced by these surface patches on V_1 and V_2 via Eq. 3.13. The intersection, $V_1 \cap V_2$, equals the two disjoint strips denoted by A and B in Fig. 3.26 (colored copper and gold).

We will argue by contradiction. Suppose there exists a global unit normal field, N , on S . Since V_1 is connected, N and N_1 must either agree or differ by a sign on V_1 according to Lemma 3.50. We can assume without loss of generality that they agree (if not, change the sign of N). Since V_2 is connected, N and N_2 must either agree or differ by a sign on V_2 . Thus, the sign of N_2 can be changed if necessary to make it agree with N on V_2 , and hence also agree with N_1 on $A \cup B$. But it is straightforward to compute what is apparent in the figure: N_1 and N_2 agree on B but disagree on A , so

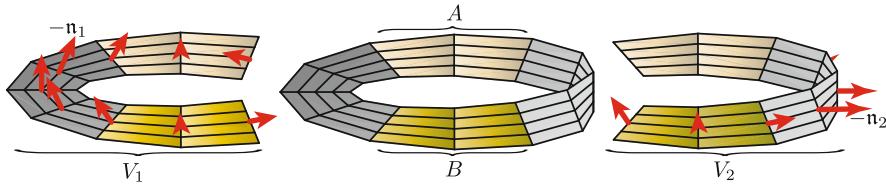


FIGURE 3.26. The Möbius strip is nonorientable because unit normal fields on V_1 and V_2 cannot be fit together \square

changing the sign of N_2 would not make it agree with N_1 on $A \cup B$. This contradiction proves that S does not admit a global unit normal field. \square

EXERCISES

EXERCISE 3.59. Let G be the graph of the function $f(x, y) = xy$. Let $p = (2, 3, 6) \in G$.

- (1) Find a basis for $T_p G$.
- (2) Find a unit normal vector to G at p .
- (3) Find a smooth function $h : \mathbb{R}^3 \rightarrow \mathbb{R}$ such that $G = h^{-1}(0)$. Compute the gradient of h at p , and use this as an alternative method to find a unit normal vector to G at p .

EXERCISE 3.60. Prove Proposition 3.52.

EXERCISE 3.61. Prove that a continuous unit normal field on a regular surface must be smooth.

EXERCISE 3.62. Consider the ellipsoid

$$E = \{(x, y, z) \in \mathbb{R}^3 \mid (x/a)^2 + (y/b)^2 + (z/c)^2 = 1\}$$

oriented by the outward-pointing unit normal field. Is the diffeomorphism $f : E \rightarrow E$ defined as $f(x, y, z) = (\frac{ay}{b}, \frac{bz}{c}, \frac{cx}{a})$ orientation-preserving or orientation-reversing? Assuming $a > b > c > 0$, at which points $p \in E$ is $\|df_p\|$ minimal? At which points is it maximal?

EXERCISE 3.63. Let $f : S \rightarrow \tilde{S}$ be a diffeomorphism between regular surfaces. Describe how an orientation for S naturally induces a well-defined orientation for \tilde{S} with respect to which f is orientation-preserving. Conclude that an orientable regular surface cannot be diffeomorphic to a nonorientable regular surface.

EXERCISE 3.64. Let S be an orientable regular surface, N an orientation for S , and f a rigid motion of \mathbb{R}^3 .

- (1) Show that the image $\tilde{S} = f(S)$ is a regular surface.
- (2) Define $\tilde{N} : \tilde{S} \rightarrow \mathbb{R}^3$ such that $\tilde{N}(f(p)) = df_p(N)$ for all $p \in S$. Show that \tilde{N} is an orientation of \tilde{S} .

- (3) Since the restriction of f to S is a diffeomorphism between S and \tilde{S} , the solution to Exercise 3.63 also provides an induced orientation for \tilde{S} ; under what conditions is it the same as \tilde{N} ?
- (4) Are any modifications needed in (1), (2), and (3) if f is instead a dilation, which means a function from \mathbb{R}^3 to \mathbb{R}^3 of the form $f(p) = \lambda p$ for some constant $\lambda > 0$?

EXERCISE 3.65. Let $f : S \rightarrow \tilde{S}$ be a diffeomorphism between oriented regular surfaces, with orientations denoted by N and \tilde{N} respectively. Let $p \in S$. Since f is smooth, there exist a neighborhood \mathcal{O} of p in \mathbb{R}^3 and a smooth function $F : \mathcal{O} \rightarrow \mathbb{R}^3$ that agrees with f on $\mathcal{O} \cap S$. If f is orientation-preserving, must $dF_p(N(p))$ and $\tilde{N}(f(p))$ point in the same direction, or at least must they form an acute angle? Or does it depend on the choice of F ?

EXERCISE 3.66. Construct an explicit parametrization of a “generalized Möbius strip” that returns with n half-twists. Under what condition on the integer n is the strip orientable? Under what condition on n_1, n_2 is the strip with n_1 half-twists diffeomorphic to the strip with n_2 half-twists?

EXERCISE 3.67 (Normal Neighborhood of a Surface). Let S be a compact regular surface. Let N be an orientation for S . For small $\epsilon > 0$, define $f : S \times (-\epsilon, \epsilon) \rightarrow \mathbb{R}^3$ as $f(p, t) = p + t \cdot N(p)$.

- (1) Prove that for sufficiently small ϵ , f is a diffeomorphism onto its image, which makes sense if you regard its domain as a subset of \mathbb{R}^4 . *HINT: Modify the hint for Exercise 3.11 on page 124.*
- (2) Let $\epsilon > 0$ be as in part (1). Prove that for every $r \in (-\epsilon, \epsilon)$, the set $S_r = \{f(p, r) \mid p \in S\}$ is a regular surface diffeomorphic to S , and that $S_r \cup S_{-r} = \{p \in \mathbb{R}^3 \mid \text{dist}(p, S) = r\}$.

EXERCISE 3.68. Let S be an orientable regular surface and let N be an orientation for S . Let $\varphi : S \rightarrow \mathbb{R}$ be a smooth function. Assume that φ has *compact support*, which means that there exists a compact subset $K \subset S$ such that φ equals zero at every point of $S - K$. For $t \in \mathbb{R}$, define

$$S_t = \{p + t\varphi(p)N(p) \mid p \in S\}.$$

Prove that for sufficiently small t , S_t is a regular surface.

EXERCISE 3.69. Let S be a regular surface, $p \in S$, $v \in T_p S$ a nonzero vector, and N a unit normal vector to S at p . Prove that there exists a neighborhood, V , of p in S such that the intersection of V with the plane $p + \text{span}\{v, N\}$ is the trace of a regular curve.

EXERCISE 3.70. Let S be a connected oriented regular surface. Suppose there exists a line $L \subset \mathbb{R}^3$ that intersects all normal lines to S (that is, for every $p \in S$, the trace of the normal line $t \mapsto p + tN(p)$ intersects L). Prove that S is a portion of a surface of revolution.

EXERCISE 3.71. Let \mathcal{P} denote the set of all two-dimensional subspaces of \mathbb{R}^3 . We saw in Example 3.45 that every $P \in \mathcal{P}$ is oriented by a single

vector, namely a unit normal vector to P . But the unit normal vector computed in that example depended on the choice of basis of P —a different basis choice might change its sign. In this problem, we will verify that there is no basis-independent method for choosing a preferred unit normal vector (or a preferred basis) for all members of \mathcal{P} .

- (1) Prove there does not exist a continuous map $\mathcal{P} \rightarrow S^2$ that associates to each plane a unit normal vector to that plane. For the term “continuous” to make sense here (via Definition A.13 on page 350 of the appendix), we will say that a sequence of subspaces converges to a limit subspace if they have bases that converge to a basis of the limit. *HINT: Such a map could orient the Möbius strip.*
- (2) Precisely state and prove the assertion that there does not exist a continuous method for associating a basis to each $P \in \mathcal{P}$.

This problem helps explain why linear algebra texts emphasize basis-independent measurements.

EXERCISE 3.72. Let S be a regular surface and $f : S \rightarrow \mathbb{R}$ a smooth function. Prove that there exists a unique tangent field on S , denoted by ∇f and called the **gradient** of f , with the following property:

$$df_p(X) = \langle (\nabla f)(p), X \rangle \quad \text{for all } p \in S \text{ and all } X \in T_p S.$$

Notice that this generalizes the definition in Sect. 4 (Chap. 2) of the gradient of a vector field on \mathbb{R}^2 .

EXERCISE 3.73. Redo the proof of Proposition 3.29 (on page 135) with the single vector N replaced by a unit normal field on a portion of S .



6. Surface Area

In this section we define and study the area of a region on a surface. Suppose that $\sigma : U \subset \mathbb{R}^2 \rightarrow S$ is a surface patch. Let $q \in U$ and define $p = \sigma(q)$. By choosing the basis $\{e_1, e_2\}$ in Definition 3.37 on page 147, we learn that the area distortion of the linear transformation $d\sigma_q : \mathbb{R}^2 \rightarrow T_p S$ is

$$\|d\sigma_q\| = |\sigma_u(q) \times \sigma_v(q)|.$$

This area distortion factor depends on $q \in U$, so we think of $\|d\sigma\|$ as a function from U to the positive reals.

To avoid technical issues with integration, we'll consider only integrals of smooth functions over certain types of regions.

DEFINITION 3.55.

Let S be a regular surface and let $R \subset S$. We call R a **polygonal region** if R is covered by a single coordinate chart $\sigma : U \subset \mathbb{R}^2 \rightarrow S$, and $\sigma^{-1}(R)$ equals the interior plus boundary of a piecewise-regular simple closed curve in \mathbb{R}^2 . In this case:

- (1) We define the **area** (also called the **surface area**) of R as

$$\text{Area}(R) = \iint_{\sigma^{-1}(R)} \|d\sigma\| dA.$$

- (2) We define the integral over R of a smooth function $f : R \rightarrow \mathbb{R}$ as

$$\iint_R f dA = \iint_{\sigma^{-1}(R)} (f \circ \sigma) \cdot \|d\sigma\| dA.$$

If R is the union of finitely many polygonal regions intersecting only along boundaries, then we define the area of R (or the integral of f over R) as the sum over all of these polygonal regions of the area (or the integral).

The value $\|d\sigma_q\|$ is the area distortion of the derivative of σ at $q \in U$, but it could alternatively be called the **infinitesimal area distortion** of σ itself at q , because if $\sigma^{-1}(R) \ni q$ is small enough that $\|d\sigma\|$ is approximately constant over it, then the above definition gives

$$\|d\sigma_q\| \approx \frac{\text{Area}(R)}{\text{Area}(\sigma^{-1}(R))}.$$

Why is $\iint_{\sigma^{-1}(R)} \|d\sigma\| dA$ a reasonable definition of the area of R ? In other words, why should the area of R be defined as the integral over $\sigma^{-1}(R)$ of the infinitesimal area distortion of σ ? Because one term of a Riemann sum approximation of this integral (in rectangular coordinates) looks like

$$\begin{aligned} \|d\sigma\| \Delta u \Delta v &= |\sigma_u \times \sigma_v| \Delta u \Delta v = |(\Delta u \sigma_u) \times (\Delta v \sigma_v)| \\ &= \text{area of parallelogram spanned by } (\Delta u \sigma_u) \text{ and } (\Delta v \sigma_v), \end{aligned}$$

where σ_u and σ_v are evaluated at a sample point (u_i, v_i) from the i th rectangle \mathcal{R}_i into which the domain $\sigma^{-1}(R)$ has been subdivided. As depicted in Fig. 3.27, this parallelogram approximates $\sigma(\mathcal{R}_i)$, so the whole Riemann sum approximates the area of all of R .

Why is $\iint_{\sigma^{-1}(R)} (f \circ \sigma) \cdot \|d\sigma\| dA$ a reasonable definition of the integral of f over R ? Imagine that S is the Earth, R is England, and f is a population density function representing the number of insects per unit area. One term of a Riemann sum approximation of this integral equals the approximate area of $\sigma(\mathcal{R}_i)$ times the population density of insects at a sample point in $\sigma(\mathcal{R}_i)$, so this term approximates the number of insects living in $\sigma(\mathcal{R}_i)$. The whole Riemann sum therefore approximates the total population of English insects,

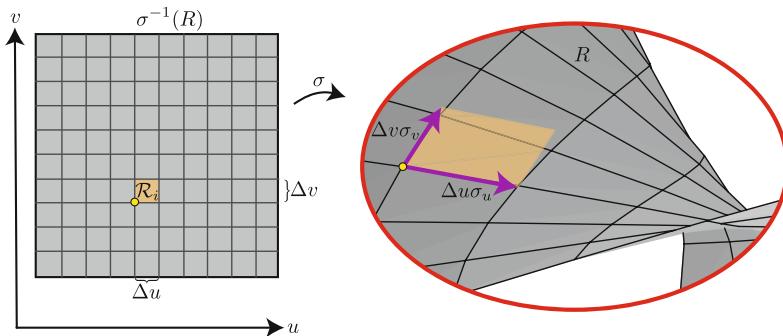


FIGURE 3.27. $|(\Delta u \sigma_u) \times (\Delta v \sigma_v)|$ approximates the area of $\sigma(R_i)$

exactly what $\int_R f dA$ should represent. Since definitions don't require proofs, the above discussion is intended only to make the definition seem reasonable. However, the following is required:

PROPOSITION 3.56.

Definition 3.55 is well defined. In other words, if there are two surface patches $\sigma_i : U_i \rightarrow V_i$ ($i \in \{1, 2\}$), each of which covers R , then $\iint_R f dA$ has the same value when computed with respect to either surface patch.

The proof relies on the following “change of coordinates” rule for double integration, found in any analysis or multivariable calculus book:

LEMMA 3.57.

If $\Psi : U_1 \rightarrow U_2$ is a diffeomorphism between open sets in \mathbb{R}^2 , $\varphi : U_2 \rightarrow \mathbb{R}$ is smooth, and $K \subset U_2$ is a polygonal region, then

$$\iint_K \varphi dA = \iint_{\Psi^{-1}(K)} (\varphi \circ \Psi) \cdot \|d\Psi\| dA.$$

PROOF OF PROPOSITION 3.56. Consider the **transition map** $\psi = \sigma_2^{-1} \circ \sigma_1$. Technically, ψ is a diffeomorphism from $\sigma_1^{-1}(V_1 \cap V_2)$ to $\sigma_2^{-1}(V_1 \cap V_2)$. However, we lose no generality if we assume that $V_1 = V_2$, so that ψ becomes a diffeomorphism from U_1 to U_2 . To achieve this simplification, just redefine U_1 to equal $\sigma_1^{-1}(V_1 \cap V_2)$ and redefine U_2 to equal $\sigma_2^{-1}(V_1 \cap V_2)$; see Fig. 3.28.

Since $\sigma_1 = \sigma_2 \circ \psi$, the chain rule (Proposition 3.9 on page 119) and Exercise 3.56 (on page 151) give

$$\|d(\sigma_1)\| = \|d(\sigma_2) \circ d\psi\| = \|d(\sigma_2)\| \cdot \|d\psi\|.$$

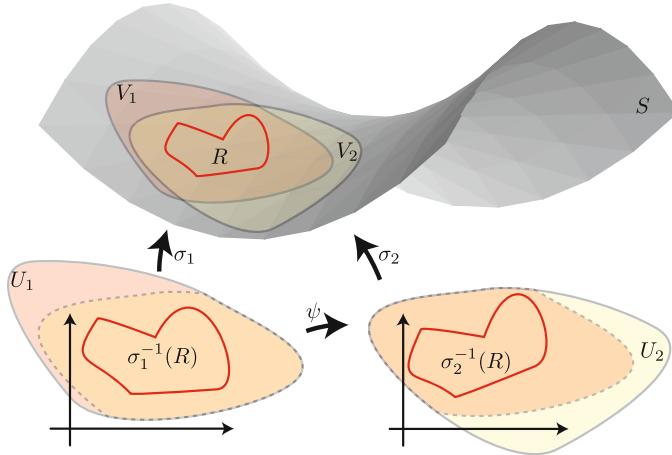


FIGURE 3.28. After shrinking U_1 and U_2 to the portions thereof shaded orange, we can assume that $V_1 = V_2$, so that $\psi : U_1 \rightarrow U_2$

The integral of f over R computed with respect to σ_1 becomes

$$\begin{aligned} \int_R f \, dA &= \iint_{\sigma_1^{-1}(R)} (f \circ \sigma_1) \cdot \|d(\sigma_1)\| \, dA \\ &= \iint_{\sigma_1^{-1}(R)} (f \circ \sigma_2 \circ \psi) \cdot \|d(\sigma_2)\| \cdot \|d\psi\| \, dA \\ &= \iint_{\sigma_2^{-1}(R)} (f \circ \sigma_2) \cdot \|d(\sigma_2)\| \, dA, \end{aligned}$$

which equals the integral of f over R computed with respect to σ_2 . The final equality is an application of Lemma 3.57. \square

EXAMPLE 3.58 (The Area of a Graph). *The graph G of a smooth function $f : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}$ is covered by the single surface patch $\sigma : U \rightarrow G$ defined as $\sigma(x, y) = (x, y, f(x, y))$. We computed in Example 3.48 on page 153 that*

$$|\sigma_x \times \sigma_y| = \sqrt{f_x^2 + f_y^2 + 1}.$$

Therefore, the area of a polygonal region $R \subset G$ equals

$$\text{area}(R) = \iint_{\sigma^{-1}(R)} \sqrt{f_x^2 + f_y^2 + 1} \, dA.$$

EXAMPLE 3.59 (The Area of the Sphere). *Recall from Example 3.24 on page 129 that (all but an arc of) the sphere S^2 is covered by the spherical coordinate chart $\sigma : \underbrace{(0, 2\pi)}_U \times (0, \pi) \rightarrow S^2$ defined as*

$$\sigma(\theta, \phi) = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi).$$

We computed in that example that $\|d\sigma\| = |\sigma_\theta \times \sigma_\phi| = \sin \phi$. Since the excluded arc has no area (it is one-dimensional), the surface area of (all of) S^2 equals

$$\iint_U (\sin \phi) dA = \int_0^{2\pi} \int_0^\pi (\sin \phi) d\phi d\theta = 2\pi \int_0^\pi (\sin \phi) d\phi = 4\pi.$$

We cheated a bit by treating the boundary of U as if it were part of U . We really should have computed $\int_{0+\epsilon}^{2\pi-\epsilon} \int_{0+\epsilon}^{\pi-\epsilon} (\sin \phi) d\phi d\theta$ for small $\epsilon > 0$ and taken the limit as $\epsilon \rightarrow 0$, but this clearly gives the same answer.

It is instructive to look back at Fig. 3.11 on page 130 for a visual understanding of the above formula. The area distortion factor $\|d\sigma\| = \sin \phi$ approaches zero at the north and south poles, where the coordinate grid lines appear to become more and more crowded together.

Although we have thus far mentioned only the area distortion of the derivative of a surface patch, the area distortions of the derivatives of other types of functions are also important, such as a diffeomorphism between surfaces:

PROPOSITION 3.60.

If $f : S_1 \rightarrow S_2$ is a diffeomorphism between regular surfaces and R is a polygonal region of S_1 , then its image $f(R)$ is a polygonal region of S_2 , and

$$\text{Area}(f(R)) = \iint_R \|df\| dA.$$

PROOF. Let $\sigma : U \subset \mathbb{R}^2 \rightarrow S_1$ be a surface patch that covers R . Then $f(R)$ is covered by the surface patch $f \circ \sigma : U \rightarrow S_2$. By the chain rule for surfaces (Exercise 3.46 on page 146) and Exercise 3.56 (on page 151),

$$\|d(f \circ \sigma)\| = \|df \circ d\sigma\| = \|df\| \cdot \|d\sigma\|.$$

Thus,

$$\text{Area}(f(R)) = \iint_{\sigma^{-1}(R)} \|d(f \circ \sigma)\| dA = \iint_{\sigma^{-1}(R)} \|df\| \cdot \|d\sigma\| dA = \iint_R \|df\| dA.$$

□

For $p \in S$, the value $\|df_p\|$ is the area distortion of the derivative of f at p , but it could alternatively be called the **infinitesimal area distortion** of f itself at p , because if $R \ni p$ is small enough that $\|df\|$ is approximately constant over it, then the above proposition gives

$$(3.15) \quad \|df_p\| \approx \frac{\text{Area}(f(R))}{\text{Area}(R)}.$$

EXERCISES

EXERCISE 3.74. Let S be the surface of revolution obtained when the graph of $y = e^{-x}$, $x \in (0, \infty)$, is revolved about the x -axis. Use an improper integral to decide whether S has finite or infinite surface area.

EXERCISE 3.75. For fixed $\phi_0 \in (0, \pi)$, derive the following formula for the area of the **spherical cap** consisting of all points of S^2 whose spherical coordinates (θ, ϕ) (as in Example 3.24 on page 129) satisfy $\phi \leq \phi_0$:

$$\text{Area} = 2\pi(1 - \cos \phi_0).$$

EXERCISE 3.76. Write a general formula for the area of a surface of revolution (of the type described in Example 3.25 on page 131). What is the area of the torus obtained by revolving about the z -axis a circle of radius 1 in the xz -plane centered at $(2, 0, 0)$ (pictured in Fig. 3.16 on page 138)?

EXERCISE 3.77. What is the area of the Möbius strip (defined in Eq. 3.14 on page 156)?

EXERCISE 3.78. Let S be a regular surface. Every pair of overlapping coordinate charts from a given atlas will determine a transition function Ψ , defined as in the proof of Proposition 3.56. Prove that S is orientable if and only if there exists an atlas for S such for each transition function Ψ satisfies $\det(d\Psi_q) > 0$ for all q in its domain.

EXERCISE 3.79. Combine Lemma 3.57 with Example 3.39 (on page 148) to describe and justify the method (familiar from multivariable calculus) of evaluating a double integral in polar coordinates.

7. Isometries and the First Fundamental Form

In this section, we define the first fundamental form of a surface, and we begin to explore the question of which measurements can be defined purely in terms of it.

Recall from Sect. 2 of Chap. 1 that the **inner product** of a pair of vectors $x = (x_1, x_2, \dots, x_n), y = (y_1, y_2, \dots, y_n) \in \mathbb{R}^n$ is

$$\langle x, y \rangle = x_1 y_1 + x_2 y_2 + \cdots + x_n y_n \in \mathbb{R}.$$

Algebraically, the inner product is a symmetric bilinear function on \mathbb{R}^n . Geometrically, it is possible to define norms and angles purely in terms of the inner product:

$$(3.16) \quad |x|^2 = \langle x, x \rangle, \quad \angle(x, y) = \cos^{-1} \left(\frac{\langle x, y \rangle}{|x||y|} \right).$$

Conversely, the inner product can be expressed purely in terms of norms. To see this, solve the equation

$$|x - y|^2 = \langle x - y, x - y \rangle = \langle x, x \rangle + \langle y, y \rangle - 2 \langle x, y \rangle$$

for $\langle x, y \rangle$, to learn that

$$(3.17) \quad \langle x, y \rangle = \frac{1}{2} (|x|^2 + |y|^2 - |x - y|^2).$$

Thus, the inner product and the norm contain exactly the same information, just packaged in different ways. Equation 3.17 was mentioned previously in Sect. 8 of Chap. 1.

Suppose that S is a regular surface and $p \in S$. We sometimes write $\langle x, y \rangle_p$ (instead of just $\langle x, y \rangle$) when $x, y \in T_p S$. In other words, $\langle \cdot, \cdot \rangle_p : T_p S \times T_p S \rightarrow \mathbb{R}$ denotes the restriction of the inner product in \mathbb{R}^3 to pairs of vectors in $T_p S$. Similarly, we will sometimes write $|x|_p$ (instead of just $|x|$) when $x \in T_p S$. In other words, $|\cdot|_p : T_p S \rightarrow \mathbb{R}$ denotes the restriction of the norm in \mathbb{R}^3 to vectors in $T_p S$. The composition of this restriction with the squaring function has an important-sounding name:

DEFINITION 3.61.

The **first fundamental form** of S assigns to each $p \in S$ the restriction to $T_p S$ of the squared norm function in \mathbb{R}^3 , i.e., the map from $T_p S$ to \mathbb{R} defined as $x \mapsto |x|_p^2$.

The term **form** can be precisely defined on its own, but we will not require this. In vague terms, a *form* on S associates to each $p \in S$ a function that inputs one or more vectors in $T_p S$. In the case of the first fundamental form, this function inputs one vector and outputs its squared norm.

This seems a grand title for such a simple definition. What's really fundamental is the following question: *What measurements on S can be described purely in terms of the first fundamental form?* A measurement on a regular surface is called **intrinsic** if it can be described purely in terms of the first fundamental form. This is an important distinction. Throughout the book, we will repeatedly ask whether newly defined measurements are intrinsic.

For now, it's clear from Eq. 3.17 that the inner product (of a pair of vectors tangent to S at an arbitrary point) is intrinsic. Further, Eq. 3.16 allows one to describe angles (between pairs of vectors tangent to S at an arbitrary point) purely in terms of the first fundamental form, so angles are intrinsic.

A less obvious example is that the area of a parallelogram (spanned a pair of vectors x, y that are tangent to S at a point) is intrinsic. This area equals $|x \times y|$, which at first glance seems to depend on more than just the first fundamental form—it involves the cross product and also the norm of a vector that is not tangent to S . But Lemma 1.43 (on page 41) allows us to reexpress the formula as

$$|x \times y| = \sqrt{|x|^2 |y|^2 - \langle x, y \rangle^2},$$

which depends only on the first fundamental form.

There is a more precise way to define “intrinsic.” What we’re really looking for are measurements that are preserved by first-fundamental-form-preserving diffeomorphisms between surfaces. Such diffeomorphisms are called *isometries*:

Let S_1, S_2 be a pair of regular surfaces. A diffeomorphism $f : S_1 \rightarrow S_2$ is called an **isometry** if df preserves their first fundamental forms, that is, if

$$|x|_p^2 = |df_p(x)|_{f(p)}^2 \quad \text{for all } p \in S_1 \text{ and all } x \in T_p S_1.$$

This is equivalent to saying that df preserves their inner products:

$$\langle x, y \rangle_p = \langle df_p(x), df_p(y) \rangle_{f(p)} \quad \text{for all } p \in S_1 \text{ and all } x, y \in T_p S_1.$$

Two regular surfaces are called **isometric** if there exists an isometry between them. A measurement or construction on surfaces is called **intrinsic** if it is preserved by isometries.

The equivalence is an immediate consequence of Eq. 3.17, which expresses the inner product in terms of the norm. This definition of “intrinsic” is consistent with the previous, but more precise. For example, the above claim that “the area of a parallelogram is intrinsic” really means this: if $f : S_1 \rightarrow S_2$ is an isometry, then $|df_p(x) \times df_p(y)| = |x \times y|$ for all $p \in S_1$ and all $x, y \in T_p S_1$.

The main source of examples of isometries is the following:

LEMMA 3.62.

If f is a rigid motion of \mathbb{R}^3 and S is a regular surface, then the image $f(S)$ is also a regular surface, and f restricts to an isometry from S to $f(S)$.

PROOF. Exercise 3.80. □

In this lemma, if $f(S) = S$, then f induces an isometry from S to itself, called an **isometry of S** . For example, every rotation about the z -axis is an isometry of an arbitrary surface of revolution (of the type constructed in Example 3.25 on page 131). Similarly, if $A \in O(3)$ is an orthogonal matrix, then L_A restricts to an isometry of the sphere S^2 .

The next example exhibits an isometry between a pair of surfaces that does not come from a rigid motions of \mathbb{R}^3 .

EXAMPLE 3.63 (The Cylinder Surface Patch Is an Isometry). In Example 3.23 on page 128, we covered all but a line of the cylinder, C , by the surface patch $\sigma : \underbrace{(-\pi, \pi)}_U \times \mathbb{R} \rightarrow C$ defined as $\sigma(u, v) = (\cos(u), \sin(u), v)$. At an arbitrary point $q = (u, v) \in U$, we computed that

$$\sigma_u(q) = (-\sin(u), \cos(u), 0), \quad \sigma_v(q) = (0, 0, 1).$$

These two vectors form an orthonormal basis of the tangent plane $T_{\sigma(q)} C$. In other words, for each $q \in U$, we have identified an orthonormal basis of

$T_q U = \mathbb{R}^2$ (namely $\{e_1, e_2\}$) that the linear transformation $d\sigma_q$ sends to an orthonormal basis of $T_{\sigma(q)} C$. This implies that σ is an isometry between U and its image.

An equivalent definition of *isometry* is a diffeomorphism that preserves the lengths of all curves (Exercise 3.81). The above surface patch for the cylinder can be visualized as the bending of a flat sheet of metal around an axis. Intuitively, the bending does not change the length of any curve drawn on the flat metal sheet, so it is visually believable that it is an isometry.

This cylinder example improves our developing intuition about which measurements are intrinsic. Do isometries preserve measurements of how a surface bends and curves within the ambient \mathbb{R}^3 ? We will later return to this important question. In fact, we will later prove that a fundamentally important measurement called the *Gaussian curvature* of a surface is intrinsic. But for now, we mention a more obvious measurement of bending/curving that is not:

EXAMPLE 3.64 (Minimal Curvature Is Not Intrinsic). Let S be a regular surface, let $p \in S$, and let $v \in T_p S$ be of unit length. The “minimal curvature” of S at p in the direction v means the minimum possible curvature at p of any regular curve γ in S with $\gamma(0) = p$ and $\gamma'(0) = v$. For example, the minimal curvature of an open subset of \mathbb{R}^2 at any point in any direction equals zero, because γ can be chosen as a straight line segment. On the other hand, Exercise 1.41 on page 31 implies that the cylinder has nonzero minimal curvatures—every curve tangent to the waist of the cylinder must have curvature of at least 1 in order to bend sharply enough to remain in the cylinder. Therefore, Example 3.63 implies that minimal curvature is not preserved by isometries. Although the term “minimal curvature” will not appear anywhere else in this book, Chap. 4 will discuss “normal curvature,” which is really just a signed version of minimal curvature.

There are even isometries between *compact* surfaces that do not come from rigid motions of \mathbb{R}^3 :

EXAMPLE 3.65 (An Isometry Between Compact Surfaces That Does Not Come from a Rigid Motion). Consider the surface of revolution obtained when the red-plus-yellow generating curve in Fig. 3.29 is rotated about the vertical axis shown. Two versions of the generating curve are illustrated—they differ in whether the yellow portion is flipped before being attached to the red portion. To ensure smoothness of both surfaces, the red and yellow curves are horizontal in neighborhoods of their attachment point and their intersections with the axis, as the figure illustrates. The resultant pair of surfaces of revolution have an obvious isometry between them. The obvious map is an isometry, because it’s the identity map between their identical red portions, while on their yellow portions it comes from a reflection across a plane (which is a rigid motion of \mathbb{R}^3).

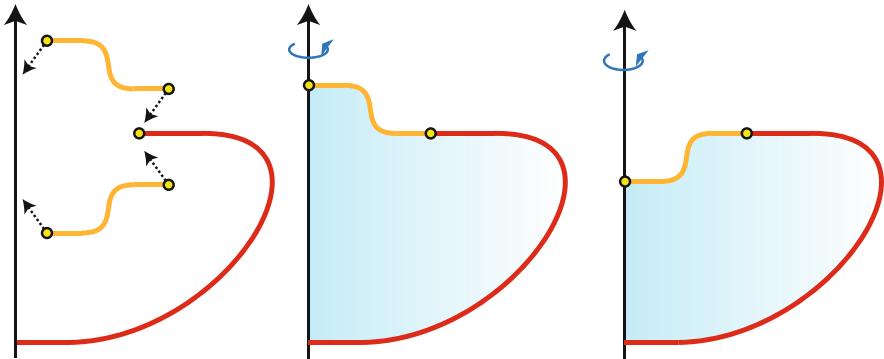


FIGURE 3.29. The volume inside the surface of revolution depends on whether the yellow portion of the generating curve is flipped before being attached to the red portion

You might have incorrectly guessed that “volume of interior” is isometry-invariant (at least among compact regular surfaces, where “interior” makes sense via Theorem 3.49 on page 154). But the pair of surfaces in the above example are isometric, yet they bound different volumes. Thus, “volume of interior” is *not* an intrinsic measurement.

EXERCISES

EXERCISE 3.80. Prove Lemma 3.62.

EXERCISE 3.81. Let $f : S_1 \rightarrow S_2$ be a diffeomorphism between regular surfaces. Prove that f is an isometry if and only if for every regular curve $\gamma : [a, b] \rightarrow S_1$, the length of γ equals the length of $f \circ \gamma$.

EXERCISE 3.82. A function $f : S_1 \rightarrow S_2$ between regular surfaces is called a **local diffeomorphism** if for all $p \in S_1$, there exists a neighborhood U of p in S_1 such that the restriction of f to U is a diffeomorphism onto its image $f(U)$. If additionally $\langle df_p(x), df_p(y) \rangle_{f(p)} = \langle x, y \rangle_p$ for all $p \in S_1$ and all $x, y \in T_p S_1$, then f is called a **local isometry**.

- (1) Show that the map from \mathbb{R}^2 to the cylinder $C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}$ defined as $f(x, y) = (\cos x, \sin x, y)$ is a local isometry.
It is not a (global) diffeomorphism because it is not injective.

- (2) Construct a local diffeomorphism from the cylinder

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1 \text{ and } -1/2 < z < 1/2\}$$

to the Möbius strip defined in Sect. 5. Does there exist a (global) diffeomorphism between these surfaces?

EXERCISE 3.83. Show that every generalized cylinder (Exercise 3.18(1) on page 136) is isometric to a standard cylinder of the form

$$\{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = R^2\}$$

for some $R > 0$.

EXERCISE 3.84. Show that every point of a generalized cone (Exercise 3.19(1) on page 137) is covered by a surface patch that is an isometry.

EXERCISE 3.85. Let S be the graph of the equation $z = xy$. Classify the linear rigid motions of \mathbb{R}^3 that induce isometries of S (that is, the ones that map S onto itself).

□

8. Equiareal and Conformal Maps (Optional)

This optional section is about two types of diffeomorphisms that are more general than isometries, namely, equiareal and conformal maps. The section is grounded in its application to a fundamental problem of cartography: what is the best method for constructing a flat map of the spherical Earth?

Here is the first important way to weaken the definition of isometry:

DEFINITION 3.66.

Let S_1, S_2 be a pair of regular surfaces. A diffeomorphism $f : S_1 \rightarrow S_2$ is called **equiareal** if for all $p \in S_1$, the area distortion $\|df_p\|$ equals 1.

A synonym for “equiareal” is “area-preserving,” because of the following result:

PROPOSITION 3.67.

A diffeomorphism $f : S_1 \rightarrow S_2$ is equiareal if and only if it is area-preserving in the following sense: if $R \subset S_1$ is any polygonal region, then $\text{Area}(R) = \text{Area}(f(R))$.

PROOF. One direction follows immediately from Proposition 3.60, and the other is left to the reader in Exercise 3.86. □

Isometries are equiareal, but not all equiareal maps are isometries, as the following example shows:

EXAMPLE 3.68. Let $\lambda > 0$. Let $f = L_A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denote the linear transformation represented by the matrix $A = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ with respect to the standard orthonormal basis $\{e_1, e_2\}$. For every $p \in \mathbb{R}^2$, we have $df_p = f$ (by Exercise 3.1 on page 123). Therefore

$$\|df_p\| = \|f\| = |\det(A)| = 1,$$

which shows that f is equiareal. Figure 3.30 illustrates the effect of f (with $\lambda = 2$) on a circle and on some angles. Although f maintains the 90° angle

between the horizontal and vertical directions, it increases the purple–green angle and decreases the green–red angle. Thus, not all angles are preserved.

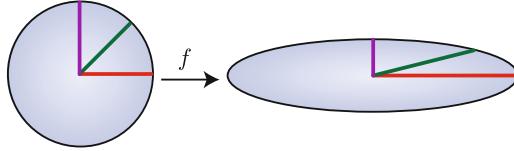


FIGURE 3.30. A linear equiareal map with distortion factor $\lambda = 2$

The above example is very special, because f is linear, but it illustrates the infinitesimal picture for a general equiareal map. That is, if $f : S_1 \rightarrow S_2$ is any equiareal map, then for every $p \in S_1$, orthonormal bases of $T_p S_1$ and $T_{f(p)} S_2$ can be chosen such that the matrix representing the linear transformation $df_p : T_p S_1 \rightarrow T_{f(p)} S_2$ has the form $A = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ for some $\lambda > 0$ (Exercise 3.88). Thus, an equiareal map avoids distorting infinitesimal area through the balancing act of expanding one direction while compressing the other to compensate. The expansion–contraction factor λ is a function of p .

The most famous example of an equiareal map comes from Archimedes. The domain is the sphere with the north and south poles removed: $S^2 - \{(0, 0, \pm 1)\}$. The range is the following cylinder:

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1 \text{ and } -1 < z < 1\}.$$

The function $f : S^2 - \{(0, 0, \pm 1)\} \rightarrow C$ is pictured in Fig. 3.31. It sends each $p \in S^2 - \{(0, 0, \pm 1)\}$ to the point $f(p) \in C$ that is nearest to p . This nearest point will have the same z -value as p . If you imagine that the z -axis emits light, and the sphere is made of glass with a speck of paint at p , then $f(p)$ is the speck's shadow on the cylinder. A formula for this shadowing of the sphere onto the cylinder is

$$(3.18) \quad f(x, y, z) = \left(\frac{x}{\sqrt{x^2 + y^2}}, \frac{y}{\sqrt{x^2 + y^2}}, z \right).$$

THEOREM 3.69 (Archimedes's Theorem).

The function f in Eq. 3.18 is equiareal.

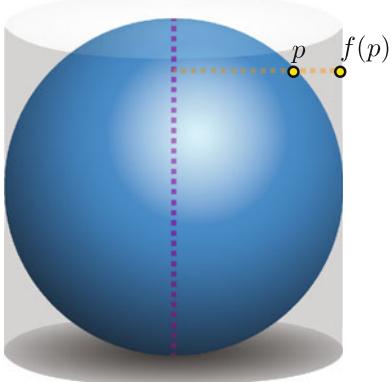


FIGURE 3.31. An equiareal map from the sphere to the cylinder

Archimedes used this theorem to discover the area of a sphere (it equals the area of the cylinder, which was previously known). According to legend, Archimedes asked that an image of his proof be inscribed on his tombstone. Since he predated the invention of calculus, his proof was much more involved than this one:

PROOF. The spherical coordinate chart $\sigma : \underbrace{(0, 2\pi) \times (0, \pi)}_U \rightarrow S^2$ was defined in Example 3.24 (on page 129) as

$$\sigma(\theta, \phi) = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi).$$

Let $q = (\theta, \phi) \in U$ and let $p = \sigma(q)$. We computed in that example that $|\sigma_\theta \times \sigma_\phi| = \sin \phi$. Since $\{\sigma_\theta, \sigma_\phi\}$ is a basis of $T_p S^2$, the definition of area distortion gives

$$\|df_p\| = \frac{|df_p(\sigma_\theta) \times df_p(\sigma_\phi)|}{|\sigma_\theta \times \sigma_\phi|} = \frac{|df_p(\sigma_\theta) \times df_p(\sigma_\phi)|}{\sin \phi}.$$

It remains to demonstrate that $|df_p(\sigma_\theta) \times df_p(\sigma_\phi)| = \sin \phi$, so that $\|df_p\| = 1$, as desired.

The composition $\tilde{\sigma} = f \circ \sigma : U \rightarrow C$ is a surface patch for C , whose formula simplifies to

$$\tilde{\sigma}(\theta, \phi) = (\cos(\theta), \sin(\theta), \cos \phi).$$

Since $df_p(\sigma_\theta) = \tilde{\sigma}_\theta$ and $df_p(\sigma_\phi) = \tilde{\sigma}_\phi$, we learn that

$|df_p(\sigma_\theta) \times df_p(\sigma_\phi)| = |\tilde{\sigma}_\theta \times \tilde{\sigma}_\phi| = |(-\sin \theta, \cos \theta, 0) \times (0, 0, -\sin \phi)| = \sin \phi$, which completes the proof. \square

From the formulas in the previous proof, you can verify what is apparent in Fig. 3.32; namely, $\|df_p\| = 1$, because an expansion of length in the latitudinal (σ_θ) direction is balanced by a contraction of length in the longitudinal (σ_ϕ) direction. In fact, with respect to the orthonormal bases $\left\{ \frac{\sigma_\theta}{|\sigma_\theta|}, \frac{\sigma_\phi}{|\sigma_\phi|} \right\}$ and $\left\{ \frac{\tilde{\sigma}_\theta}{|\tilde{\sigma}_\theta|}, \frac{\tilde{\sigma}_\phi}{|\tilde{\sigma}_\phi|} \right\}$, the linear transformation df_p is represented by the matrix $\begin{pmatrix} 1/\sin \phi & 0 \\ 0 & \sin \phi \end{pmatrix}$; compare to Example 3.68. The expansion–contraction factor, $\lambda = 1/\sin \phi$, equals 1 along the equator and approaches infinity as p approaches the poles.

Archimedes's theorem can be used to create a flat rectangular map of the spherical Earth with desirable properties. Specifically, the **Lambert projection**

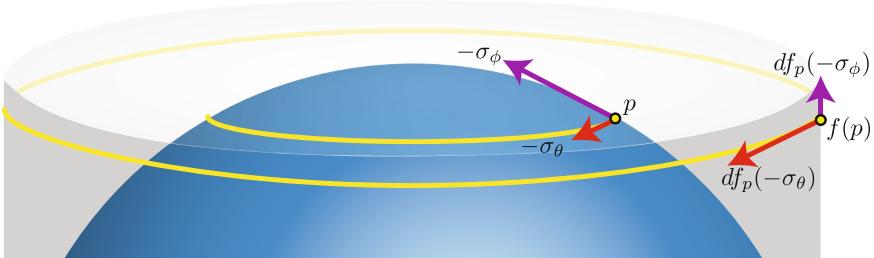


FIGURE 3.32. $\|df_p\| = 1$ because the vectors at p span the same area as the vectors at $f(p)$

tion is obtained by projecting a globe onto the cylinder (via the function f from Eq. 3.18) and then unwrapping the cylinder into a flat rectangle (via the inverse of the cylinder's surface patch in Example 3.23 on page 128). This process and the resulting map are pictured in Fig. 3.33. The orange ellipses help you visualize the separate vertical and horizontal distortion factors of the projection at several points. Technically, they are the images of equal-sized round circles under the derivative of the projection (from globe to map) at a grid of equally spaced points along the map.

According to Archimedes's theorem, the Lambert projection creates an “equal-area” map. The area of every continent or country on the spherical globe equals the area of its projection onto this rectangular map. The map therefore correctly displays relative areas, such as the fact that Africa is about 3.7 times the size of Australia. A common trivial modification is to compress the Lambert map horizontally. For example, the **Gall–Peters map** is nothing more than the Lambert map horizontally compressed until the expansion–contraction factor equals 1 along the 45° parallels north and south ($\phi = \pi/4$ and $\phi = 3\pi/4$), rather than along the equator. Shapes appear undistorted along these two special parallels, which are shown as thick orange lines in Fig. 3.33.

Although these equal-area maps represent area accurately, we saw in Fig. 3.30 that they distort angles and shapes. The distortion is extreme near the poles. Could a map of the Earth preserve angles? This question leads to our second generalization of isometries:

DEFINITION 3.70.

Let S_1, S_2 be a pair of regular surfaces. A diffeomorphism $f : S_1 \rightarrow S_2$ is called **conformal** if df preserves angles, that is, if

$$\angle(x, y) = \angle(df_p(x), df_p(y)) \text{ for all } p \in S_1 \text{ and all } x, y \in T_p S_1.$$

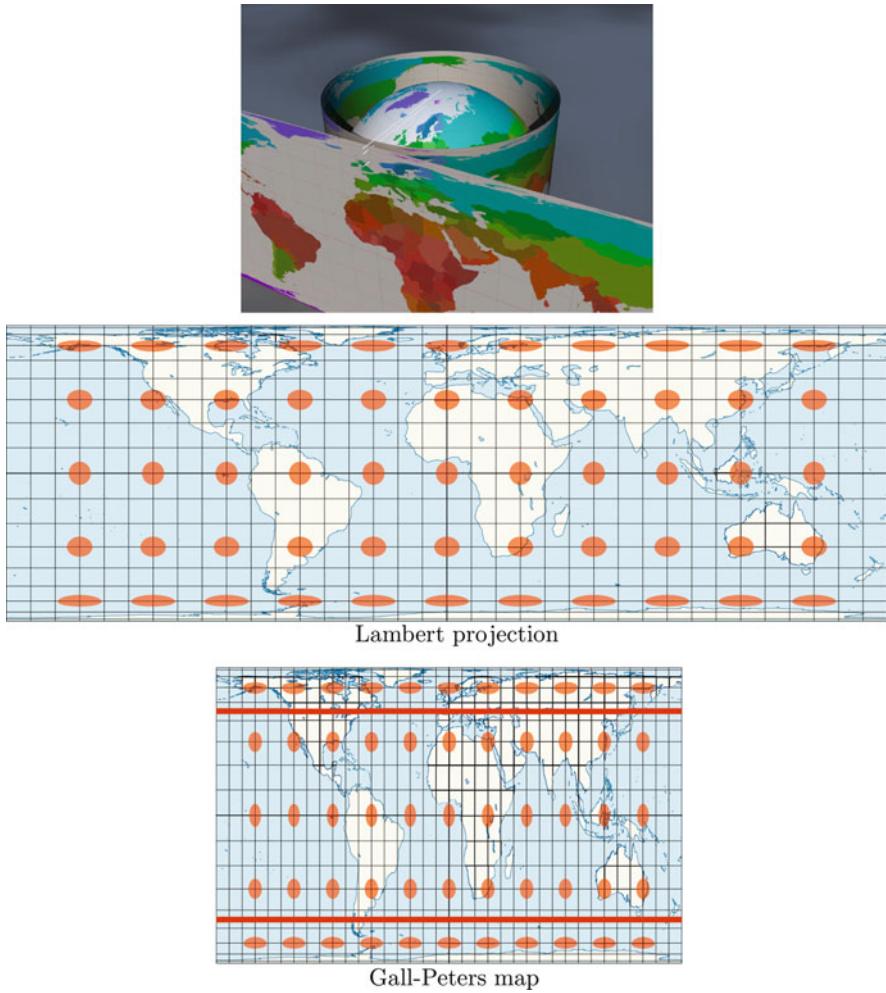


FIGURE 3.33. Equal-area maps of the Earth

EXAMPLE 3.71. Let $\lambda > 0$. Let $f = L_A : \mathbb{R}^2 \rightarrow \mathbb{R}^2$ denote the linear transformation represented by the matrix $A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$, that is, $f(p) = \lambda p$. Let $p \in \mathbb{R}^2$ and $x, y \in T_p \mathbb{R}^2 = \mathbb{R}^2$. Notice that $df_p(x) = f(x) = \lambda x$. Thus

$$\begin{aligned} \angle(df_p(x), df_p(y)) &= \cos^{-1} \left(\frac{\langle df_p(x), df_p(y) \rangle}{|df_p(x)| |df_p(y)|} \right) = \cos^{-1} \left(\frac{\langle \lambda x, \lambda y \rangle}{|\lambda x| |\lambda y|} \right) \\ &= \cos^{-1} \left(\frac{\langle x, y \rangle}{|x| |y|} \right) = \angle(x, y), \end{aligned}$$

so f is conformal. Figure 3.34 shows the effect of f (with $\lambda = 2$) on a circle and some angles.

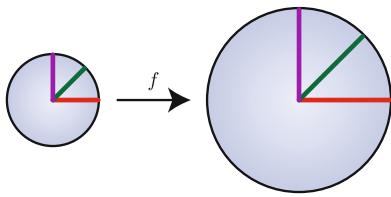


FIGURE 3.34. A linear conformal map with distortion factor $\lambda = 2$

The above example is very special, because f is linear, but it illustrates the infinitesimal picture for a general conformal map. That is, if $f : S_1 \rightarrow S_2$ is any conformal map, then for every $p \in S_1$, there exist orthonormal bases of $T_p S_1$ and $T_{f(p)} S_2$ with respect to which the linear transformation df_p is represented by a

matrix of the form $A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$. The distortion factor λ typically depends on p , and is the factor by which df_p uniformly scales the norms of all vectors. In summary:

PROPOSITION 3.72.

A diffeomorphism $f : S_1 \rightarrow S_2$ is conformal if and only if there exists a smooth positive-valued function $\lambda : S_1 \rightarrow \mathbb{R}$ such that

$$\langle df_p(x), df_p(y) \rangle_{f(p)} = \lambda(p)^2 \cdot \langle x, y \rangle_p \text{ for all } p \in S_1 \text{ and all } x, y \in T_p S_1.$$

PROOF. Follows from Exercise 3.89. □

The best-known example of a conformal map is the **stereographic projection** function $f : S^2 - \{(0, 0, 1)\} \rightarrow \mathbb{R}^2$. By definition, $f(p)$ gives the xy -coordinates of the intersection of the plane $z = -1$ with the line containing $(0, 0, 1)$ and p ; see Fig. 3.35 (some authors instead project onto the plane $z = 0$; see Exercise 3.93).

PROPOSITION 3.73.

The stereographic projection function f is conformal.

PROOF. We first must describe a formula for f . An arbitrary point $p \in S^2$ can be expressed as $p = (\underbrace{r \cos \theta}_x, \underbrace{r \sin \theta}_y, \underbrace{\pm \sqrt{1 - r^2}}_z)$ for some angle θ and some $r > 0$. Its image $f(p)$ will have the form $(R \cos \theta, R \sin \theta)$ for some $R > 0$. To complete our description of f (and to understand f^{-1}), we must relate R and r . By similar triangles,

$$(3.19) \quad R = \frac{2r}{1 - z} = \frac{2r}{1 \pm \sqrt{1 - r^2}} \quad \text{solve for } r \quad r = \frac{4R}{R^2 + 4}.$$

In summary, a formula for f^{-1} is

$$f^{-1}(R \cos \theta, R \sin \theta) = \left(r \cos \theta, r \sin \theta, \pm \sqrt{1 - r^2} \right), \quad \text{where } r = \frac{4R}{R^2 + 4}.$$

The symbol “ \pm ” equals 1 if $R > 2$ and equals -1 if $R < 2$.

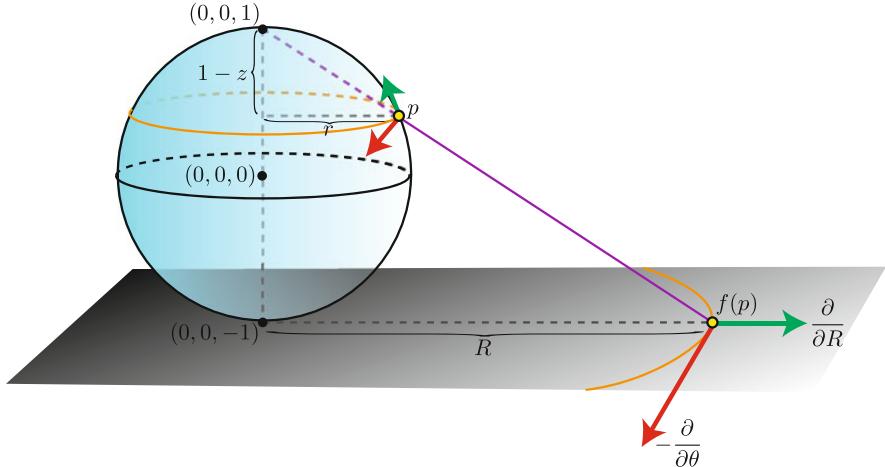


FIGURE 3.35. Stereographic projection is conformal because its derivative scales the green and red vectors by the same factor

Now we prove that f is conformal. It is equivalent to prove that its inverse $f^{-1} : \mathbb{R}^2 \rightarrow S^2 - \{(0,0,1)\}$ is conformal. Let $q = (R \cos \theta, R \sin \theta) \in \mathbb{R}^2$. A basis for $T_q \mathbb{R}^2 = \mathbb{R}^2$ is

$$\frac{\partial}{\partial \theta} = (-R \sin \theta, R \cos \theta), \quad \frac{\partial}{\partial R} = \frac{q}{|q|} = (\cos \theta, \sin \theta).$$

These two vectors are orthogonal. Their images under $d(f^{-1})_q$ are also orthogonal:

$$d(f^{-1})_q \left(\frac{\partial}{\partial \theta} \right) = (-r \sin \theta, r \cos \theta, 0), \quad d(f^{-1})_q \left(\frac{\partial}{\partial R} \right) = \frac{dr}{dR} \left(\cos \theta, \sin \theta, \frac{\pm r}{\sqrt{1-r^2}} \right).$$

We have identified a pair of orthogonal vectors sent by $d(f^{-1})_q$ to a pair of orthogonal vectors, so it will suffice to verify that these vectors are scaled by the same amount; think about why this will suffice, or see Exercise 3.89 for more details. For this, it is straightforward to compute

$$(3.20) \quad \lambda = \frac{|d(f^{-1})_q \left(\frac{\partial}{\partial \theta} \right)|}{\left| \frac{\partial}{\partial \theta} \right|} = \frac{|d(f^{-1})_q \left(\frac{\partial}{\partial R} \right)|}{\left| \frac{\partial}{\partial R} \right|} = \frac{4}{4+R^2}.$$

□

Figure 3.36 shows a map made from a stereographic projection of the globe (turned upside down in order to depict the northern hemisphere). This map even includes part of the southern hemisphere (overlaid white), although an unbounded map would be required to include all of the southern hemisphere. This figure also depicts an artistic representation of stereographic

projection as the shadowing from a light source positioned at the pole. In addition to these still images, we recommend the YouTube video *Stereographic projection of the Riemann sphere* for visualizing the conformal property of stereographic projection.

Stereographic projection is important within mathematics, but not within map-making. Cartographers care much more about a different conformal map—the **Mercator projection**, illustrated in Fig. 3.37. This well-recognized projection has greatly influenced our collective mental image of the Earth.



FIGURE 3.36. Stereographic projection map of northern hemisphere

EXAMPLE 3.74 (The Mercator Projection). *The idea is to modify the spherical coordinate chart so it becomes conformal. Recall from Example 3.24 on page 129 the spherical coordinate chart $\sigma : \underbrace{(0, 2\pi) \times (0, \pi)}_U \rightarrow S^2$ defined as*

$$\sigma(\theta, \phi) = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi).$$

We computed in that example that

$$\sigma_\theta = (-\sin \phi \sin \theta, \sin \phi \cos \theta, 0), \quad \sigma_\phi = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi).$$

The vectors σ_θ and σ_ϕ are orthogonal, but σ is not conformal, because

$$\sin \phi = |\sigma_\theta| \neq |\sigma_\phi| = 1.$$

We can fix this problem by precomposing σ with a function that appropriately warps the vertical axis of the domain. That is, we define the function

$$\Psi : \underbrace{(0, 2\pi) \times (a, b)}_{\tilde{U}} \rightarrow \underbrace{(0, 2\pi) \times (0, \pi)}_U$$

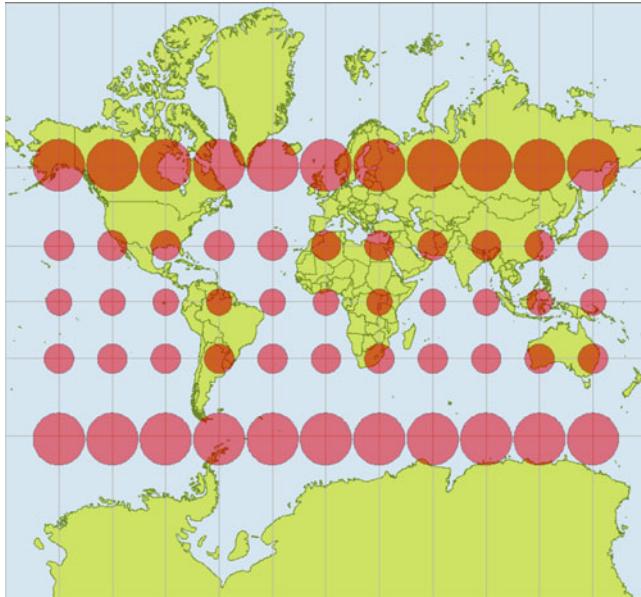


FIGURE 3.37. The Mercator projection

such that $\Psi(\theta, y) = (\theta, \phi(y))$ for some yet-to-be-determined numbers $a, b \in \mathbb{R}$ and function $\phi : (a, b) \rightarrow (0, \pi)$. Here we're thinking of \tilde{U} as the θy -plane, and of U as the $\theta\phi$ -plane.

The composition $\tilde{\sigma} = \sigma \circ \Psi : \tilde{U} \rightarrow S^2$ will be our modified coordinate chart for the sphere. Notice that at every $q = (\theta, y) \in \tilde{U}$, we have

$$|\tilde{\sigma}_\theta| = \sin(\phi(y)), \quad |\tilde{\sigma}_y| = |\phi'(y)|.$$

Since the orthonormal basis $\{e_1, e_2\}$ of $T_q \tilde{U} = \mathbb{R}^2$ is sent by $d\tilde{\sigma}_q$ to the orthogonal basis $\{\tilde{\sigma}_\theta, \tilde{\sigma}_y\}$ of $T_{\tilde{\sigma}(q)} S^2$, to prove that $\tilde{\sigma}$ is conformal, it suffices to ensure that $|\tilde{\sigma}_\theta| = |\tilde{\sigma}_y|$ (see Exercise 3.89). For this, we must choose the function $\phi(y)$ so solve the differential equation

$$\phi'(y) = \sin(\phi(y)).$$

Set $a = -\infty$ and $b = \infty$, and verify that $\phi(y) = 2 \cot^{-1}(e^{-y})$ is a solution with the natural initial condition $\phi(0) = \pi/2$. To obtain a bounded map of the Earth, one chooses finite values $a = -L, b = L$, and thereby settles for leaving neighborhoods of the poles uncharted.

In summary, the spherical coordinate chart preserves vertical lengths but compresses horizontal lengths. The Mercator projection is obtained by modifying it so that it compresses vertical lengths just as badly as horizontal. Since Mercator predated the invention of calculus, he probably constructed his map through the less-precise mechanical process of simply adjusting the spacing between latitude lines until angles seemed to be correctly represented;

more precisely, until every line on his map making an arbitrary angle α with all latitudes corresponded to a “rhumb line” on the globe—a curve that really did meet all latitudes of the globe at the angle α . This “rhumb line” property is desirable for navigation purposes, and will be explored further in Exercise 3.99.

Like any conformal map, Mercator’s map represents angles accurately. Equivalently, it represents small shapes accurately—a truly round island will appear round on the Mercator map, provided the island is small enough so that the function λ in Proposition 3.72 is approximately constant over it. However, the Mercator projection distorts areas, with a distortion factor that becomes more and more severe towards the poles. This has led countless school children to believe that Antarctica is much larger than it actually is.

Is it possible to have the best of both worlds—a map that represents angles and areas accurately? According to the following proposition, this could be achieved only by an *isometric* surface patch for the sphere:

PROPOSITION 3.75.

A diffeomorphism $f : S_1 \rightarrow S_2$ is an isometry if and only if it is equiareal and conformal.

PROOF. Exercise 3.90 □

We will later prove that there does not exist an isometry between open subsets of the plane and of the sphere (Exercise 5.33 on page 274). Intuitively, if a flat rubber sheet is stretched over any part of the sphere, then distances must be distorted—there must be regular curves on the flat rubber sheet whose lengths are different after the stretching. Thus, it’s impossible for a flat map of the Earth to be both equiareal and conformal. The question of which map is best depends on the intended use. The Mercator projection has properties that historically made it ideally suited for navigation at sea, while equiareal maps are much better suited for statistical purposes.

We end this section by collecting and summarizing in a uniform way the various equivalent definitions of the three types of maps studied in this and the previous section. A diffeomorphism $f : S_1 \rightarrow S_2$ is **an isometry/equareal/conformal** if for each $p \in S_1$, $df_p : T_p S_1 \rightarrow T_{f(p)} S_2$ preserves **inner products/areas/angles**. We therefore can understand these three classes of maps by understanding what it means for a linear transformation to preserve these three measurements:

PROPOSITION 3.76.

Let $\mathcal{V}_1, \mathcal{V}_2 \subset \mathbb{R}^3$ be a pair of two-dimensional subspaces of \mathbb{R}^3 . Let $g : \mathcal{V}_1 \rightarrow \mathcal{V}_2$ be a linear transformation. Let $\mathcal{B}_1 = \{x_1, y_1\}$ and $\mathcal{B}_2 = \{x_2, y_2\}$ denote orthonormal bases for \mathcal{V}_1 and \mathcal{V}_2 respectively. Let A be the matrix that represents g with respect to these bases.

- (1) Equivalent characterizations of “inner-product-preserving”:
- $\langle g(x), g(y) \rangle = \langle x, y \rangle$ for all $x, y \in \mathcal{V}_1$.
 - $|g(x)| = |x|$ for all $x \in \mathcal{V}_1$.
 - A is an orthogonal matrix.
 - \mathcal{B}_2 can be rechosen so that A is the identity matrix.
- (2) Equivalent characterizations of “area-preserving”:
- $|g(x) \times g(y)| = |x \times y|$ for all $x, y \in \mathcal{V}_1$.
 - $\|g\| = 1$.
 - $\det(A) = \pm 1$.
 - \mathcal{B}_1 and \mathcal{B}_2 can be rechosen so that $A = \begin{pmatrix} \lambda & 0 \\ 0 & 1/\lambda \end{pmatrix}$ for some $\lambda > 0$.
- (3) Equivalent characterizations of “angle-preserving”:
- $\angle(g(x), g(y)) = \angle(x, y)$ for all $x, y \in \mathcal{V}_1$.
 - $\exists \lambda > 0$ such that $\langle g(x), g(y) \rangle = \lambda^2 \langle x, y \rangle$ for all $x, y \in \mathcal{V}_1$.
 - $\exists \lambda > 0$ such that $|g(x)| = \lambda|x|$ for all $x \in \mathcal{V}_1$.
 - A is a scalar multiple of an orthogonal matrix.
 - \mathcal{B}_2 can be rechosen so that $A = \begin{pmatrix} \lambda & 0 \\ 0 & \lambda \end{pmatrix}$ for some $\lambda > 0$.
 - There exists an orthogonal pair of nonzero vectors $x, y \in \mathcal{V}_1$ such that $g(x), g(y)$ are orthogonal and $\frac{|g(x)|}{|x|} = \frac{|g(y)|}{|y|}$.

The proof is left for the exercises.

EXERCISES

EXERCISE 3.86. Prove Proposition 3.67.

EXERCISE 3.87. Prove part (1) of Proposition 3.76 with these added characterizations:

- There exists an orthonormal basis $\{u, v\}$ of \mathcal{V}_1 such that $\{g(u), g(v)\}$ is an orthonormal basis of \mathcal{V}_2 .
- For every orthonormal basis $\{u, v\}$ of \mathcal{V}_1 , $\{g(u), g(v)\}$ is an orthonormal basis of \mathcal{V}_2 .

Are (e) and (f) still equivalent to each other if “orthonormal” is replaced with “both of unit length”? Prove or give a counterexample.

EXERCISE 3.88. Prove part (2) of Proposition 3.76.

HINT: To show that (b) implies (d), choose x_1 as a global maximum (among all unit vectors in \mathcal{V}_1) of the function $v \mapsto |g(v)|^2$. Choose y_1 to be of unit length and orthogonal to x_1 . Define $x_2 = \frac{g(x_1)}{|g(x_1)|}$, and choose y_2 to be of unit length and orthogonal to x_2 . Now A has the form $A = \begin{pmatrix} \lambda & \mu \\ 0 & 1/\lambda \end{pmatrix}$. Show that a nonzero value of μ would contradict the maximality assumption.

EXERCISE 3.89. Prove part (3) of Proposition 3.76.

EXERCISE 3.90. Prove Proposition 3.75.

EXERCISE 3.91. Verify the following formulas that express the stereographic projection function f and its inverse in rectangular coordinates:

$$f(x, y, z) = \frac{2}{1-z}(x, y), \quad f^{-1}(x, y) = \frac{(4x, 4y, x^2 + y^2 - 4)}{x^2 + y^2 + 4}.$$

EXERCISE 3.92. A “circle in S^2 ” means the intersection of S^2 with a plane (not necessarily passing through the origin). Prove that stereographic projection matches

$$(a \text{ circle in } S^2) \leftrightarrow (a \text{ circle or a line in } \mathbb{R}^2).$$

HINT: By the previous exercise, if $(X, Y, Z) = f^{-1}(x, y) \in S^2$ lies on the plane described by $AX + BY + CZ = D$, then $(x, y) \in \mathbb{R}^2$ satisfies the equation

$$(C - D)(x^2 + y^2) + 2Ax + 2By = C + D.$$

Visually confirm this using a computer graphing application to plot the image under f^{-1} of several circles and lines in \mathbb{R}^2 .

EXERCISE 3.93 (Alternative Stereographic Projection). Some authors instead define “stereographic projection” as the function $F : S^2 - \{(1, 0, 0)\} \rightarrow \mathbb{R}^2$ such that $F(p)$ equals the xy -coordinates of the intersection of the plane $z = 0$ with the line containing $(1, 0, 0)$ and p . That is, they shadow onto the plane $z = 0$ rather than the plane $z = -1$. Verify that:

- (1) $F = \frac{1}{2}f$, where f is the stereographic projection map defined in this section.
- (2) Exercise 3.91 for F becomes

$$F(x, y, z) = \frac{1}{1-z}(x, y), \quad F^{-1}(x, y) = \frac{(2x, 2y, x^2 + y^2 - 1)}{x^2 + y^2 + 1}.$$

- (3) Equations 3.19 and 3.20 for F become

$$r = \frac{2R}{R^2 + 1}, \quad \lambda = \frac{2}{1 + R^2}.$$

EXERCISE 3.94. In defining stereographic projection, what happens if you shine light from $(0, 0, 0)$ rather than $(0, 0, 1)$? That is, let $H = \{(x, y, z) \in S^2 \mid z < 0\}$ denote the southern hemisphere of the sphere, and define the **gnomonic projection** $g : H \rightarrow \mathbb{R}^2$ so that $g(p)$ gives the xy -coordinates of the intersection of the plane $z = -1$ with the line containing $(0, 0, 0)$ and p . Is g conformal? Show that the intersection with H of an arbitrary great circle in S^2 is projected to a straight line in \mathbb{R}^2 . Because of this property, a gnomonic map is useful for charting paths on the Earth that follow great circles.

EXERCISE 3.95. Recall that the Gall–Peters map is obtained from the Lambert map by horizontally scaling until the expansion–contraction factor equals 1 along the 45° parallels north and south ($\phi = \pi/4$ and $\phi = 3\pi/4$), rather than along the equator. What is the correct horizontal scaling factor?

EXERCISE 3.96. Prove that every surface of revolution (as in Example 3.25 on page 131) has an atlas of conformal surface patches. *HINT: Generalize the construction of the Mercator projection from Example 3.74.*

EXERCISE 3.97. Prove that every surface of revolution (as in Example 3.25 on page 131) has an atlas of equiareal surface patches.

EXERCISE 3.98. Prove that Enneper's surface (the parametrized surface σ defined in Exercise 3.27 on page 139) is conformal.

EXERCISE 3.99. Let $\gamma : \mathbb{R} \rightarrow \mathbb{R}^2$ be a logarithmic spiral (defined in Exercise 1.3 on page 6).

- (1) A **rhumb line** (also called a **loxodrome**) means a curve in S^2 that crosses all longitudes at the same angle. If f denotes stereographic projection, prove that $f^{-1} \circ \gamma$ is a rhumb line. *HINT: Use Exercise 1.16 on page 15 and the fact that f is conformal.*
- (2) Explain why rhumb lines correspond to straight lines on a Mercator map.
- (3) Use a computer graphing application to plot the rhumb lines corresponding to several choices of the parameters c and λ , similar to the one displayed in Fig. 3.38.

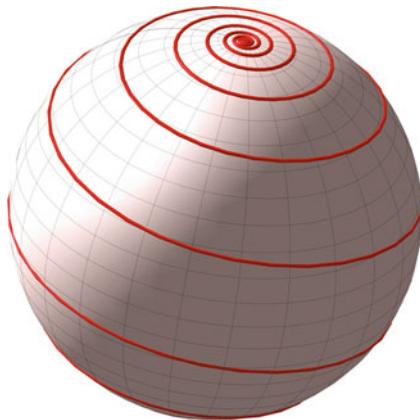


FIGURE 3.38. A rhumb line of S^2



9. The First Fundamental Form in Local Coordinates

In this section, we present the classical notation for expressing the first fundamental form in local coordinates, which is a convenient notation convention that will facilitate local coordinate calculations in future chapters.

Let S be a regular surface and let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch. Define the functions $E, F, G : U \rightarrow \mathbb{R}$ such that for all $q \in U$,

$$E(q) = |\sigma_u(q)|_p^2, \quad F(q) = \langle \sigma_u(q), \sigma_v(q) \rangle_p, \quad G(q) = |\sigma_v(q)|_p^2,$$

where $p = \sigma(q)$. We will frequently suppress inputs and subscripts, simply writing

$$\boxed{E = |\sigma_u|^2, \quad F = \langle \sigma_u, \sigma_v \rangle, \quad G = |\sigma_v|^2.}$$

These three functions encode the first fundamental form at the points of V . In terms of them, we can compute any intrinsic measurement (any measurement that depends only on the first fundamental form), including lengths, angles, and areas. Here's how:

LEMMA 3.77.

(1) If $x = (a, b), y = (c, d) \in \mathbb{R}^2$, then

$$\begin{aligned} \langle d\sigma_q(x), d\sigma_q(y) \rangle_p &= \langle a\sigma_u(q) + b\sigma_v(q), c\sigma_u(q) + d\sigma_v(q) \rangle_p \\ &= (ac)E(q) + (ad + bc)F(q) + (bd)G(q). \end{aligned}$$

(2) If $x = (a, b) \in \mathbb{R}^2$, then $|d\sigma_q(x)|_p^2 = a^2E(q) + 2abF(q) + b^2G(q)$.

(3) If $\gamma : [0, l] \rightarrow U$ is a regular curve with component functions denoted by $\gamma(t) = (u(t), v(t))$, then

$$\text{Arclength}(\sigma \circ \gamma) = \int_0^l \sqrt{u'(t)^2 E(\gamma(t)) + 2u'(t)v'(t)F(\gamma(t)) + v'(t)^2 G(\gamma(t))} dt.$$

(4) If $R \subset V$ is a polygonal region, then

$$\text{Area}(R) = \iint_{\sigma^{-1}(R)} \sqrt{EG - F^2} dA.$$

The angle $\angle(d\sigma_q(x), d\sigma_q(y))$ can be also expressed in terms of $\{E, F, G\}$ by combining (1) and (2) with Eq. 3.16 on page 165.

PROOF. For part (1), use the bilinear and symmetric properties of the inner product. Part (2) is the special case $x = y$ of part (1). Part (3) follows from (2) with $\gamma'(t) = (u'(t), v'(t))$ playing the role of x . Part (4) follows from the definition of area together with Lemma 1.43(2) on page 41, since

$$\|d\sigma\| = |\sigma_u \times \sigma_v| = \sqrt{|\sigma_u|^2 |\sigma_v|^2 - \langle \sigma_u, \sigma_v \rangle^2} = \sqrt{EG - F^2}.$$

□

This lemma leads us to the following definition:

DEFINITION 3.78.

The **first fundamental form in the local coordinates** $\{\mathbf{u}, \mathbf{v}\}$ (also called “the first fundamental form of σ ”) is the expression

$$\mathcal{F}_1 = E du^2 + 2F du dv + G dv^2.$$

You might initially think of this expression as nothing more than a memory trick for part (3) of the above lemma:

$$\begin{aligned}\text{Arclength}(\sigma \circ \gamma) &= \int_0^l \sqrt{\mathcal{F}_1} \\ &= \int_0^l \sqrt{E du^2 + 2F du dv + G dv^2} \\ &= \int_0^l \sqrt{E \left(\frac{du}{dt}\right)^2 + 2F \left(\frac{du}{dt}\right) \left(\frac{dv}{dt}\right) + G \left(\frac{dv}{dt}\right)^2} dt.\end{aligned}$$

The “ dt ”-cancelation is a symbol manipulation that doesn’t require justification if you regard this simply as a memory trick for remembering Lemma 3.77(3).

But in truth, the first fundamental form in local coordinates is more than a memory trick. The expression “ du ” should be interpreted as the derivative of the function on U that maps $(u, v) \mapsto u$. Similarly, “ dv ” is the derivative of the function on U that maps $(u, v) \mapsto v$. So if $x = (a, b)$, then $du_q(x) = a$ and $dv_q(x) = b$ for every $q \in U$. With this understanding, the first fundamental form in local coordinates at $q \in U$ is more than just a formal expression—it’s the function that sends $x \in T_q U = \mathbb{R}^2$ to $|d\sigma_q(x)|^2$. To summarize:

$$(\mathcal{F}_1)_q(x) = |d\sigma_q(x)|^2 \text{ for all } q \in U \text{ and all } x \in T_q U = \mathbb{R}^2.$$

In words, \mathcal{F}_1 associates to each $q \in U$ the function that inputs a tangent vector, x , to U at q , and outputs the number $|d\sigma_q(x)|^2$. This is sometimes summarized by saying that \mathcal{F}_1 is the σ -pullback to U of the first fundamental form on S .

Indeed, the point is to pull back information about S to U via the surface patch σ . Imagine inhabitants of U who know that their home is identified with (part of) some surface S via some surface patch σ , but they don’t know a formula for S or for σ . What can they learn about S ? Absolutely nothing. But what if they also know a formula for the first fundamental form in these local coordinates? This means they know formulas for the E, F, G functions. Equivalently, it means they have a formula to compute $|d\sigma_q(x)|^2$ for every $q \in U$ and $x \in T_q U = \mathbb{R}^2$. Now they can learn quite a bit about S . For every regular curve γ in their world, they can compute the length of $\sigma \circ \gamma$. For every region in their world, they can compute the area of the region in S with which σ identifies it. Our previous question—which measurements on S depend only on the first fundamental form?—is essentially asking which measurements these inhabitants are capable of computing. An *intrinsic* measurement is essentially a measurement that can be expressed in local coordinates in terms only of the functions E, F , and G .

EXAMPLE 3.79. The spherical coordinate chart $\sigma : \underbrace{(0, 2\pi) \times (0, \pi)}_U \rightarrow S^2$

was defined as $\sigma(\theta, \phi) = (\sin \phi \cos \theta, \sin \phi \sin \theta, \cos \phi)$. In Example 3.24 on page 129, we computed

$$\sigma_\theta = (-\sin \phi \sin \theta, \sin \phi \cos \theta, 0), \quad \sigma_\phi = (\cos \phi \cos \theta, \cos \phi \sin \theta, -\sin \phi).$$

From this, we can compute

$$E = |\sigma_\theta|^2 = \sin^2 \phi, \quad F = \langle \sigma_\theta, \sigma_\phi \rangle = 0, \quad G = |\sigma_\phi|^2 = 1.$$

In summary, the first fundamental form of S^2 in spherical coordinates is the expression

$$(3.21) \quad \underbrace{\sin^2 \phi}_{E} d\theta^2 + \underbrace{0}_{F} d\theta d\phi + \underbrace{1}_{G} d\phi^2,$$

or more concisely, $\sin^2 \phi d\theta^2 + d\phi^2$. This expression simultaneously communicates formulas for E , F , and G , and also tells you which of these corresponds to which local variable.

The arc length of a curve in S^2 of the form $\gamma(t) = \sigma(\theta(t), \phi(t))$, $t \in [0, l]$, is found by integrating the square root of this expression:

$$\text{arclength}(\gamma) = \int_0^l \sqrt{\sin^2(\phi(t)) \theta'(t)^2 + \phi'(t)^2} dt.$$

The area of a region $R \subset S^2$ is

$$\text{area}(R) = \iint_{\sigma^{-1}(R)} \sqrt{EG - F^2} dA = \iint_{\sigma^{-1}(R)} \sin(\phi) dA,$$

as we previously discovered in Example 3.59 on page 163.

Look back at Fig. 3.11 on page 130. The value $E = \sin^2 \phi$ is the squared speed at which the pink latitude is traversed, while $G = 1$ is the squared speed at which the green longitude is traversed. Furthermore, $F = 0$ reflects the fact that the latitudes are everywhere orthogonal to the longitudes. At least in this example, the first fundamental form in local coordinates contains a lot of geometric information about the surface.

EXAMPLE 3.80 (Graphs). If $U \subset \mathbb{R}^2$ is open, and $f : U \rightarrow \mathbb{R}$ is a smooth function, then its graph $G = \{(x, y, f(x, y)) \mid (x, y) \in U\}$ is a regular surface according to Lemma 3.17 on page 123. It is covered by the single surface patch $\sigma : U \rightarrow G$ defined as $\sigma(x, y) = (x, y, f(x, y))$. At $q = (x, y) \in U$, we have

$$\sigma_x(q) = (1, 0, f_x(q)), \quad \sigma_y(q) = (0, 1, f_y(q)).$$

So the first fundamental form of G in these coordinates is

$$\underbrace{(1 + f_x^2)}_E dx^2 + \underbrace{(f_x \cdot f_y)}_F dx dy + \underbrace{(1 + f_y^2)}_G dy^2.$$

EXERCISES

EXERCISE 3.100. If $\sigma : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is a parametrized surface (rather than a surface patch of a regular surface), it still makes sense to define $E, F, G : U \rightarrow \mathbb{R}$ and to define \mathcal{F}_1 exactly as in this section. For each function, describe the largest domain on which it is a parametrized surface, and compute \mathcal{F}_1 :

- (1) $\sigma(u, v) = (u^2, v^2, u^2 + v^2 + u + v)$,
- (2) $\sigma(u, v) = (u, u^2, v^3)$,
- (3) $\sigma(u, v) = (u - v, u + v, u^2 + v^2)$.

EXERCISE 3.101. Prove that the following are equivalent for a diffeomorphism $f : S \rightarrow \tilde{S}$ between regular surfaces:

- (1) f is an isometry.
- (2) For every surface patch $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$, the first fundamental form of σ equals the first fundamental form of $f \circ \sigma$.
- (3) Every $p \in S$ is covered by a surface patch σ such that the first fundamental form of σ equals the first fundamental form of $f \circ \sigma$.

EXERCISE 3.102. If $\sigma : U \rightarrow V \subset S$ and $\tilde{\sigma} : U \rightarrow \tilde{V} \subset \tilde{S}$ are surface patches for regular surfaces (with the same domain $U \subset \mathbb{R}^2$), and they have the same first fundamental form (the same E, F, G functions), prove that V and \tilde{V} are isometric.

EXERCISE 3.103. Let $\gamma : \mathbb{R} \rightarrow \mathbb{R}^3$ be a helix of the form $\gamma(\theta) = (\cos \theta, \sin \theta, c\theta)$, where $c \neq 0$ is a constant, shown green in Fig. 3.39. For each value of θ , consider the infinite line (shown red) through $\gamma(\theta)$ that is parallel to the xy -plane and intersects the z -axis. The union of all these lines is called a **helicoid**, visualized as the surface swept out by the propeller of a rising helicopter (or lowering if $c < 0$). It can be covered by the single surface patch

$$\sigma(\theta, t) = (t \cos \theta, t \sin \theta, c\theta), \quad t, \theta \in (-\infty, \infty).$$

- (1) Describe the first fundamental form in these coordinates.
- (2) What is the area of the portion of the helicoid corresponding to $0 < t < 1$ and $0 < \theta < 4\pi$?
- (3) At a point p of the helicoid, how does the angle that a unit normal vector at p makes with the z -axis depend on the distance of p to the z -axis?

EXERCISE 3.104. Describe the first fundamental form of a generalized cylinder (Exercise 3.18(1) on page 136) in the natural coordinates.

EXERCISE 3.105. Let σ denote the coordinate chart for the surface of revolution defined in Example 3.25 on page 131.

- (1) Verify that the first fundamental form σ is

$$\underbrace{(x'(t)^2 + z'(t)^2)}_E dt^2 + \underbrace{x(t)^2}_G d\theta^2.$$

- (2) Verify that $E_\theta = 0$, while $G_t \neq 0$. In Fig. 3.40, how is this related to the fact that the green vertical edges of the rectangle in U are mapped by σ to equal-length curves on S , while the red horizontal edges are mapped to unequal-length curves on S ? Formulate a general rule that applies to an arbitrary coordinate chart for an arbitrary regular surface.
- (3) Now let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch for an arbitrary regular surface with first fundamental form denoted by $E du^2 + 2F du dv + G dv^2$. Prove that $E_v = G_u = 0$ on U if and only if $\sigma_{uv}(q)$ is normal to $T_{\sigma(q)}S$ for all $q \in U$.

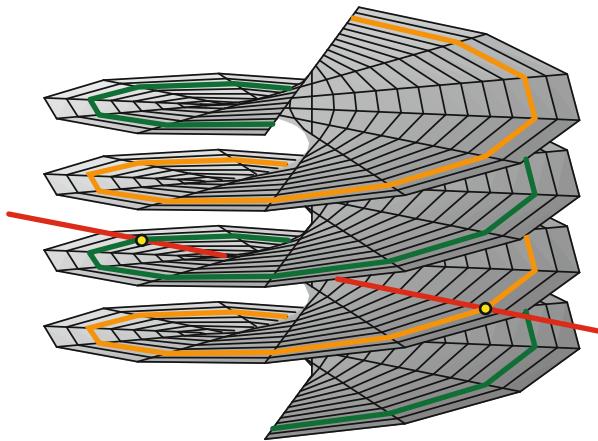
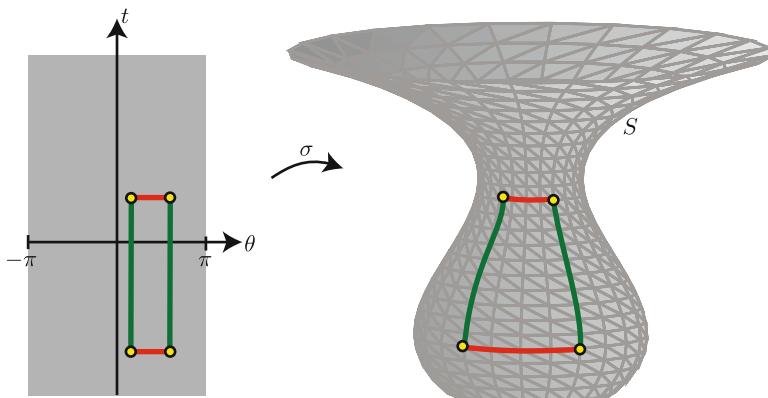


FIGURE 3.39. A helicoid

FIGURE 3.40. The green edges on S have the same length, while the red edges do not

EXERCISE 3.106. Describe the first fundamental form of a generalized cone (Exercise 3.19(1) on page 137) in the natural coordinates.

EXERCISE 3.107. Find the first fundamental form of S^2 in the local coordinates determined by stereographic projection.

EXERCISE 3.108. Let S be a regular surface, and $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ a surface patch with first fundamental form $Edu^2 + 2Fdu\,dv + Gdv^2$.

- (1) Prove that σ is an isometry if and only if $E = G = 1$ and $F = 0$.
- (2) Prove that σ is equiareal if and only if $\sqrt{EG - F^2} = 1$.
- (3) Prove that σ is conformal if and only if $E = G$ and $F = 0$.

EXERCISE 3.109. Classify the functions $f(x, y)$ for which the surface patch in Example 3.80 for the graph of f is: (1) an isometry, (2) equiareal, (3) conformal.

EXERCISE 3.110. Let σ be the surface patch in Example 3.25 on page 131 for the surface of revolution generated by $\gamma = (x(t), z(t))$.

- (1) Classify the choices of γ for which σ is an isometry.
- (2) Write the differential equation that must be satisfied for σ to become conformal after replacing γ with the reparametrization $\gamma \circ \phi$.
- (3) Repeat (2) with “equiareal” instead of “conformal.”

EXERCISE 3.111. Let S be a **tangent developable** (Exercise 3.24 on page 138). Compute the first fundamental form of S in the given local coordinates. Prove that every point of S is covered by a surface patch that is an isometry. *HINT: If C lies in the xy -plane, then S is an open set of the xy -plane, so the result is obvious. By Theorem 1.65 on page 56, there exists a plane curve $\tilde{\gamma}$ with the same curvature function as γ . Since the first fundamental form depends only on the curvature function, the tangent developables of γ and $\tilde{\gamma}$ must be isometric to each other.*

EXERCISE 3.112. Compute the first fundamental form of the **Möbius strip** with respect to the coordinate chart given in this chapter. Verify that this coordinate chart is *not* an isometry (your experience constructing a paper model might have misled your intuition, but we will later prove that this Möbius strip cannot be covered by surface patches that are isometries).

EXERCISE 3.113. Let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch for the regular surface S , and let $g : \tilde{U} \rightarrow U$ be a diffeomorphism between open sets in \mathbb{R}^2 . Derive a formula for the first fundamental form of the surface patch $\sigma \circ g$ in terms of the first fundamental form of σ and the derivative of g .



10. An Alternative Characterization of Regular Surfaces (Optional)

This optional section embodies a primary theme from real analysis: in addition to beautiful theorems about what’s true, we also need pathological examples that illuminate the edge between true and false. The first goal of

this section is to exhibit two strange sets that are not quite regular surfaces. The second goal is to find a simpler criterion for testing whether a set is a regular surface. For this, the strange examples will help us avoid a tempting oversimplification.

This section is intended primarily for readers who are harboring a suspicion that our previous definition of a *regular surface* might include redundancies, or might be possible to simplify. For example, if you check that a purported surface patch σ is smooth and satisfies the rank-2 condition, is it really necessary also to verify that its inverse is smooth? Shouldn't this follow from Proposition 3.29 on page 135?

EXAMPLE 3.81. *We will unsuccessfully try to prove that the following is a regular surface:*

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 < (1/3)^2 \text{ and } z \text{ is rational}\}.$$

For each positive integer $n \in \mathbb{N}$, let U_n be the open ball in \mathbb{R}^2 centered at $(n, 0)$ with radius $1/3$, and let U denote their union; that is,

$$U_n = B((n, 0), 1/3) = \{q \in \mathbb{R}^2 \mid \text{dist}(q, (n, 0)) < 1/3\}, \quad U = \bigcup_{n \in \mathbb{N}} U_n.$$

Let $V = S$, which is open in S . Let r_1, r_2, r_3, \dots denote an enumeration of the rational numbers. Define $\sigma : U \rightarrow V$ such that for each n , U_n is mapped in the obvious way onto $V_n = \{(x, y, z) \in V \mid z = r_n\}$. It is straightforward to see that $\sigma : U \rightarrow V$ is a smooth bijective function.

For an arbitrary integer n , the restriction of σ to U_n is a diffeomorphism from U_n to V_n . In particular, σ satisfies the rank-2 conclusion of Proposition 3.21 on page 126. The inverse of this restriction is a smooth function from V_n to U_n . Even though the inverse of this restriction of σ is smooth, this does not mean that the inverse of all of σ is smooth. In fact, $\sigma^{-1} : V \rightarrow U$ is not even continuous—it does not pull back open sets to open sets. For example, the open set U_n pulls back to the set V_n , which is not open in V .

Neither does it work to use the larger atlas in which the restriction of σ to each U_n is considered a separate surface patch. This strategy makes each surface patch become a diffeomorphism onto its image, but its image is not open in S .

It turns out that the set S from this example is not a regular surface (or even a surface). The basic problem is that the topology that S inherits as a subset of \mathbb{R}^3 is different from the topology it inherits through its identification with U . The individual sheets V_n comprising S are like the layers of a phyllo pastry, infinitely densely packed so that every neighborhood of every point of S intersects infinitely many sheets.

The next example is even stranger, because the phyllo-pastry-like sheets are all connected to each other. The example will help to clarify the limit to which the hypotheses of the surface of revolution example (on page 131) can be weakened.

EXAMPLE 3.82. The curve $\gamma : (0, 3) \rightarrow \mathbb{R}^2$ pictured in Fig. 3.41 is parametrized in such a way that:

- For $t \in (0, 1)$ (brown), $\gamma(t) = (6t + 1, 0, \sin(1/t))$.
- For $t \in [1, 2]$ (blue), γ smoothly connects the brown and green curves.
- For $t \in (2, 3)$ (green), $\gamma(t) = (1, 0, 5 - 2t)$.

It is possible to arrange the blue segment so that γ becomes a regular curve with no self-intersections. If the trace, C , of γ is revolved around the z -axis, as in Example 3.25 on page 131, the resulting set S does not deserve the title “surface of revolution,” because it is not a regular surface (or even a surface). The patch $\sigma : U \rightarrow V \subset S$ constructed in that example is a bijective smooth function that satisfies the rank-2 condition, but its inverse is not smooth (or even continuous). In fact, $\sigma^{-1} : V \rightarrow U$ pulls back the open set $(-\pi, \pi) \times (2, 3)$ in U to the revolution of the green curve, which is not open in S .

These examples show that a smooth bijection $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ satisfying the rank-2 condition is not necessarily a diffeomorphism, and so is not necessarily a valid surface patch. This issue is not about differentiability, but it is purely topological. With a mild hypothesis to rule out the phyllopastry phenomenon, the rank-2 condition becomes enough to verify that a smooth bijection is a valid surface patch:

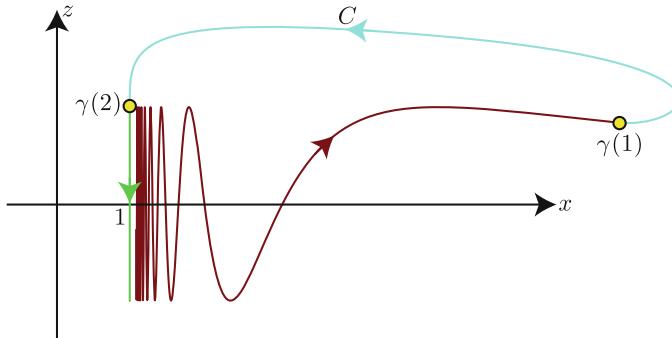


FIGURE 3.41. Rotating C about the z -axis does not result in a regular surface

PROPOSITION 3.83.

A set $S \subset \mathbb{R}^3$ is a regular surface if and only if for each $p \in S$, there exist an open set $U \subset \mathbb{R}^2$ and a smooth injective function $\sigma : U \rightarrow S$ whose image, $V = \sigma(U)$, contains p such that:

- (1) For all $q \in U$, $d\sigma_q$ has rank 2.
- (2) If $\tilde{U} \subset U$ is open, then $\sigma(\tilde{U})$ is open in S .

It is straightforward to see that it would be equivalent to replace condition (2) with the following:

(2') V is open in S and σ^{-1} is continuous.

Thus, once you verify the rank-2 condition, instead of confirming that σ^{-1} is smooth, you need only confirm that σ^{-1} is continuous to prove that σ is a valid surface patch.

Nevertheless, the above phrasing of (2) matches nicely with the examples in this section. The hypothesis that $\sigma(\tilde{U})$ is open in S is equivalent to this: if $\{p_k\}$ is a convergent sequence of points of S whose limit is in $\sigma(\tilde{U})$, then $p_k \in \sigma(\tilde{U})$ for sufficiently large k . Thus, this hypothesis really fails only for pathological examples such as Examples 3.81 and 3.82, where the points p_k can be taken from different phyllo sheets.

PROOF. If S is a regular surface, then every surface patch $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ satisfies (1) by Proposition 3.21 on page 126. If $\tilde{U} \subset U$ is open, then $\sigma(\tilde{U})$ is open in V because $\sigma^{-1} : V \rightarrow U$ is continuous. Since $\sigma(\tilde{U})$ is open in V , and V is open in S , it follows that $\sigma(\tilde{U})$ is open in S .

For the other direction, let $p \in S$, and assume that $\sigma : U \subset \mathbb{R}^2 \rightarrow S$ is a smooth injective function satisfying conditions (1) and (2), whose image, $V = \sigma(U)$, contains p . Define $q = \sigma^{-1}(p)$. By Proposition 3.29 on page 135, there exists an open set $\tilde{U} \subset U$ containing q such that the restriction of σ to \tilde{U} is a diffeomorphism onto its image, $\tilde{V} = \sigma(\tilde{U})$. By hypothesis (2), \tilde{V} is open in S , so $\sigma : \tilde{U} \rightarrow \tilde{V}$ is a valid surface patch covering p . \square

EXERCISES

EXERCISE 3.114. Prove that the set S in Example 3.81 is not a surface.

EXERCISE 3.115. Prove that the set S in Example 3.82 is not a surface.

EXERCISE 3.116. Define C as in Example 3.82, and define

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid (x, 0, z) \in C\}.$$

Is S a regular surface?

EXERCISE 3.117. What is wrong with this reasoning: *The set $S = \mathbb{R}^3$ is a regular surface with an atlas comprising one surface patch for every $\lambda \in \mathbb{R}$, namely the function $\sigma_\lambda : \mathbb{R}^2 \rightarrow S$ defined as $\sigma_\lambda(x, y) = (x, y, \lambda)$.*



This catenoid-shaped bubble curves inward so as to minimize its surface area. This chapter presents natural measurements of how a surface curves in space.

The Curvature of a Surface

A regular surface looks like a plane if one zooms sufficiently in near any point, but if one zooms back out, it might curve and bend through the ambient \mathbb{R}^3 . That's what makes differential geometry so much richer than Euclidean geometry.

The goal of this chapter is to measure the curvature of a regular surface. We will soon define the *Gaussian curvature*, $K(p)$, at each point p of a regular surface S as a real number that reflects how sharply S curves away from $T_p S$ into the ambient \mathbb{R}^3 near p ; see Fig. 4.1.

Here is a brief overview of how we'll do it. First recall from Proposition 1.35 on page 28 that the curvature of a unit-speed curve γ in \mathbb{R}^n is

$$\kappa = \langle -\mathbf{n}', \mathbf{t} \rangle,$$

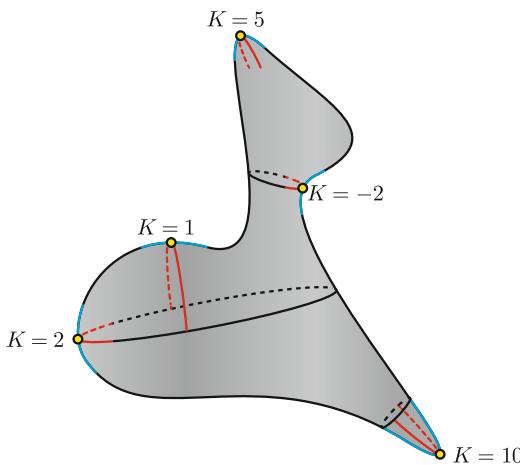


FIGURE 4.1. The Gaussian curvature measures how sharply the surface curves away from its tangent plane at each point

$N(\gamma(t))$, and the measurement

$$II_p(v) = \langle -N'(0), v \rangle$$

will be called the *normal curvature* of S at p in the direction of v ; see Fig. 4.2. Since N is a normal vector to the tangent plane, it encodes the tilt of the tangent plane, so $II_p(v)$ roughly measures the rate at which the tangent plane's tilt changes in moving away from p in the direction v . If I carry a flagpole along γ , then $II_p(v)$ measures the rate at which the changing terrain forces me to tilt it backward or forward in order to keep its tip pointing directly up from the surface. This is part of the answer—at each point of Fig. 4.1 labeled with a large positive K value, the tilt of the tangent plane seems to change rapidly along all curves passing through the point.

You might ask, shouldn't we measure $|N'(0)|$ instead of $\langle -N'(0), v \rangle$? Since $-N'(0)$ is not necessarily parallel to v , we seem to be ignoring its component orthogonal to v . In other words, shouldn't we also measure the rate at which I must tilt the flagpole right and left, not just forward and backward? Don't worry, we will initially keep track of all components of the vector $-N'(0)$. In fact, we will begin by studying the *Weingarten map*, which is the linear transformation sending $v \in T_p S$ to the vector $-N'(0)$ (not just its inner product with v). But it will turn out that no information is lost in restricting attention to normal curvatures. We'll see that one can determine the full Weingarten map just by knowing $II_p(v)$ for all $v \in T_p S$. A little bit of linear algebra will go a long way here.

where \mathbf{t} and \mathbf{n} denote the unit tangent and unit normal vectors along γ . Since \mathbf{n} is orthogonal to the tangent line, it encodes its direction, so this essentially says that curvature is the rate at which the tangent line's direction changes.

To generalize this idea to surfaces, we'll replace \mathbf{n} with a unit normal field *to the surface*, denoted by N . That is, given a regular surface S that is oriented by a unit normal field N , a point $p \in S$, and a unit vector $v \in T_p S$, we'll choose a curve γ in S with $\gamma(0) = p$ and $\gamma'(0) = v$. We set $N(t) =$

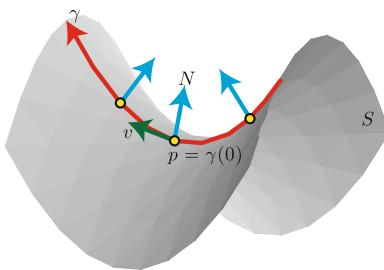


FIGURE 4.2. $II_p(v) = \langle -N'(0), v \rangle$ is called the *normal curvature* of S at p in the direction of v

ture to be the determinant of the Weingarten map, which turns out to equal the product of the maximal and the minimal normal curvatures.

The above paragraphs give only a very brief overview of the strategy that we will use to measure curvature, with the key concepts (normal curvature, the Weingarten map, Gaussian curvature) introduced in the order that best highlights the analogy between the curvature of a curve and the Gaussian curvature of a surface. We will now start over, and use the next few sections to define and explain each concept more precisely, this time in the order required for a sound rigorous development.

1. The Gauss Map

In this section, we begin implementing the above-sketched strategy by defining the Gauss map and the Weingarten map.

Let S be an oriented surface. Let $N : S \rightarrow \mathbb{R}^3$ denote the orientation; that is, N is a unit normal field on S . The fact that it's a *unit* normal field means that it outputs only unit-length vectors in \mathbb{R}^3 ; that is, it outputs only elements of the sphere $S^2 \subset \mathbb{R}^3$. Thus, N is a function from S to S^2 . Regarded in this way (as a smooth map from the regular surface S to the regular surface S^2), N is called the *Gauss map*. In other words, the Gauss map is nothing other than the given unit normal field N , but it's best to imagine the output vectors drawn based at the origin (rather than at points of the surface), so you can visualize these outputs as points of S^2 .

For $p \in S$, consider the derivative $dN_p : T_p S \rightarrow T_{N(p)} S^2$. The domain and codomain of this linear transformation are really the same subspace:

$$T_p S = T_{N(p)} S^2,$$

You might also ask, since $II_p(v)$ depends on v , how can we discuss the curvature of a surface “at a point” rather than “at a point in a direction”? At each labeled point in Fig. 4.1, one could measure the normal curvature in the direction of the red or blue or any other curve through the point. Which single number deserves to be called the Gaussian curvature? Linear algebra again provides the solution—we will define the Gaussian curvature

because $N(p)$ is a normal vector to both. We can therefore regard $dN_p : T_p S \rightarrow T_p S$. Regarded in this way, the negative of this linear transformation is called the *Weingarten map* of S at p . Its determinant is called the *Gaussian curvature* of S at p . Half its trace is called the *mean curvature* of S at p . Both of these measurements are well defined, because the determinant and trace of a linear transformation from a vector space to itself do not depend on the basis with respect to which the linear transformation is represented as a matrix. In summary:

DEFINITION 4.1.

Let S be an oriented surface. The **Gauss map** of S means the unit normal field N regarded as a function from S to S^2 . For every $p \in S$, the linear transformation

$$\mathcal{W}_p = -dN_p : T_p S \rightarrow T_p S$$

is called the **Weingarten map** of S at p . Its determinant and half its trace,

$$K(p) = \det(\mathcal{W}_p) \quad \text{and} \quad H(p) = \frac{1}{2} \operatorname{trace}(\mathcal{W}_p),$$

are respectively called the **Gaussian curvature** and the **mean curvature** of S at p .

The negative sign in the definition of \mathcal{W}_p might initially seem like a nuisance, but its logic will be explained soon. We will now describe the geometric meaning of the Weingarten map (which relies on the interpretation of derivatives found in Proposition 3.7 on page 118). Let γ be a regular curve in S with $\gamma(0) = p$ and $\gamma'(0) = v$. Let $N(t) = N(\gamma(t))$ be the restriction of the unit normal field to γ . Picture each $N(t)$ drawn with tail at the origin, so the tips determine a path on S^2 (shown in blue in Fig. 4.3). The initial velocity vector (shown in purple) of this blue path is

$$dN_p(v) = N'(0) = -\mathcal{W}_p(v).$$

We can picture this purple vector as tangent to S at p (rather than tangent to S^2 at $N(t)$), because these two tangent planes are the same. Visually, if I carry a flag pole along γ , then $-\mathcal{W}_p(v)$ roughly represents the speed and direction in which I must tilt it (as I pass p) to keep its tip pointing directly up from the surface.

Geometric interpretations of the Gaussian curvature will be a major theme of the remainder of this book, while the geometry of the mean curvature will be studied only in Sect. 6 of this chapter. But for now, since the Weingarten map has such a natural geometric meaning, and since the determinant and the trace are essentially the only measurements of a linear transformation of a two-dimensional vector space that are well defined (independent of the choice of basis), we have sufficient motivation for defining the

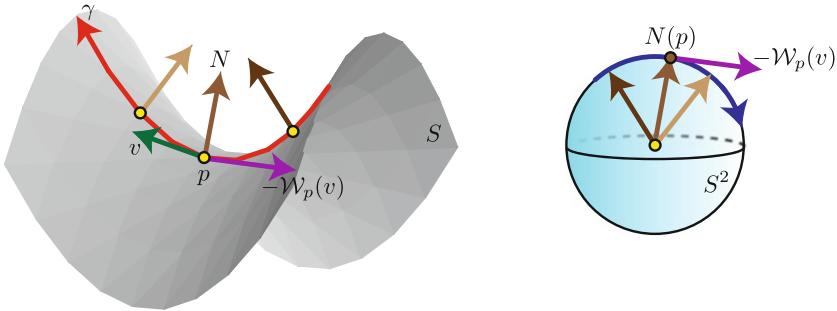


FIGURE 4.3. $N'(0) = -\mathcal{W}_p(v)$ is the initial derivative of the restriction of N to any curve in S through p with initial velocity v

Gaussian and mean curvatures and for expecting to eventually find natural geometric interpretations of these measurements.

In Fig. 4.3, the vectors v and $-\mathcal{W}_p(v)$ appear to be parallel. In other words, v appears to be an eigenvector of \mathcal{W}_p . This generally happens only for special choices of $v \in T_p S$.

Notice that if S were given the other possible orientation, namely $-N$, then for every $v \in T_p S$, the value of $\mathcal{W}_p(v)$ would change to its negative. What would happen to the Gaussian and mean curvatures? All four entries of the 2×2 matrix representing \mathcal{W}_p (with respect to any fixed basis) would change sign. Since 2 is even, it is straightforward to show that this change has no effect on the determinant of the matrix, but it changes the sign of the trace.

Thus, Gaussian curvature (but not mean curvature) is independent of the choice of orientation of S , and is therefore well defined for nonoriented (or even nonorientable) surfaces. If S is any surface, and $p \in S$, then every coordinate chart covering p induces an orientation of a neighborhood of p in S . Although the linear transformation \mathcal{W}_p and the sign of $H(p)$ depend on which orientation the coordinate chart induces, the number $K(p)$ does not, and is therefore well defined.

EXAMPLE 4.2 (A Plane). Let S denote a two-dimensional subspace of \mathbb{R}^3 . According to Example 3.45 on page 153, S can be oriented by a constant unit normal field N . Since N is constant, the derivative of its restriction to any path in S equals zero. Therefore, $\mathcal{W}_p(v) = \mathbf{0}$ for every $p \in S$ and every $v \in T_p S$. Thus, the Gaussian curvature and the mean curvature of S at every point equal 0.

EXAMPLE 4.3 (A Sphere). The sphere of radius r ,

$$S^2(r) = \{p \in \mathbb{R}^3 \mid |p| = r\},$$

has an outward-pointing unit normal field described as

$$N(p) = \frac{p}{|p|} = \frac{p}{r} \quad \text{for all } p \in S^2(r).$$

If γ is a regular curve in $S^2(r)$ with $\gamma(0) = p$ and $\gamma'(0) = v$, then

$$\frac{d}{dt} \Big|_{t=0} N(\gamma(t)) = \frac{d}{dt} \Big|_{t=0} \frac{\gamma(t)}{r} = \frac{v}{r},$$

so $\mathcal{W}_p(v) = -\frac{v}{r}$. Therefore, the matrix for \mathcal{W}_p (with respect to any basis of $T_p S$) is $-\frac{1}{r} \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$. Since p was arbitrary, $S^2(r)$ has constant Gaussian curvature equal to the determinant of this matrix, which is $K = \frac{1}{r^2}$. Furthermore, $S^2(r)$ has constant mean curvature equal to half the trace of this matrix, which is $H = -\frac{1}{r}$. With respect to the other (inward-pointing) orientation, $S^2(r)$ has constant mean curvature $H = \frac{1}{r}$.

EXAMPLE 4.4 (A Critical Point of a Graph). Let $U \subset \mathbb{R}^2$ be open and $f : U \rightarrow \mathbb{R}$ a smooth function. The graph, G , of f is a regular surface covered by the single surface patch $\sigma : U \rightarrow G$ defined as $\sigma(x, y) = (x, y, f(x, y))$. In Example 3.48 on page 153, we computed that

$$\sigma_x = (1, 0, f_x), \quad \sigma_y = (0, 1, f_y), \quad N = \frac{\sigma_x \times \sigma_y}{|\sigma_x \times \sigma_y|} = \frac{(-f_x, -f_y, 1)}{\sqrt{f_x^2 + f_y^2 + 1}}.$$

Let $q = (x_0, y_0) \in U$. We wish to compute the Weingarten map at $p = \sigma(q) = (x_0, y_0, f(x_0, y_0))$ under the added assumption that f has a critical point at q ; in other words, $f_x(x_0, y_0) = f_y(x_0, y_0) = 0$, so that $N(p) = (0, 0, 1)$ and $T_p G = \text{span}\{e_1, e_2\}$, where $e_1 = (1, 0, 0)$ and $e_2 = (0, 1, 0)$.

Define $\gamma_1(t) = (x_0 + t, y_0, f(x_0 + t, y_0))$, which is a regular curve in G with $\gamma_1(0) = p$ and $\gamma'_1(0) = e_1$. The restriction of N to γ_1 is

$$N_1(t) = N(\gamma_1(t)) = \frac{(-f_x(x_0 + t, y_0), -f_y(x_0 + t, y_0), 1)}{\sqrt{f_x^2(x_0 + t, y_0) + f_y^2(x_0 + t, y_0) + 1}}.$$

At time $t = 0$, notice that the denominator of this expression for $N_1(t)$ has value 1 and derivative 0 (because of the critical point assumption). Therefore,

$$N'_1(0) = \frac{d}{dt} \Big|_{t=0} (-f_x(x_0 + t, y_0), -f_y(x_0 + t, y_0), 1) = (-f_{xx}(x_0, y_0), -f_{yx}(x_0, y_0), 0).$$

Thus,

$$\mathcal{W}_p(e_1) = -dN_p(e_1) = -N'_1(0) = (f_{xx}(x_0, y_0), f_{yx}(x_0, y_0), 0).$$

Next, we similarly define $\gamma_2(t) = (x_0, y_0 + t, f(x_0, y_0 + t))$, which is a regular curve in G with $\gamma_2(0) = p$ and $\gamma'_2(0) = e_2$, and let $N_2(t) = N(\gamma_2(t))$. The analogous calculation gives

$$\mathcal{W}_p(e_2) = -dN_p(e_2) = -N'_2(0) = (f_{xy}(x_0, y_0), f_{yy}(x_0, y_0), 0).$$

In summary,

$$\begin{aligned}\mathcal{W}_p(e_1) &= f_{xx}(x_0, y_0) \cdot e_1 + f_{yx}(x_0, y_0) \cdot e_2, \\ \mathcal{W}_p(e_2) &= f_{xy}(x_0, y_0) \cdot e_1 + f_{yy}(x_0, y_0) \cdot e_2.\end{aligned}$$

This says that the matrix for \mathcal{W}_p with respect to the basis $\{e_1, e_2\}$ equals

$$\mathcal{W}_p = \begin{pmatrix} f_{xx}(x_0, y_0) & f_{xy}(x_0, y_0) \\ f_{yx}(x_0, y_0) & f_{yy}(x_0, y_0) \end{pmatrix}.$$

The Gaussian curvature is the determinant

$$K(p) = \det(\mathcal{W}_p) = f_{xx}(x_0, y_0)f_{yy}(x_0, y_0) - f_{xy}(x_0, y_0)^2.$$

Recall the second derivative test from multivariable calculus: if $K(p) > 0$, then the critical point is a local extremum, and if $K(p) < 0$, then it is a saddle point.

The conclusion of the above example is that the Weingarten map (regarded as a matrix with respect to the natural basis) at a critical point of the graph of a function (with the orientation induced by the graph coordinate chart) is just the matrix of second-order partial derivatives of the function. Figure 4.4 shows how this works for four particular quadratic functions all of which have a critical point at $(x_0, y_0) = (0, 0)$. In each graph, the path $\gamma_1(t)$ and the restricted field $N(\gamma_1(t))$ are shown in green, while $\gamma_2(t)$ and $N(\gamma_2(t))$ are shown in red. For each graph, try to visually understand the sign of all entries of the matrix for the Weingarten map.

What can we learn from these examples? For three of them (all except $f = 2xy$), the matrix for \mathcal{W}_p is diagonal; in other words, e_1 and e_2 are eigenvectors of \mathcal{W}_p . Visually, along the red path, the red vectors are tilting only in the direction of the red path (not in the orthogonal green direction), and vice versa. If I carry a flagpole along either path, then the changing terrain forces me to tilt it forward or backward (not side to side) in order to keep its tip pointing directly up from the surface.

The fourth example, $f = 2xy$, is different. Here \mathcal{W}_p has only off-diagonal entries. Walking along the red or green path, I must tilt my flagpole side to side (not forward or backward) to keep its tip pointing directly up from the surface. This is not a bad function, it's just a bad choice of basis. Verify that with respect to the 45° rotated basis

$$(4.1) \quad \{R_{45}(e_1), R_{45}(e_2)\} = \left\{ \left(\sqrt{2}/2, \sqrt{2}/2 \right), \left(-\sqrt{2}/2, \sqrt{2}/2 \right) \right\},$$

the Weingarten map is represented by the matrix $\mathcal{W}_p = \begin{pmatrix} 2 & 0 \\ 0 & -2 \end{pmatrix}$. This is the same matrix we found for $f = x^2 - y^2$ (top right). The reason is that the graph of $f = 2xy$ is obtained by rotating the graph of $f = x^2 - y^2$ counterclockwise 45° about the z -axis (Exercise 4.1).

These examples might lead you to guess that the matrix for the Weingarten map (1) is symmetric with respect to every orthonormal basis, and (2) is diagonal with respect to a properly chosen orthonormal basis. We will now verify (1), which is a step toward proving (2) in the next section.

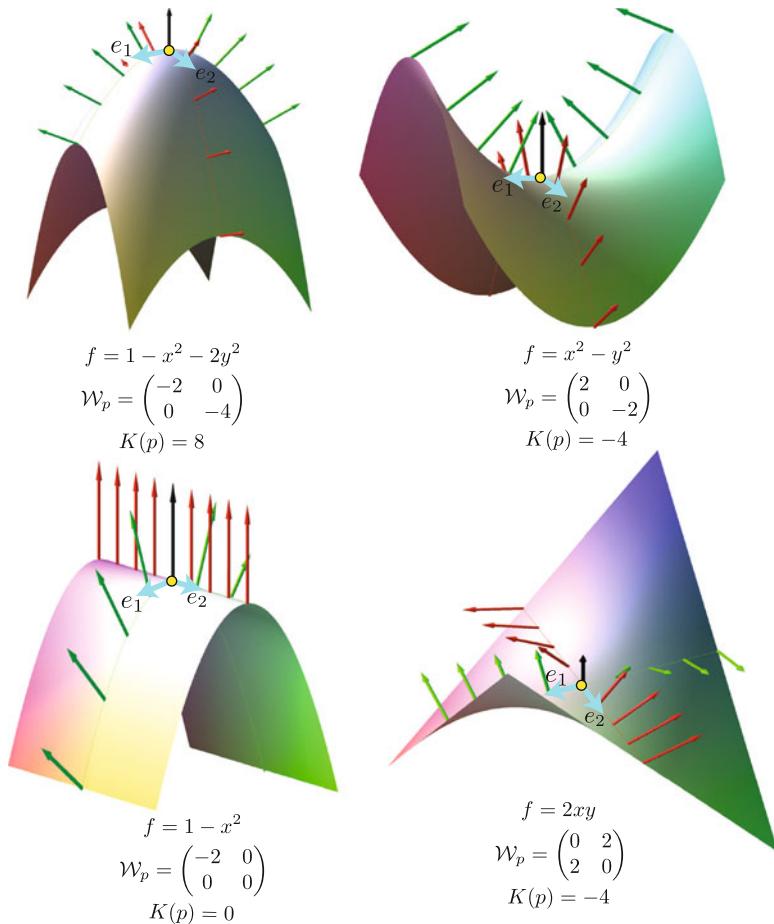


FIGURE 4.4. The Weingarten map at $p = (0, 0, f(0, 0))$ with respect to the basis $\{e_1, e_2\}$ for four quadratic functions

LEMMA 4.5.

If S is an oriented regular surface and $p \in S$, then the Weingarten map \mathcal{W}_p is represented by a symmetric matrix with respect to any orthonormal basis of $T_p S$.

We will give an elementary proof of this lemma at the end of the next section, but for now we will use the inverse function theorem to get the job done. In fact, the following proof reveals the significance of Example 4.4: every point of a regular surface looks like a critical point of a graph, after applying a rigid motion.

PROOF. Let $\{v_1, v_2\}$ be an arbitrary orthonormal basis of $T_p S$. Let $A \in O(3)$ be the orthogonal matrix whose columns are $\{v_1, v_2, N(p)\}$, and

let $g = L_A$ be the corresponding rigid motion of \mathbb{R}^3 . Define $\tilde{S} = g^{-1}(S)$ and $\tilde{p} = (\tilde{x}_0, \tilde{y}_0, \tilde{z}_0) = g^{-1}(p)$. Notice that \tilde{S} is a regular surface for which $T_{\tilde{p}}\tilde{S} = \text{span}\{e_1, e_2\}$.

By Exercise 3.47 on page 146 (which required the inverse function theorem), a neighborhood of \tilde{p} in \tilde{S} equals the graph of a smooth function $f(x, y)$. Example 4.4 now implies that the Weingarten map for \tilde{S} at \tilde{p} with respect to the basis $\{e_1, e_2\}$ is $\begin{pmatrix} f_{xx}(\tilde{x}_0, \tilde{y}_0) & f_{xy}(\tilde{x}_0, \tilde{y}_0) \\ f_{yx}(\tilde{x}_0, \tilde{y}_0) & f_{yy}(\tilde{x}_0, \tilde{y}_0) \end{pmatrix}$, which is symmetric because $f_{xy} = f_{yx}$.

It is straightforward to show that the Weingarten map for S at p with respect to the basis $\{v_1 = g(e_1), v_2 = g(e_2)\}$ is represented by exactly this same matrix (with all entries multiplied by -1 in the case $\det(A) = -1$). \square

To select a basis with respect to which \mathcal{W}_p is a *diagonal* (rather than just symmetric) matrix, we'll need to understand self-adjoint linear transformations, introduced in the next section.

EXERCISES

EXERCISE 4.1. Prove that the graph of $f = 2xy$ is obtained by rotating the graph of $f = x^2 - y^2$ counterclockwise 45° about the z -axis.

EXERCISE 4.2. The level sets of the function $f(x, y, z) = x^2 + y^2 - z^2$ corresponding to the values $f = -1$, $f = 0$, and $f = 1$ were illustrated in Fig. 3.14 on page 134. Describe the image of the Gauss map for each of these level sets (for $f = 0$, the origin must first be removed in order for the set to be a regular surface).

EXERCISE 4.3. Let S be a connected regular surface such that the image of the Gauss map is a great circle of S^2 (which means the intersection of S^2 with a plane through the origin). What is the strongest conclusion you can draw about S ?

EXERCISE 4.4. Let S be an oriented regular surface and let f be a rigid motion of \mathbb{R}^3 . Recall that the image $f(S)$ is also a regular surface with a natural induced orientation (Exercise 3.64(2) on page 158). Describe how the Gaussian and mean curvatures of S and $f(S)$ are related. Repeat if instead, f is the dilation map $p \mapsto \lambda \cdot p$ for some nonzero $\lambda \in \mathbb{R}$.

2. Self-Adjoint Linear Transformations (Linear Algebra Background)

In addition to geometric insight, differential geometry sometimes requires a bit of linear algebra to move forward. In this section, we will abstractly study linear transformations that have the symmetric property of the Weingarten map revealed in Lemma 4.5.

PROPOSITION AND DEFINITION 4.6.

Let $\mathcal{V} \subset \mathbb{R}^3$ be a two-dimensional subspace. A linear transformation $W : \mathcal{V} \rightarrow \mathcal{V}$ is called **self-adjoint** if it satisfies the following equivalent properties:

- (1) There exists an orthonormal basis for \mathcal{V} with respect to which W is represented by a symmetric matrix.
- (2) With respect to any orthonormal basis for \mathcal{V} , W is represented by a symmetric matrix.
- (3) $\langle W(x), y \rangle = \langle x, W(y) \rangle$ for all $x, y \in \mathcal{V}$.

PROOF OF EQUIVALENCE. Suppose that $A = \begin{pmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{pmatrix}$ represents W with respect to the basis $\{v_1, v_2\}$. As reviewed in Sect. 4 of Chap. 3, this just means that

$$W(v_1) = a_{11}v_1 + a_{21}v_2 \quad \text{and} \quad W(v_2) = a_{12}v_1 + a_{22}v_2.$$

If this basis is orthonormal, then as discussed in Sect. 2 of Chap. 1, these coefficients can be expressed as $a_{ij} = \langle v_i, W(v_j) \rangle$, so

$$(4.2) \quad A = \begin{pmatrix} \langle v_1, W(v_1) \rangle & \langle v_1, W(v_2) \rangle \\ \langle v_2, W(v_1) \rangle & \langle v_2, W(v_2) \rangle \end{pmatrix}.$$

(3) \implies (2): With respect to any orthonormal basis $\{v_1, v_2\}$ of \mathcal{V} , W is represented by the matrix in Eq. 4.2, which is symmetric by property (3).

(2) \implies (1): Obvious.

(1) \implies (3): Let $\{v_1, v_2\}$ be an orthonormal basis of \mathcal{V} with respect to which the matrix representing W is symmetric. This matrix is given by Eq. 4.2, so $\langle v_1, W(v_2) \rangle = \langle v_2, W(v_1) \rangle$. An arbitrary pair of vectors in \mathcal{V} can be expressed as $x = av_1 + bv_2$ and $y = cv_1 + dv_2$ for some $a, b, c, d \in \mathbb{R}$. By linearity,

$$\begin{aligned} \langle W(x), y \rangle &= \langle W(av_1 + bv_2), cv_1 + dv_2 \rangle \\ &= ac \langle W(v_1), v_1 \rangle + bc \langle W(v_2), v_1 \rangle + ad \langle W(v_1), v_2 \rangle + bd \langle W(v_2), v_2 \rangle \\ &= ac \langle v_1, W(v_1) \rangle + bc \langle v_2, W(v_1) \rangle + ad \langle v_1, W(v_2) \rangle + bd \langle v_2, W(v_2) \rangle \\ &= \langle av_1 + bv_2, W(cv_1 + dv_2) \rangle = \langle x, W(y) \rangle. \end{aligned}$$

□

Even better than symmetric, we can make the matrix diagonal:

PROPOSITION 4.7.

If $W : \mathcal{V} \rightarrow \mathcal{V}$ is a self-adjoint linear transformation, then there exists an orthonormal basis $\{v_1, v_2\}$ of \mathcal{V} with respect to which W is represented by a diagonal matrix $\begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix}$. In other words, v_1 and v_2 are eigenvectors of W with corresponding eigenvalues λ_1 and λ_2 .

PROOF. Consider the function $Q : \mathcal{V} \rightarrow \mathbb{R}$ defined as $Q(v) = \langle v, W(v) \rangle$. Let $S^1 \subset \mathcal{V}$ denote the circle of all unit-length vectors in \mathcal{V} . Since S^1 is compact, the continuous function Q restricted to S^1 achieves a minimum value, $\lambda_1 \in \mathbb{R}$, at some $v_1 \in S^1$ (Lemma A.25 on page 356 of the appendix). Let $v_2 \in S^1$ be orthogonal to v_1 , and define $\lambda_2 = Q(v_2)$. By Eq. 4.2 on page 202, W is represented by the matrix $\begin{pmatrix} \lambda_1 & b \\ b & \lambda_2 \end{pmatrix}$ with respect to the basis $\{v_1, v_2\}$, where $b = \langle W(v_1), v_2 \rangle = \langle v_1, W(v_2) \rangle$.

An arbitrary $v \in S^1$ has the form $v = (\cos \theta)v_1 + (\sin \theta)v_2$ for some angle θ . Furthermore:

$$Q(v) = \lambda_1 \cos^2 \theta + 2b \cos \theta \sin \theta + \lambda_2 \sin^2 \theta.$$

This function of θ has a global minimum (and hence a critical point) at $\theta = 0$ (which corresponds to $v = v_1$). Therefore,

$$0 = \frac{d}{d\theta} \Big|_{\theta=0} (\lambda_1 \cos^2 \theta + 2b \cos \theta \sin \theta + \lambda_2 \sin^2 \theta) = 2b.$$

Thus, $b = 0$, so the matrix is diagonal. \square

Some ideas from the previous proof deserve to be named and summarized.

PROPOSITION AND DEFINITION 4.8.

The **quadratic form** associated to a self-adjoint linear transformation $W : \mathcal{V} \rightarrow \mathcal{V}$ is the function $Q : \mathcal{V} \rightarrow \mathbb{R}$ defined as

$$Q(v) = \langle v, W(v) \rangle \quad \text{for all } v \in \mathcal{V}.$$

If $\{v_1, v_2\}$ is an orthonormal basis of eigenvectors of W with corresponding eigenvalues $\lambda_1 \leq \lambda_2$ (whose existence is guaranteed by Proposition 4.7), then Q acts on an arbitrary unit vector in \mathcal{V} as

$$Q((\cos \theta)v_1 + (\sin \theta)v_2) = \lambda_1 \cos^2 \theta + \lambda_2 \sin^2 \theta.$$

In particular, λ_1 and λ_2 are respectively the minimum and maximum values that Q attains on the circle of unit vectors in \mathcal{V} .

The last general fact worth mentioning is that knowledge of Q completely determines W by a trick analogous to Eq. 3.17 on page 166; namely, solve the equation

$$Q(x - y) = \langle x - y, W(x - y) \rangle = \underbrace{\langle x, W(x) \rangle}_{Q(x)} + \underbrace{\langle y, W(y) \rangle}_{Q(y)} - 2 \langle x, W(y) \rangle$$

for $\langle x, W(y) \rangle$, to learn that

$$(4.3) \quad \langle x, W(y) \rangle = \frac{1}{2} (Q(x) + Q(y) - Q(x - y)).$$

Thus, W and Q contain exactly the same information, just packaged in different ways: W is a linear transformation from \mathcal{V} to \mathcal{V} , while Q is a quadratic function from \mathcal{V} to \mathbb{R} . A formula for either can be derived from the other.

We end this section with a more elementary proof of Lemma 4.5 from page 200 that does not require the inverse function theorem.

ALTERNATIVE PROOF OF LEMMA 4.5. Let $\sigma : U \subset \mathbb{R}^2 \rightarrow S$ be a surface patch. For every $q \in U$, $(N \circ \sigma)(q)$ is orthogonal to $\sigma_u(q)$; in other words, the function $q \mapsto \langle (N \circ \sigma)(q), \sigma_u(q) \rangle$ equals the zero function on U . In particular, the partial derivative of this function with respect to v vanishes:

$$0 = \frac{\partial}{\partial v} \langle N \circ \sigma, \sigma_u \rangle = \langle (N \circ \sigma)_v, \sigma_u \rangle + \langle N \circ \sigma, \sigma_{uv} \rangle.$$

Therefore,

$$\begin{aligned} \langle (N \circ \sigma)_v, \sigma_u \rangle &= -\langle N \circ \sigma, \sigma_{uv} \rangle, \\ \langle (N \circ \sigma)_u, \sigma_v \rangle &= -\langle N \circ \sigma, \sigma_{vu} \rangle. \end{aligned}$$

(The second equation comes from exchanging the roles of u and v in the above steps.) The right sides of the above two equations are equal to each other, because $\sigma_{uv} = \sigma_{vu}$. Therefore, the left sides must also be equal to each other:

$$\langle (N \circ \sigma)_u, \sigma_v \rangle = \langle (N \circ \sigma)_v, \sigma_u \rangle.$$

On the other hand, it is straightforward to see that for every $q \in U$,

$$(4.4) \quad (N \circ \sigma)_u(q) = -\mathcal{W}_p(\sigma_u(q)) \quad \text{and} \quad (N \circ \sigma)_v(q) = -\mathcal{W}_p(\sigma_v(q)),$$

where $p = \sigma(q)$, simply because the Weingarten map is negative the derivative of the Gauss map. Therefore,

$$\langle \mathcal{W}_p(\sigma_u(q)), \sigma_v(q) \rangle = \langle \mathcal{W}_p(\sigma_v(q)), \sigma_u(q) \rangle.$$

To summarize, $\langle \mathcal{W}_p(x), y \rangle = \langle x, \mathcal{W}_p(y) \rangle$ for all pairs x, y chosen from the basis $\{\sigma_u(q), \sigma_v(q)\}$ of $T_p S$. Even though this basis is not necessarily orthonormal, it follows from the linearity of \mathcal{W}_p (as at the end of the proof of Proposition 4.6) that the same is true for all pairs x, y chosen from all of $T_p S$. Therefore, \mathcal{W}_p is self-adjoint. \square

We now wish to introduce and henceforth adopt a common convention that helps reduce the notational clunkiness of calculations like those found in the above proof. When $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ is a surface patch, it is common to denote by the same name a function on V and its composition with σ . For example, the function $K \circ \sigma : U \rightarrow \mathbb{R}$ will be denoted simply by K whenever this will not cause confusion. More importantly, the function $N \circ \sigma : U \rightarrow \mathbb{R}^3$ will be denoted simply by N whenever this will not cause confusion. For example, the expressions N_u and N_v will be unambiguously interpreted as shorthand for $(N \circ \sigma)_u$ and $(N \circ \sigma)_v$; this is the only possible

meaning, assuming that $\{u, v\}$ are the names of the coordinate variables in U as usual.

With this streamlined notational convention, the key steps in the above proof can be summarized as follows:

$$(4.5) \quad \underbrace{\langle N_u, \sigma_v \rangle}_{-\mathcal{W}(\sigma_u)} = -\langle N, \sigma_{vu} \rangle = -\langle N, \sigma_{uv} \rangle = \underbrace{\langle N_v, \sigma_u \rangle}_{-\mathcal{W}(\sigma_v)}.$$

The same method also verifies that

$$(4.6) \quad \underbrace{\langle N_u, \sigma_u \rangle}_{-\mathcal{W}(\sigma_u)} = -\langle N, \sigma_{uu} \rangle \quad \text{and} \quad \underbrace{\langle N_v, \sigma_v \rangle}_{-\mathcal{W}(\sigma_v)} = -\langle N, \sigma_{vv} \rangle.$$

The underbraced equalities, which come from Eq. 4.4, have their input points suppressed for brevity. For example, the expression “ $-\mathcal{W}(\sigma_u)$ ” should be interpreted as the function on U whose value at $q \in U$ equals $-\mathcal{W}_{\sigma(q)}(\sigma_u(q))$.

EXERCISES

EXERCISE 4.5. Let $\mathcal{V} \subset \mathbb{R}^3$ be a two-dimensional subspace. Let $W : \mathcal{V} \rightarrow \mathcal{V}$ be a self-adjoint linear transformation. Let Q be the quadratic form associated to W .

- (1) Prove that the trace of W equals twice the average value of Q on the circle of unit vectors in \mathcal{V} .
- (2) If $x, y \in \mathcal{V}$ are orthonormal, prove that $Q(x) + Q(y)$ equals the trace of W ; in particular, this expression does not depend on the choice of orthonormal vectors.

EXERCISE 4.6. Let $\mathcal{V} \subset \mathbb{R}^3$ be a two-dimensional subspace. Let $W : \mathcal{V} \rightarrow \mathcal{V}$ be a self-adjoint linear transformation. Define $K = \det(W)$ and $H = \frac{1}{2} \operatorname{trace}(W)$. Prove that $K \leq H^2$.

EXERCISE 4.7. Let $\mathcal{W}_p = \begin{pmatrix} 0 & 2 \\ 2 & 0 \end{pmatrix}$ be the linear transformation from the bottom-right example in Fig. 4.4 on page 200. Find an orthonormal basis of eigenvectors by explicitly performing the steps of the proof of Proposition 4.7. Compare with Eq. 4.1 on page 199.

EXERCISE 4.8. Let S be a regular surface, and $f : S \rightarrow \mathbb{R}$ a smooth function. With the terminology of Exercise 3.44 on page 146, let $p \in S$ be a critical point of S , and consider the **Hessian function** from $T_p S$ to \mathbb{R} defined as $v \mapsto \operatorname{Hess}(f)_p(v)$. Prove that this Hessian function is the quadratic form associated to a self-adjoint linear transformation $H_p : T_p S \rightarrow T_p S$.

The point p is called a **nondegenerate critical point** of f if H_p is invertible. In this case, prove that p is a local minimum of f if both eigenvalues of H_p are positive, a local maximum if both eigenvalues are negative, and a saddle point if the eigenvalues have different signs. Here “minimum,” “maximum,” and “saddle” are defined in the most natural manner that generalizes the case $S = \mathbb{R}^2$.



3. Normal Curvature

In this section, we define and study the second fundamental form and normal curvature. The story begins by applying the general linear algebra facts from the previous section to the Weingarten map.

DEFINITION 4.9.

Let S be an oriented regular surface, and let $p \in S$. Let $\{v_1, v_2\}$ be an orthonormal basis of $T_p S$ with respect to which the Weingarten map is represented by a diagonal matrix: $\mathcal{W}_p = \begin{pmatrix} k_1 & 0 \\ 0 & k_2 \end{pmatrix}$ (which exists by Proposition 4.7).

- (1) The eigenvectors $\pm v_1$ and $\pm v_2$ are called **principal directions** of S at p .
- (2) The eigenvalues k_1 and k_2 are called the **principal curvatures** of S at p . If $k_1 = k_2$, then p is called an **umbilical point**.
- (3) The quadratic form associated to \mathcal{W}_p is called the **second fundamental form** of S at p , and is denoted by II_p . That is,

$$II_p(v) = \langle \mathcal{W}_p(v), v \rangle = \langle -dN_p(v), v \rangle$$

for all $v \in T_p S$.

- (4) If $v \in T_p S$ with $|v| = 1$, then the value $II_p(v)$ is also called the **normal curvature** of S at p in the direction v .

If $k_1 \neq k_2$, then v_1 and v_2 are uniquely determined, except that either could be multiplied by -1 , which is why the “ \pm ” appears in (1) above. If p is an umbilical point ($k_1 = k_2$), then every orthonormal basis of $T_p S$ is a basis of eigenvectors of \mathcal{W}_p , so in this case, every unit vector in $T_p S$ qualifies as a principal direction.

According to Examples 4.2 and 4.3 (page 197), every point of a plane or a sphere is an umbilical point. In the graphs shown in Fig. 4.4 (page 200), the principal curvatures are the eigenvalues of the given matrices. For three of these four graphs (all except the one on the bottom right), $\pm e_1$ and $\pm e_2$ are principal directions.

According to Proposition 4.8 (page 203), II_p acts on an arbitrary unit vector in $T_p S$ as

$$(4.7) \quad II_p((\cos \theta)v_1 + (\sin \theta)v_2) = k_1 \cos^2 \theta + k_2 \sin^2 \theta.$$

In particular, the principal curvatures, k_1 and k_2 , are the minimum and maximum normal curvatures.

The Gaussian curvature, $K(p)$, equals the determinant of \mathcal{W}_p , which is just the product of the principal curvatures. The mean curvature, $H(p)$, equals half the trace of \mathcal{W}_p , which is the *average* of the principal curvatures (hence the name *mean* curvature):

$$\boxed{K(p) = k_1 k_2} \quad \text{and} \quad \boxed{H(p) = \frac{1}{2}(k_1 + k_2)}.$$

EXAMPLE 4.10 (The Curvature of a Cylinder). Denote the cylinder of radius r by

$$C(r) = \{(r \cos \theta, r \sin \theta, z) \in \mathbb{R}^3 \mid \theta \in [0, 2\pi), z \in \mathbb{R}\}.$$

The outward-pointing unit normal field is

$$N(p) = (\cos \theta, \sin \theta, 0) \text{ for all } p = (r \cos \theta, r \sin \theta, z) \in C(r).$$

Let $p_0 = (r \cos \theta_0, r \sin \theta_0, z_0) \in C_r$. Figure 4.5 leads one to guess that the extreme normal curvatures at p_0 occur in the “around” direction $v_1 = (-\sin \theta_0, \cos \theta_0, 0)$ and the “upward” direction $v_2 = (0, 0, 1)$, so we will express the Weingarten map with respect to this basis.

The curve $\gamma_1(t) = (r \cos(\theta_0 + t/r), r \sin(\theta_0 + t/r), z_0)$ in $C(r)$ satisfies $\gamma_1(0) = p_0$ and $\gamma'_1(0) = v_1$. The restriction of N to γ_1 is

$$N_1(t) = N(\gamma_1(t)) = (\cos(\theta_0 + t/r), \sin(\theta_0 + t/r), 0),$$

so

$$\mathcal{W}_{p_0}(v_1) = -N'_1(0) = -\frac{1}{r}v_1.$$

The curve $\gamma_2(t) = (r \cos \theta_0, r \sin \theta_0, z_0 + t)$ in $C(r)$ satisfies $\gamma_2(0) = p_0$ and $\gamma'_2(0) = v_2$. The restriction of N to γ_2 is $N_2(t) = N(\gamma_2(t)) = (\cos \theta_0, \sin \theta_0, 0)$. Since this is constant,

$$\mathcal{W}_{p_0}(v_2) = -N'_2(0) = (0, 0, 0) = \mathbf{0} \cdot v_2.$$

In summary, $v_1 = (-\sin \theta_0, \cos \theta_0, 0)$ and $v_2 = (0, 0, 1)$ are principal directions, while the principal curvatures are $k_1 = -\frac{1}{r}$ and $k_2 = 0$. The cylinder therefore has constant Gaussian curvature $k = k_1 k_2 = 0$ and constant mean curvature $H = \frac{1}{2}(k_1 + k_2) = -\frac{1}{2r}$.

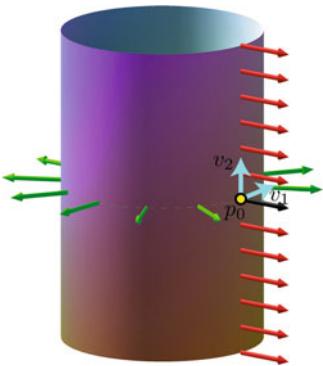


FIGURE 4.5. The outward normal field N restricted to curves through p in the principal directions v_1 and v_2

Recall from Example 3.63 on page 167 that the natural surface patch for the cylinder is an isometry. This isometry preserves Gaussian curvature ($K = 0$ for both the cylinder and the plane), but does not preserve principal curvatures (only the cylinder has one nonzero principal curvature). We will learn in Chap. 5 that every isometry preserves Gaussian curvature.

When $|v| = 1$, recall that the value $II_p(v)$ is called the “normal curvature of S at p in the direction of v .” It equals the component of $-N'(0)$ in

the direction v , where $N(t)$ is the restriction of N to a curve γ in S with $\gamma(0) = p$ and $\gamma'(0) = v$. Intuitively, this measures the rate at which I must tilt a flagpole forward/backward to keep its tip pointing directly up from the surface as I walk along γ . Since the normal field N is being differentiated, the term “normal curvature” might seem somewhat linguistically appropriate, but this is coincidental. Historically, the term “normal curvature” derived from an important alternative characterization of its value:

PROPOSITION 4.11 (Alternative Characterization of Normal Curvature).

Let S be an oriented regular surface, $p \in S$, and $v \in T_p S$ with $|v| = 1$. Consider the family of all regular curves γ in S with $\gamma(0) = p$ and $\gamma'(0) = v$.

- (1) For every curve γ in this family,

$$\langle \gamma''(0), N(p) \rangle = II_p(v).$$

That is, the normal curvature equals the normal component of the initial acceleration vector of γ .

- (2) The minimum curvature at p among curves in this family (regarded as space curves) equals $|II_p(v)|$.

An important (and perhaps surprising) consequence of (1) is that the value $\langle \gamma''(0), N(p) \rangle$ depends only on $v = \gamma'(0)$ and not on $\gamma''(0)$. The normal component of acceleration does *not* depend on the choice of γ from this family!

Part (2) says that normal curvature is a signed version of the “minimal curvature” measurement defined in Example 3.64 on page 168.

PROOF. For (1), since $\gamma'(t)$ is orthogonal to $N(\gamma(t))$ for all t , we have

$$0 = \frac{d}{dt} \Big|_{t=0} \langle \gamma'(t), N(\gamma(t)) \rangle = \langle \gamma''(0), N(p) \rangle + \langle v, -\mathcal{W}_p(v) \rangle.$$

Therefore, $\langle \gamma''(0), N(p) \rangle = \langle v, \mathcal{W}_p(v) \rangle = II_p(v)$.

For (2), the curvature, κ , at p of an arbitrary curve γ in this family equals the norm of the component of $\gamma''(0)$ orthogonal to v , so $\kappa \geq |\langle \gamma''(0), N(p) \rangle|$. It remains to prove that this lower bound is attained, or equivalently, that there exists a curve γ in this family for which $\gamma''(0)$ is parallel to $N(p)$. For this, consider the subspace $P = \text{span}\{v, N(p)\} \subset \mathbb{R}^3$. Exercise 3.69 on page 159 established that the intersection of $p + P$ with a sufficiently small neighborhood of p in S equals the trace of a regular curve γ . After γ is parametrized by arc length with $\gamma(0) = p$ and $\gamma'(0) = v$, it is called the **normal section** of S at p in the direction of v ; see Fig. 4.6 (left). Since γ lies in the plane $p + P$, its acceleration vector lies in P , and is therefore parallel to $N(p)$. Its curvature therefore equals

$$\kappa = |\gamma''(0)| = |\langle \gamma''(0), N(p) \rangle| = |II_p(v)|.$$

□

This proof accounts for the vocabulary—*the curvature of the normal section equals the absolute value of the normal curvature*. All other curves in the family have at least this much curvature, but there is no upper bound—curves in the family can be constructed with arbitrarily large curvature, like the “circles” of arbitrarily small radius indicated in Fig. 4.6 (right).

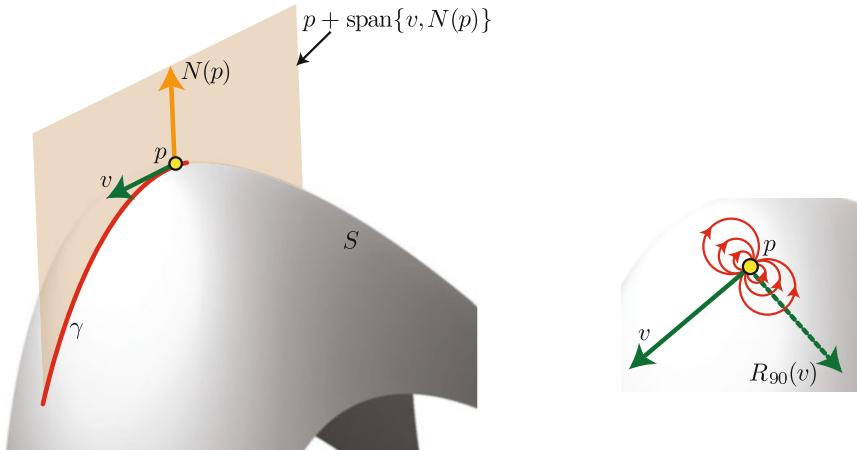


FIGURE 4.6. Among curves in S through p with initial velocity v , the curvature is minimized by a normal section (left) but has no maximum (right)

For the remainder of this section, let γ belong to the family of curves in the previous proposition. Our goal is to distinguish between the possibilities shown in Fig. 4.6 by defining a measurement of how sharply γ is turning left/right at p . For this, let $\kappa_n = II_p(v)$ denote the normal curvature, and assume for simplicity that γ is of unit speed, so that its initial acceleration vector, $a = \gamma''(0)$, is orthogonal to v . According to (1), this acceleration vector decomposes as

$$(4.8) \quad a = \kappa_n \cdot N(p) + \kappa_g \cdot R_{90}(v)$$

for some scalar $\kappa_g \in \mathbb{R}$, which we call the “**geodesic curvature of γ at p** ”; see Fig. 4.7. Here, $R_{90} : T_p S \rightarrow T_p S$ denotes the rotation by 90° in the direction that is counterclockwise with respect to the orientation, as defined in Equations 3.11 and 3.12 on page 150. To achieve more symmetric grammar, we will sometimes refer to κ_n as the “**normal curvature of γ at p** ,” even though it doesn’t depend on the choice of γ from the family.

Since the curvature, κ , at p of γ (regarded as a space curve) equals the norm of a , we have

$$(4.9) \quad \kappa^2 = \kappa_n^2 + \kappa_g^2.$$

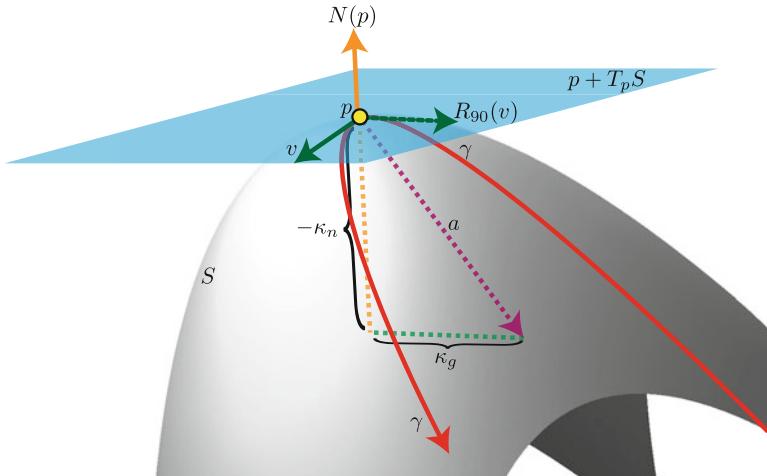


FIGURE 4.7. The normal and geodesic curvatures of γ at p measure the signed lengths of the components of the acceleration vector normal and tangent to S respectively

In particular, $\kappa \geq |\kappa_n|$, which we already learned in part (2) of the previous proposition. This quantifies the intuitive idea that the possible curvatures of regular curves in S are restricted by the curvature of S itself. In order to bend sharply enough to remain in S , γ must have curvature at least $|\kappa_n|$ at p .

On the other hand, the geodesic curvature is unrestricted:

PROPOSITION 4.12.

With the notation and assumptions of Proposition 4.11, all real numbers are attained as values of κ_g at p for curves γ in this family.

Although we will postpone a formal proof of this proposition until Exercise 5.56 (in Sect. 6 of Chap. 5), Fig. 4.6 indicates why it is believable. This figure could be recaptioned as follows: *among curves in the family, κ_g equals zero for the normal section (left), or is arbitrarily largely positive (respectively negative) for counterclockwise (respectively clockwise) circles of small radii.* But notice that the normal section is guaranteed to have vanishing geodesic curvature only at p . In the next chapter, we will do better by constructing curves in S (called geodesics) with everywhere vanishing geodesic curvature.

EXERCISES

EXERCISE 4.9. Prove that $K \leq H^2$ at every point of an oriented regular surface. *HINT: See Exercise 4.6 on page 205.*

EXERCISE 4.10. Prove that every ruled surface (Exercise 3.29 on page 139) satisfies $K \leq 0$.

EXERCISE 4.11. If γ is a unit-speed curve in S^2 , verify that its geodesic curvature function (with respect to the outward-pointing orientation) is

$$\kappa_g(t) = \langle \gamma(t) \times \gamma'(t), \gamma''(t) \rangle.$$

EXERCISE 4.12. For fixed $\phi_0 \in (0, \pi)$, let γ be a unit-speed parametrization of the ϕ_0 -latitudinal curve in the sphere S^2 (the θ -parameter curve in Example 3.24 on page 129 with $\phi = \phi_0$). With respect to the outward-pointing orientation of S^2 , show that the geodesic curvature of γ is constant at $\kappa_g = \cot(\phi_0)$.

EXERCISE 4.13. Is part (1) of Proposition 4.11 true when v has arbitrary length?

EXERCISE 4.14. In the cylinder of Example 4.10, compute the normal curvature at p_0 in an arbitrary direction via three methods:

- (1) Using Eq. 4.7.
- (2) By explicitly parametrizing a curve γ in an arbitrary direction (such as, for example, a helix) and...
 - (a) ...differentiating the restriction of N to γ .
 - (b) ...computing the normal component of $\gamma''(0)$.

EXERCISE 4.15. Let $\gamma : (-\epsilon, \epsilon) \rightarrow S$ be a regular curve in an oriented surface. For all $t \in (-\epsilon, \epsilon)$, let $\bar{\gamma}(t)$ be the projection of $\gamma(t) - \gamma(0)$ onto $T_{\gamma(0)}S$ (as defined in Sect. 2 of Chap. 1). Prove that the geodesic curvature of γ equals the signed curvature of $\bar{\gamma}$ at time 0. For “signed curvature” to make sense here, use a positively oriented orthonormal basis to identify $T_{\gamma(0)}S \cong \mathbb{R}^2$.

EXERCISE 4.16. Prove that every compact regular surface has a point of positive Gaussian curvature.

HINT: Let $p \in S$ be a point of maximum distance to the origin. By applying Exercise 1.43 on page 32 to a normal section, conclude that the normal curvature of S at p in every direction is $\geq \frac{1}{r}$, where r is the distance from p to the origin.

EXERCISE 4.17. Prove that for a unit-speed curve γ in an oriented surface S , we have $\kappa_n = \kappa \cos \theta$, where θ is the angle between the unit normal vector \mathbf{n} to γ (regarded as a space curve) and the unit normal vector N to S at p .

EXERCISE 4.18. A unit-speed curve $\gamma : I \rightarrow S$ in an oriented regular surface S is called a **line of curvature** if $\gamma'(t)$ is a principal direction for each $t \in I$.

- (1) Prove that every unit-speed curve in a plane or a sphere is a line of curvature.
- (2) Prove that every longitude and every latitude is a line of curvature of a surface of revolution (Example 3.25 on page 131).
- (3) Prove that γ is a line of curvature if and only if there exists a function $\lambda : I \rightarrow \mathbb{R}$ such that for all $t \in I$, $(N \circ \gamma)'(t) = \lambda(t)\gamma'(t)$.
- (4) Explain why the definition of “line of curvature” makes sense for a nonoriented surface.

EXERCISE 4.19. Let S be a path-connected oriented regular surface. If all points of S are umbilical points, prove that S is contained in either a plane or a sphere.

HINT: First prove that the principal curvature function $k = k_1 = k_2$ is constant on S . For this, let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a coordinate chart. At every $q \in U$, $N_u(q) = dN_{\sigma(q)}(\sigma_u(q)) = -k(q) \cdot \sigma_u(q)$. That is, $N_u = -k \cdot \sigma_u$ on all of U , and similarly, $N_v = -k \cdot \sigma_v$ on all of U . Differentiating this first equation with respect to v and the second with respect to u and subtracting yields $k_u \sigma_u = k_v \sigma_v$, which implies that $k_u = k_v = 0$; thus, k is constant. If $k \neq 0$, demonstrate that the point $\sigma + \frac{1}{k}N$ is constant on U , so V is contained in the sphere of radius $\frac{1}{k}$ about this point.

EXERCISE 4.20. At a point p of an oriented regular surface S , prove that the average of the normal curvatures at p in the directions of a pair of orthonormal vectors equals $H(p)$; in particular, this average does not depend on the pair of vectors. *HINT: Use Exercise 4.5(2) on page 205.*

EXERCISE 4.21. Let S be an oriented regular surface. A regular curve $\gamma : I \rightarrow S$ is called an **asymptotic** if $II_{\gamma(t)}(\gamma'(t)) = 0$ for all $t \in I$.

- (1) If γ has nonzero curvature, prove that it is an asymptotic if and only if its unit binormal vector $\mathbf{b}(t)$ is parallel to $N(\gamma(t))$ for all $t \in I$.
- (2) Explain why “asymptotic” makes sense for nonoriented surfaces.

EXERCISE 4.22. Let S be an oriented regular surface. For $c > 0$ consider the *dilation* map $d_c : \mathbb{R}^3 \rightarrow \mathbb{R}^3$ defined as $d_c(p) = cp$. Orient $\tilde{S} = d_c(S) = \{cp \mid p \in S\}$ as follows: $\tilde{N}(cp) = N(p)$ for all $p \in S$. How are the Gaussian curvature, principal curvatures, and principal directions of \tilde{S} related to those measurements on S ? What happens if $c < 0$?

EXERCISE 4.23. Let S be the torus obtained by revolving about the z -axis the circle in the xz -plane with radius 1 centered at $(2, 0, 0)$. This torus is illustrated in Fig. 4.8. Colored red (respectively green) is the region where $x^2 + y^2 < 4$ (respectively $x^2 + y^2 > 4$). Let N be the outward-pointing unit normal field on S .

- (1) Verify that the unit normal vector to every longitudinal curve at every point is $\mathbf{n} = -N$.

- (2) Verify that the unit normal vector \mathbf{n} to every latitudinal curve in the red (respectively green) region at every point makes an acute (respectively obtuse) angle with N .
- (3) Use Exercise 4.18(2) to conclude that $K > 0$ on the green region and $K < 0$ on the red region.
- (4) Verify that the Gauss map $N : S \rightarrow S^2$ restricted to either the red or green region is a diffeomorphism from that colored region to all of S^2 except the north and south poles.
- (5) Conclude that the integral over S of the Gaussian curvature equals zero; that is, $\iint_S K dA = 0$. *HINT: Apply Proposition 3.60 on page 164 to the restriction of the Gauss map to each colored region.*

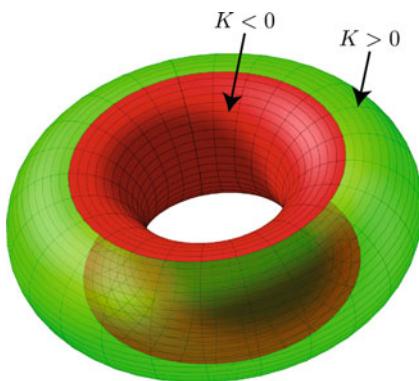


FIGURE 4.8. Regions of positive (green) and negative (red) Gaussian curvature on the torus of revolution

EXERCISE 4.24. Let γ be a unit-speed curve in an oriented surface S . How would its geodesic curvature κ_g be affected if the orientation of S were reversed? What if the orientation of γ were reversed?

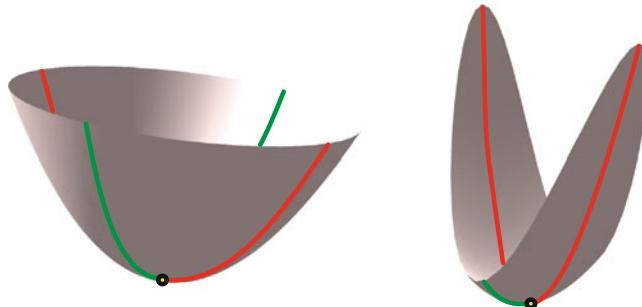
COMMENT: κ_g roughly measure how sharply γ curves “to the left,” but an orientation is necessary to make sense of the phrase “to the left.” This phrase presupposes that the curve is traversed by someone walking on the “top of” the surface, with his/her head pointing in the direction of the orientation. A person traversing the same curve walking along the other possible “top” of the surface would instead be turning to the right.

4. Geometric Characterizations of Gaussian Curvature

In this section, we discuss some geometric characterizations of the magnitude and the sign of the Gaussian curvature.

We previously advertised $K(p)$ as a measurement of “how sharply S curves away from $T_p S$ near p ” but that was an oversimplification. The value $K(p)$ is only a single number. Knowledge of both principal directions and both principal curvatures at p indicates more about how S curves away from $T_p S$ near p than any single number could indicate. For example, the plane and the cylinder both have constant Gaussian curvature zero (Examples 4.2 on page 197 and 4.10 on page 207). The obvious visual difference is explained by the cylinder’s having one nonzero principal curvature at each point. Similarly, Fig. 4.9 shows two graphs that (according to Example 4.4 on page 198) have

the same Gaussian curvature at the origin, yet have obvious visual differences accounted for by their different principal curvatures.



$$\begin{aligned} f &= x^2 + y^2 \\ k_1 &= 2, \quad k_2 = 2 \end{aligned}$$

$$\begin{aligned} f &= 3x^2 + (1/3)y^2 \\ k_1 &= 6, \quad k_2 = 2/3 \end{aligned}$$

FIGURE 4.9. Both graphs have Gaussian curvature 4 at the origin (but different principal curvatures)

In spite of these limitations, the Gaussian curvature is one of the most fundamental concepts in geometry. Much of the remainder of this book is devoted to exploring what can be done with it. We will begin this exploration now by mentioning the most obvious geometric interpretation of Gaussian curvature. For this, it is best to discuss separately its sign (positive or negative) and its absolute value.

We begin with the Gaussian curvature's sign. By definition, $K(p)$ is the determinant of the linear transformation $\mathcal{W}_p : T_p S \rightarrow T_p S$. According to the interpretation in Sect. 4 (Chap. 3) of the sign of the determinant of a linear transformation,

$$(4.10) \quad K(p) \text{ is } \left\{ \begin{array}{l} \text{positive} \\ \text{negative} \\ \text{zero} \end{array} \right\} \iff \mathcal{W}_p \text{ is } \left\{ \begin{array}{l} \text{an orientation-preserving isomorphism} \\ \text{an orientation-reversing isomorphism} \\ \text{not an isomorphism} \end{array} \right\}.$$

This fact about $T_p S$ can be upgraded to the following assertion about a neighborhood of p in S :

PROPOSITION 4.13.

Let S be an oriented regular surface, and let $p \in S$. If $K(p) \neq 0$, there exists a neighborhood of p in S restricted to which the Gauss map $N : S \rightarrow S^2$ is a diffeomorphism onto its image, and furthermore, $K(p)$ is $\begin{cases} \text{positive} \\ \text{negative} \end{cases}$ if and only if this diffeomorphism is $\begin{cases} \text{orientation-preserving} \\ \text{orientation-reversing} \end{cases}$ with respect to the given orientation of S and the outward-pointing orientation of S^2 .

PROOF. Since $K(p) = \det(\mathcal{W}_p) = \det(-dN_p) = \det(dN_p)$, we could replace \mathcal{W}_p with dN_p in Eq. 4.10. The claim that N is a diffeomorphism between neighborhoods follows immediately from the inverse function theorem for surfaces (Theorem 3.35 on page 143). Recall that the domain and codomain of $dN_p : T_p S \rightarrow T_{N(p)} S^2$ are the same vector space; that is, $T_p S = T_{N(p)} S^2$. The orientation $N(p)$ on this single vector space is simultaneously interpreted as the given orientation of S at p and as the outward-pointing orientation of S^2 at $N(p)$. After possibly shrinking the neighborhoods, we can assume that they are connected, so the result now follows from Proposition 3.52 on page 155. \square

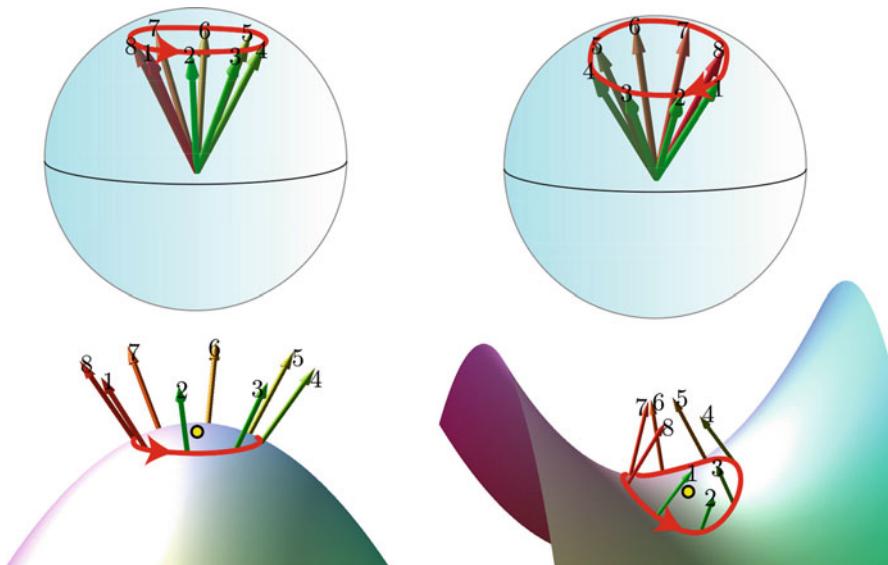


FIGURE 4.10. The sign of K determines whether the Gauss map preserves or reverses orientation

Figure 4.10 illustrates this proposition. When $K(p) \neq 0$, the Gauss map sends a small *counterclockwise* loop around p in S to a small loop in S^2 that is *counterclockwise* if $K(p) > 0$ or *clockwise* if $K(p) < 0$. In this figure, can you visualize why the conclusion would be the same with respect to the other orientation of each surface (but always using the outward-pointing orientation of S^2)?

To interpret the absolute value of the Gaussian curvature geometrically, notice that

$$(4.11) \quad |K(p)| = |\det(dN_p)| = \|dN_p\|,$$

which is the infinitesimal area distortion of the Gauss map at p . If $K(p) \neq 0$ and R is a sufficiently small polygonal region in S containing p , then Eq. 3.15 (page 164) gives

$$|K(p)| \approx \frac{\text{area}(N(R))}{\text{area}(R)}.$$

This conclusion can be visualized in Fig. 4.10 by imagining that the red loop in S is chosen small enough that its interior qualifies as such a region R .

We end this section with another characterization of the sign of the Gaussian curvature, which is particularly simple because it does not mention the Gauss map:

PROPOSITION 4.14.

Let S be a (not necessarily oriented) regular surface, and let $p \in S$. If $K(p) > 0$, then a sufficiently small neighborhood of p in S lies entirely on one side of the plane $p + T_p S$ (except for the point p itself, which lies in this plane). If $K(p) < 0$, then every neighborhood of p in S intersects both sides of $p + T_p S$.

PROOF. As in the proof of Lemma 4.5 on page 200, after applying a rigid motion, we can assume without loss of generality that $T_p S = \text{span}\{e_1, e_2\}$ and that a neighborhood of p in S equals the graph of a smooth function f . According to Example 4.4 on page 198, $K(p) = f_{xx}f_{yy} - f_{xy}^2$. The result now follows from the second derivative test from multivariable calculus, which classifies the critical point as a (strict) local extremum if $f_{xx}f_{yy} - f_{xy}^2 > 0$ and as a saddle point if $f_{xx}f_{yy} - f_{xy}^2 < 0$. \square

EXERCISES

EXERCISE 4.25. Let G denote the graph of $f(x, y) = xy$. Describe the intersection of G with its tangent plane at the origin.

EXERCISE 4.26. Let S be a *compact* regular surface. Denote the region of S with nonnegative Gaussian curvature by $S_+ = \{p \in S \mid K(p) \geq 0\}$. Prove that $N : S_+ \rightarrow S^2$ (the restriction of the Gauss map to S_+) is surjective.

HINT: For arbitrary $v_0 \in S^2$, consider a plane with normal vector v_0 that does not intersect S , and translate it in the direction of $\pm v_0$ until it first touches S .

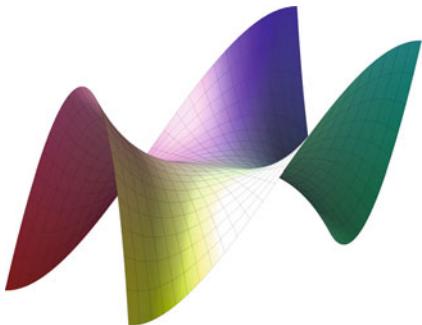


FIGURE 4.11. The monkey saddle
 $\{(u, v, u^3 - 3v^2u) \mid u, v \in \mathbb{R}^2\}$

EXERCISE 4.27. Let S be a compact regular surface with positive Gaussian curvature. Although we won't prove it here, this implies that S is **strictly convex**, which means that for all $p \in S$, the plane $p + T_p S$ intersects S only at p (the rest of S lies entirely on one side of this plane). Assuming this fact, prove:

- (1) The Gauss map $N : S \rightarrow S^2$ is a diffeomorphism.
- (2) $\int_S K dA = \text{Area}(S^2) = 4\pi$.

HINT: For (1), see the hint from the previous exercise. For (2), apply Proposition 3.60 on page 164 to the Gauss map. In Sect. 6 of Chap. 3, we defined integration only over certain types of regions. We will later learn that every compact regular surface qualifies, but for now observe that S qualifies because it's diffeomorphic to S^2 .

EXERCISE 4.28. Verify that the curvature at the origin of the **monkey saddle**, $S = \{(u, v, u^3 - 3v^2u) \mid u, v \in \mathbb{R}^2\}$, equals zero; see Fig. 4.11. In Proposition 4.14, can any conclusion be drawn when $K(p) = 0$? □

5. The Second Fundamental Form in Local Coordinates

In order to emphasize the geometric meaning of the second fundamental form and the Gaussian curvature, we have thus far worked directly from the definitions when computing these measurements, but this approach has computational limitations. For example, we were able to compute the curvature of a graph only at a critical point, not at a general point. In this section, we will greatly improve our computational prowess by deriving local coordinate expressions for the second fundamental form and the Gaussian curvature.

Throughout this section, let S be an oriented regular surface and let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch. Let N denote the orientation, and assume that the surface patch is compatible with the orientation.

Recall that in Sect. 9 of Chap. 3, we defined the functions $E, F, G : U \rightarrow \mathbb{R}$ as follows:

$$E = |\sigma_u|^2, \quad F = \langle \sigma_u, \sigma_v \rangle, \quad G = |\sigma_v|^2.$$

We then defined the *first fundamental form in the local coordinates $\{u, v\}$* (also called “the first fundamental form of σ ”) as the expression

$$\mathcal{F}_1 = E du^2 + 2F du dv + G dv^2.$$

Here we interpreted the expression “ du ” as the derivative of the function on U that maps $(u, v) \mapsto u$, and we similarly interpreted “ dv ” as the derivative of the function on U that maps $(u, v) \mapsto v$. So if $x = (a, b)$, then $du_q(x) = a$ and $dv_q(x) = b$ for every $q \in U$. With this understanding, the first fundamental form in local coordinates assigns to each $q \in U$ the function that sends $x \in T_q U = \mathbb{R}^2$ to $|d\sigma_q(x)|^2$. That is,

$$(\mathcal{F}_1)_q(x) = |d\sigma_q(x)|^2.$$

In other words, if you think of σ as identifying U with (part of) S , then \mathcal{F}_1 is the form on U that gets identified with the first fundamental form of S . The point was to pull information about S back to U . For example, given any curve γ in U , one can compute the length of the corresponding curve in S (namely $\sigma \circ \gamma$) just by knowing the first fundamental form in these local coordinates. We saw that areas and angles on S can similarly be computed on U using only \mathcal{F}_1 .

Our next goal is to express the *second* fundamental form in local coordinates. We will continue using the convention of denoting “ $N \circ \sigma$ ” simply by “ N ” whenever this does not cause confusion. Using this convention, we first define the functions $e, f, g : U \rightarrow \mathbb{R}$ as follows:

$$(4.12) \quad \begin{aligned} e &= \langle \sigma_{uu}, N \rangle = -\langle N_u, \sigma_u \rangle = \langle \mathcal{W}(\sigma_u), \sigma_u \rangle, \\ f &= \langle \sigma_{uv}, N \rangle = -\langle N_v, \sigma_u \rangle = -\langle N_u, \sigma_v \rangle = \langle \mathcal{W}(\sigma_u), \sigma_v \rangle, \\ g &= \langle \sigma_{vv}, N \rangle = -\langle N_v, \sigma_v \rangle = \langle \mathcal{W}(\sigma_v), \sigma_v \rangle. \end{aligned}$$

These equalities come from Equations 4.5 and 4.6 (page 205) and have input variables suppressed for brevity. For example, the first red equality really means that $e(q) = \langle \mathcal{W}_{\sigma(q)}(\sigma_u(q)), \sigma_u(q) \rangle$ for each $q \in U$, and similarly for the other expressions. Because of this relationship between the functions e, f, g and the Weingarten map, it is appropriate to make the following definition:

DEFINITION 4.15.

The **second fundamental form in the local coordinates $\{u, v\}$** (also called “the second fundamental form of σ ”) is the expression

$$\mathcal{F}_2 = e du^2 + 2f du dv + g dv^2.$$

This terminology is reasonable, for the following reason:

PROPOSITION 4.16.

If $q \in U$, $x \in T_q U = \mathbb{R}^2$, and we set $p = \sigma(q)$, then

$$(\mathcal{F}_2)_q(x) = II_p(d\sigma_q(x)).$$

In other words, if you think of σ as identifying U with (part of) S , then \mathcal{F}_2 is the form on U that gets identified with the second fundamental form on S .

PROOF. Express x in components as $x = (a, b)$. Recall that

$$du(x) = a, \quad dv(x) = b, \quad d\sigma_q(x) = a\sigma_u + b\sigma_v$$

(here and throughout this proof, all expressions are assumed to be evaluated at q). Thus

$$\begin{aligned} II_p(d\sigma_q(x)) &= II_p(a\sigma_u + b\sigma_v) = \langle \mathcal{W}_p(a\sigma_u + b\sigma_v), a\sigma_u + b\sigma_v \rangle \\ &= a^2 \langle \mathcal{W}_p(\sigma_u), \sigma_u \rangle + 2ab \langle \mathcal{W}_p(\sigma_u), \sigma_v \rangle + b^2 \langle \mathcal{W}_p(\sigma_v), \sigma_v \rangle \\ &= a^2 e + 2abf + b^2 g \\ &= e du(x)^2 + 2f du(x) dv(x) + g dv(x)^2 = (\mathcal{F}_2)_q(x). \end{aligned}$$

□

This proposition says that $(\mathcal{F}_2)_q$ expresses II_p in local coordinates. Recall that II_p is the quadratic form associated to the Weingarten map, \mathcal{W}_p . In particular, II_p and \mathcal{W}_p contain exactly the same information, just packaged in different ways; that is, either can be expressed in terms of the other, as described in Sect. 2. It must therefore be possible to describe a matrix for \mathcal{W}_p in terms of the functions e, f, g (and also the functions E, F, G because of the inner product that appears in Eq. 4.3 on page 204). The following proposition shows how it's done. All functions are assumed to be evaluated at $q = \sigma^{-1}(p)$.

PROPOSITION 4.17.

The matrix $\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ that represents \mathcal{W}_p with respect to the basis $\{\sigma_u, \sigma_v\}$ of $T_p S$ is given by

$$\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = \frac{1}{EG - F^2} \begin{pmatrix} eG - fF & fG - gF \\ fE - eF & gE - fF \end{pmatrix}.$$

In the very special case that $\{\sigma_u, \sigma_v\}$ is orthonormal (so $E = G = 1$ and $F = 0$), the formula from this proposition simplifies to $\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix} = \begin{pmatrix} e & f \\ f & g \end{pmatrix}$. This special case of the formula also follows immediately from Eq. 4.12. But in the general case, $\begin{pmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{pmatrix}$ is not necessarily a symmetric matrix, much less exactly equal to $\begin{pmatrix} e & f \\ f & g \end{pmatrix}$.

PROOF. From the red portions of Eq. 4.12,

$$e = \langle \mathcal{W}_p(\sigma_u), \sigma_u \rangle = \langle w_{11}\sigma_u + w_{21}\sigma_v, \sigma_u \rangle = w_{11}E + w_{21}F,$$

$$f = \langle \mathcal{W}_p(\sigma_u), \sigma_v \rangle = \langle w_{11}\sigma_u + w_{21}\sigma_v, \sigma_v \rangle = w_{11}F + w_{21}G,$$

$$f = \langle \mathcal{W}_p(\sigma_v), \sigma_u \rangle = \langle w_{12}\sigma_u + w_{22}\sigma_v, \sigma_u \rangle = w_{12}E + w_{22}F,$$

$$g = \langle \mathcal{W}_p(\sigma_v), \sigma_v \rangle = \langle w_{12}\sigma_u + w_{22}\sigma_v, \sigma_v \rangle = w_{12}F + w_{22}G.$$

This is a system of four linear equations and four variables $\{w_{11}, w_{12}, w_{21}, w_{22}\}$. Instead of solving it with a four-by-four matrix, it is simpler

to notice that the system is equivalent to the following equation involving two-by-two matrices:

$$\begin{pmatrix} e & f \\ f & g \end{pmatrix} = \begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}.$$

Therefore,

$$\begin{pmatrix} w_{11} & w_{21} \\ w_{12} & w_{22} \end{pmatrix} = \begin{pmatrix} e & f \\ f & g \end{pmatrix} \cdot \begin{pmatrix} E & F \\ F & G \end{pmatrix}^{-1}.$$

Notice that the left side is the transpose of the matrix we are after. The result follows from explicitly working out the matrix product on the right side. \square

The Gaussian curvature is the determinant of the matrix in Proposition 4.17, while the mean curvature is half the trace, which works out to

$$(4.13) \quad K = \frac{eg - f^2}{EG - F^2} \quad \text{and} \quad 2H = \frac{eG - 2fF + gE}{EG - F^2}.$$

The principal curvatures can be found by computing the eigenvalues of this matrix, but there is an alternative method that is generally simpler. Once Eq. 4.13 is used to determine K and H at a point, we know the sum and the product of the principal curvatures: $k_1 + k_2 = 2H$ and $k_1 k_2 = K$. Solving this system of two equations, we learn that the principal curvatures are

$$(4.14) \quad \{k_1, k_2\} = \left\{ H - \sqrt{H^2 - K}, H + \sqrt{H^2 - K} \right\}.$$

EXAMPLE 4.18 (A General Point of a Graph). Let $U \subset \mathbb{R}^2$ be open and $\varphi : U \rightarrow \mathbb{R}$ a smooth function. The graph, G , of φ is a regular surface covered by the single surface patch $\sigma : U \rightarrow G$ defined as $\sigma(x, y) = (x, y, \varphi(x, y))$. In Example 3.48 on page 153, we computed that

$$\sigma_x = (1, 0, \varphi_x), \quad \sigma_y = (0, 1, \varphi_y), \quad N = \frac{\sigma_x \times \sigma_y}{|\sigma_x \times \sigma_y|} = \frac{(-\varphi_x, -\varphi_y, 1)}{\sqrt{\varphi_x^2 + \varphi_y^2 + 1}}.$$

The coefficients of the first fundamental form are

$$E = \langle \sigma_x, \sigma_x \rangle = 1 + \varphi_x^2, \quad F = \langle \sigma_x, \sigma_y \rangle = \varphi_x \varphi_y, \quad G = \langle \sigma_y, \sigma_y \rangle = 1 + \varphi_y^2.$$

The second-order partial derivatives of σ are

$$\sigma_{xx} = (0, 0, \varphi_{xx}), \quad \sigma_{xy} = (0, 0, \varphi_{xy}), \quad \sigma_{yy} = (0, 0, \varphi_{yy}),$$

so the coefficients of the second fundamental form are

$$e = \frac{\varphi_{xx}}{\sqrt{1 + \varphi_x^2 + \varphi_y^2}}, \quad f = \frac{\varphi_{xy}}{\sqrt{1 + \varphi_x^2 + \varphi_y^2}}, \quad g = \frac{\varphi_{yy}}{\sqrt{1 + \varphi_x^2 + \varphi_y^2}}.$$

The Gaussian curvature at a general point is

$$K = \frac{eg - f^2}{EG - F^2} = \frac{\left(\frac{\varphi_{xx}\varphi_{yy} - \varphi_{xy}^2}{1 + \varphi_x^2 + \varphi_y^2} \right)}{(1 + \varphi_x^2)(1 + \varphi_y^2) - (\varphi_x \varphi_y)^2} = \frac{\varphi_{xx}\varphi_{yy} - \varphi_{xy}^2}{(1 + \varphi_x^2 + \varphi_y^2)^2}.$$

At a critical point, the denominator equals 1, so this formula for K reduces to the simpler formula from Example 4.4 on page 198. The mean curvature is

$$2H = \frac{eG - 2fF + gE}{EG - F^2} = \frac{\varphi_{xx}(1 + \varphi_y^2) - 2\varphi_{xy}\varphi_x\varphi_y + \varphi_{yy}(1 + \varphi_x^2)}{(1 + \varphi_x^2 + \varphi_y^2)^{3/2}}.$$

EXAMPLE 4.19 (A Surface of Revolution). In Example 3.25 on page 131, we studied the surface obtained by revolving the trace of $\gamma(t) = (x(t), 0, z(t))$ about the z -axis. We constructed the surface patch $\sigma(\theta, t) = (x(t)\cos(\theta), x(t)\sin(\theta), z(t))$, and we computed at an arbitrary point $q = (\theta, t)$ that

$$\sigma_\theta(q) = (-x\sin(\theta), x\cos(\theta), 0), \quad \sigma_t(q) = (x'\cos(\theta), x'\sin(\theta), z').$$

Here we are suppressing the input variable, for example by writing z' rather than $z'(t)$. We will assume that γ is parametrized by arc length, that is, that $(x')^2 + (z')^2 = 1$. The coefficients of the first fundamental form are

$$E = \langle \sigma_\theta, \sigma_\theta \rangle = x^2, \quad F = \langle \sigma_\theta, \sigma_t \rangle = 0, \quad G = \langle \sigma_t, \sigma_t \rangle = 1.$$

Since $\sigma_\theta \times \sigma_t = (xz' \cos \theta, xz' \sin \theta, -xx')$ and $|\sigma_\theta \times \sigma_t|^2 = (xz')^2 + (xx')^2 = x^2$, the unit normal field induced by this surface patch is

$$N = \frac{\sigma_\theta \times \sigma_t}{|\sigma_\theta \times \sigma_t|} = (z' \cos \theta, z' \sin \theta, -x').$$

When $\theta = 0$, notice that $N = (z', 0, -x')$ is the normal vector to the generating curve, obtained by rotating $\gamma' = (x', 0, z')$ clockwise 90° in the xz -plane. The second-order partial derivatives are

$$\sigma_{\theta\theta} = (-x \cos \theta, -x \sin \theta, 0), \quad \sigma_{\theta t} = (-x' \sin \theta, x' \cos \theta, 0), \quad \sigma_{tt} = (x'' \cos \theta, x'' \sin \theta, z''),$$

so the coefficients of the second fundamental form are

$$e = -xz', \quad f = 0, \quad g = x''z' - x'z''.$$

In summary, the first and second fundamental forms in the local coordinates $\{\theta, t\}$ are

$$\mathcal{F}_1 = x^2 d\theta^2 + dt^2, \quad \mathcal{F}_2 = (-xz') d\theta^2 + (x''z' - x'z'') dt^2.$$

The Gaussian curvature equals

$$\begin{aligned} K &= \frac{eg - f^2}{EG - F^2} = \frac{-xz'(x''z' - x'z'')}{x^2} = \frac{-x''(z')^2 + x'z'z''}{x} = \frac{-x''(z')^2 + x'(-x'x'')}{x} \\ &= \frac{-x''((z')^2 + (x')^2)}{x} = -\frac{x''}{x}. \end{aligned}$$

The equality of the red expressions follows from differentiating $(x')^2 + (z')^2 = 1$. In summary,

$$(4.15) \quad K = -\frac{x''}{x}.$$

The mean curvature equals

$$2H = \frac{eG - 2fF + gE}{EG - F^2} = \frac{-xz' + x^2(x''z' - x'z'')}{x^2} = \frac{-z' + x(x''z' - x'z'')}{x}.$$

The principal curvatures could be calculated from Eq. 4.14, but it is better instead to use the logic from which this equation was derived. Since $f = F = 0$, Eq. 4.13 simplifies to

$$K = \frac{eg}{EG}, \quad 2H = \frac{eG + gE}{EG},$$

so one can identify by sight the pair of numbers whose product is K and whose sum is $2H$, namely

$$(4.16) \quad \{k_1, k_2\} = \left\{ \frac{e}{E}, \frac{g}{G} \right\} = \left\{ -\frac{z'}{x}, \frac{x''z' - x'z''}{1} \right\}.$$

EXAMPLE 4.20 (Fake Spheres). We wish to construct surfaces of revolution with constant Gaussian curvature equal to 1. If $K = 1$, then Eq. 4.15 says that $x'' = -x$. The following is a solution for every $a > 0$:

$$(4.17) \quad x(t) = a \cos(t).$$

The hypothesis that γ is parametrized by arc length, so $(x')^2 + (z')^2 = 1$, means that z is determined by x as follows:

$$(4.18) \quad z(t) = \int_0^t \sqrt{1 - x'(s)^2} ds = \int_0^t \sqrt{1 - a^2 \sin^2(s)} ds.$$

We will choose the domain, I , of the generating curve $\gamma(t) = (x(t), 0, z(t))$ to be the largest possible interval about 0. More precisely, if $0 < a < 1$, then we choose $I = (-\pi/2, \pi/2)$ to ensure that $x(t) > 0$, so that the generating curve does not intersect with the axis of revolution. If $a > 1$, then we choose $I \subset (-\pi/2, \pi/2)$ to be the largest interval on which the expression under the square root in Eq. 4.18 is nonnegative.

The choice $a = 1$ gives $x(t) = \cos(t)$, $z(t) = \sin(t)$, and $I = (-\pi/2, \pi/2)$, which generates the sphere S^2 . This is the only choice of a for which the surface closes up into a compact regular surface. For other choices of a , Eq. 4.18 cannot be simplified in terms of elementary functions, but is nonetheless a smooth function on I . The choices $a = \frac{1}{2}$ and $a = 2$ are illustrated in Fig. 4.12. These fake spheres have no umbilical points (Exercise 4.36), which makes them genuinely different from S^2 (even locally). In particular, a rigid motion of space could not match any neighborhood of a fake sphere with any neighborhood of S^2 . Did you expect surfaces with constant curvature 1 to exist that are genuinely different from S^2 in this sense? But we will later learn in Sect. 3 of Chap. 5 that each fake sphere is locally isometric to S^2 for the same reason that a cylinder is locally isometric to a plane; namely, every pair of surfaces with the same constant Gaussian curvature must be locally isometric. Do you think that a compact surface with constant curvature 1 could be genuinely different from S^2 ?

Our primary reason for computing the coefficients $\{e, f, g\}$ of \mathcal{F}_2 in this section was as a step toward computing the Gaussian, mean, and principal curvatures in particular examples. But it is worth ending this section by mentioning that \mathcal{F}_2 itself has the following visual interpretation:

PROPOSITION 4.21.

If $q \in U$, $p = \sigma(q)$, and $x \in T_q U = \mathbb{R}^2$, then

$$\langle \sigma(q + tx) - \sigma(q), N(p) \rangle \approx \frac{t^2}{2} (\mathcal{F}_2)_q(x)$$

is a good approximation in the sense that the norm of the difference between the left and right sides is equal to an error function $E(t)$ with the property that $\lim_{t \rightarrow 0} \frac{E(t)}{t^2} = 0$.

PROOF. Express the components of q as $q = (u_0, v_0)$ and express the components of x as $x = (du_q(x), dv_q(x)) = (a, b)$. Using the second-order Taylor approximation for σ at q (Eq. 3.6 on page 118), we have

$$\begin{aligned} \langle \sigma(q + tx) - \sigma(q), N(p) \rangle &= \langle \sigma(u_0 + ta, v_0 + tb) - \sigma(u_0, v_0), N(p) \rangle \\ &\approx \left\langle t \underbrace{(a\sigma_u(q) + b\sigma_v(q))}_{\text{orthogonal to } N(p)} + \frac{t^2}{2} (a^2\sigma_{uu}(q) + 2ab\sigma_{uv}(q) + b^2\sigma_{vv}(q)), N(p) \right\rangle \\ &= \frac{t^2}{2} \langle du_q(x)^2 \sigma_{uu}(q) + 2du_q(x)dv_q(x)\sigma_{uv}(q) + dv_q(x)^2 \sigma_{vv}(q), N(p) \rangle \\ &= \frac{t^2}{2} (du_q(x)^2 e(q) + 2du_q(x)dv_q(x)f(q) + dv_q(x)^2 g(q)) = \frac{t^2}{2} (\mathcal{F}_2)_q(x). \end{aligned}$$

□

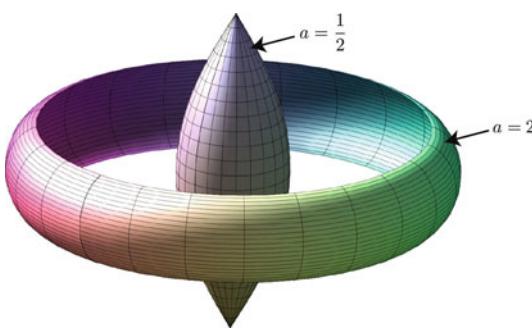


FIGURE 4.12. Fake spheres: surfaces of revolution with constant Gaussian curvature 1

$\frac{t^2}{2} (\mathcal{F}_2)_q(x)$ approximates the “height” above $T_p S$ of the point of S corresponding to a displacement in U away from q a distance t in the direction of x ; see Fig. 4.13.

Recall from Proposition 4.16 that $(\mathcal{F}_2)_q(x) = II_p(d\sigma_q(x))$; this says that \mathcal{F}_2 is a local coordinate expression for the second fundamental form, which was originally defined in a geometric manner that did not require the mention of any coordinate chart. By contrast, Proposition 4.21 is entirely about the coordinate chart σ . It says that

EXERCISES

EXERCISE 4.29. Use Proposition 4.21 to give an alternative proof of Proposition 4.14 on page 216.

EXERCISE 4.30. Show that the mean curvature of the surface $z = xy$ at the origin equals zero.

EXERCISE 4.31. Exercise 4.23 on page 212 described and illustrated the regions of positive and negative Gaussian curvature for a torus of revolution. Use Eq. 4.15 to give an alternative proof that this description is valid.

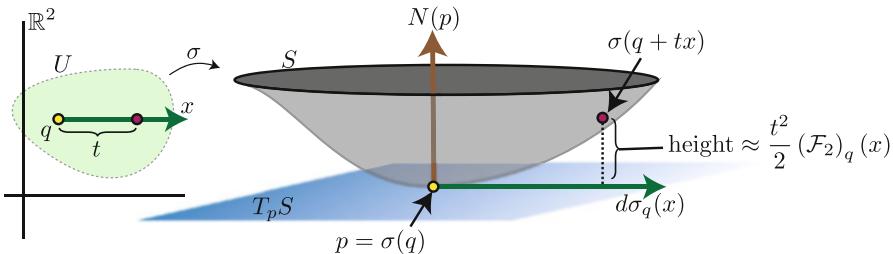


FIGURE 4.13. $\frac{t^2}{2} (\mathcal{F}_2)_q(x)$ approximates the “height” above $T_p S$ of the point of S corresponding to a displacement in U away from q a distance t in the direction of x

EXERCISE 4.32. Prove that the mean curvature $H(p)$ at a point p of an oriented regular surface S equals the average value of the normal curvature at p over the circle of unit-length directions in $T_p S$. *HINT: Use Exercise 4.5(1) on page 205.*

EXERCISE 4.33. Prove that the coordinate lines of a surface patch are **lines of curvature** (as defined in Exercise 4.18 on page 211) if and only if $f = F = 0$.

EXERCISE 4.34. Classify the surfaces of revolution with constant Gaussian curvature zero.

EXERCISE 4.35. (*Requires differential equations*) Classify the surfaces of revolution with constant Gaussian curvature -1 . Verify that the pseudosphere from Exercise 4.44 is included among the solutions.

EXERCISE 4.36. In Example 4.20, use Eq. 4.16 to verify that the fake spheres with $a = 2$ and $a = \frac{1}{2}$ have no umbilical points.

EXERCISE 4.37. Let S be an oriented regular surface and let $p \in S$. Define the “height” function $f : S \rightarrow \mathbb{R}$ such that $f(q) = \langle q - p, N(p) \rangle$ for all $q \in S$. With notation as in Exercise 3.44 (page 146), prove that p is a critical point of f and that $\text{Hess}(f)_p(v) = II_p(v)$ for all $v \in T_p S$.

EXERCISE 4.38. Equation 4.16 describes the principal curvatures of a surface of revolution as

$$\{k_1, k_2\} = \left\{ -\frac{z'}{x}, x''z' - x'z'' \right\}.$$

Which of these two values corresponds to the principal direction $\frac{\sigma_\theta}{|\sigma_\theta|}$ (tangent to the latitude) and which corresponds to the principal direction $\frac{\sigma_t}{|\sigma_t|}$ (tangent to the longitude)?

EXERCISE 4.39. Prove that a **tangent developable** (Exercise 3.24 on page 138 and Exercise 3.111 on page 188) has constant Gaussian curvature zero. Also compute its mean curvature.

Use a computer algebra application as needed for the calculations in the remaining exercises from this section.

EXERCISE 4.40. Calculate the Gaussian curvature and principal curvatures of the **helicoid** from Example 3.103 on page 186. In particular, verify that $K < 0$, and also that K depends only on the distance, t , to the z -axis, and is monotonically increasing in t (that is, K strictly increases towards zero as $t \rightarrow \infty$).

EXERCISE 4.41. The function $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as

$$\sigma(u, v) = \left(u - \frac{u^3}{3} + uv^2, v - \frac{v^3}{3} + vu^2, u^2 - v^2 \right)$$

is called **Enneper's surface**. Exercise 3.27 (page 139) verified that σ is a parametrized surface, and Exercise 3.98 (page 182) proved that σ is conformal.

- (1) Calculate the coefficients E, F, G of the first fundamental form in the local coordinates u, v .
- (2) Calculate the coefficients e, f, g of the second fundamental form in the local coordinates u, v .
- (3) Calculate the principal curvatures and the Gaussian curvature.

EXERCISE 4.42. Compute the Gaussian curvature of the **Möbius strip** defined by Eq. 3.14 on page 156.

EXERCISE 4.43. Let $\gamma_1, \gamma_2 : I \rightarrow \mathbb{R}^3$ be a pair of regular space curves with the same domain, I . Define $\sigma : I \times (0, 1) \rightarrow \mathbb{R}^3$ as

$$\sigma(t, u) = u \cdot \gamma_1(t) + (1 - u) \cdot \gamma_2(t).$$

Assume that γ_1, γ_2 are chosen such that σ is regular. Prove that $K(t, 1/2) = 0$ for all $t \in I$.

EXERCISE 4.44. The tractrix, $\gamma(t) = (\sin(t), 0, \cos(t) + \ln(\tan(t/2)))$, $t \in (\pi/2, \pi)$, was illustrated and studied in Exercise 1.4 on page 6. The surface of revolution obtained by revolving the trace of γ about the z -axis (as in Example 3.25 on page 131) is called the **pseudosphere**, illustrated in Fig. 4.14. Prove that it has constant Gaussian curvature $K = -1$.

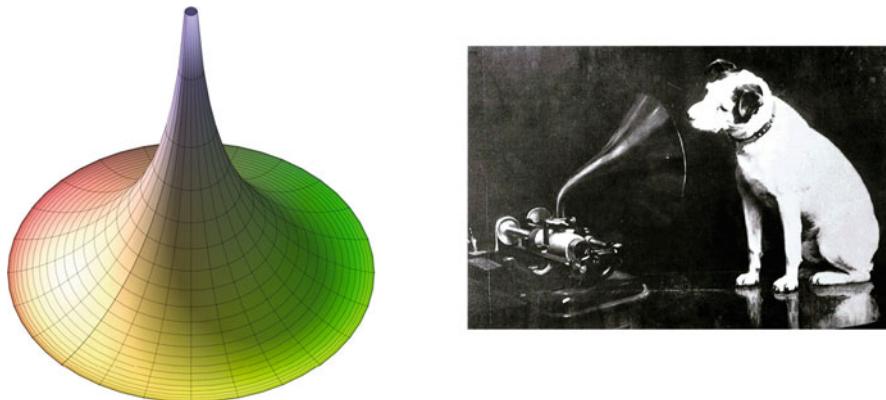


FIGURE 4.14. The pseudosphere is the surface of revolution generated by the tractrix. Horn loudspeakers designed in this shape have certain acoustic advantages. *Right:* Francis Barraud's photograph of Nipper looking into an Edison–Bell phonograph

EXERCISE 4.45. The monkey saddle was defined in Exercise 4.28 on page 217 as the graph of $f(u, v) = u^3 - 3uv^2$.

- (1) Verify that the Gaussian curvature at an arbitrary point of this graph is given by

$$K = \frac{-36(u^2 + v^2)}{(1 + 9u^4 + 18u^2v^2 + 9v^4)^2}.$$

- (2) Verify that this formula for K is invariant under a rotation of the uv -plane about the origin by any angle, but that the function f does not have this invariance.

EXERCISE 4.46 (Generalized Helicoids). Let $\gamma(t) = (x(t), 0, z(t))$, $t \in (a, b)$, be a regular curve. Recall from Example 3.25 on page 131 that revolving the trace, C , of γ about the z -axis yields a surface of revolution parametrized as

$$\sigma(\theta, t) = (x(t) \cos \theta, x(t) \sin \theta, z(t)).$$

For a constant $c \geq 0$, consider the following generalization:

$$\sigma_c(\theta, t) = (x(t) \cos \theta, x(t) \sin \theta, z(t) + c\theta),$$

which parameterizes the surface obtained by revolving C about the z -axis while simultaneously translating it in the positive z -direction. The parameterized surface σ_c is called a “generalized helicoid,” because the case $x(t) = t$ and $z(t) = 0$ is the helicoid from Exercise 3.103 on page 186, although it would be equally appropriate to call it a “generalized surface of revolution.”

- (1) Find a formula for the Gaussian curvature of σ_c .
- (2) If $x(t) = \cos(t)$ and $z(t) = \sin(t)$, $t \in (-\pi, \pi)$, then σ_c is called a **twisted sphere**, as illustrated in Fig. 4.15 (left). Verify that the Gaussian curvature of the twisted sphere is *not* constant.
- (3) If γ is the tractrix from Exercise 4.44, then σ_c is called a **twisted pseudosphere**, or **Dini's surface**, as illustrated in Fig. 4.15 (middle). Verify that this surface has constant negative Gaussian curvature.
- (4) If $x(t) = t$ and $z(t) = t\sqrt{a^2 - \frac{c^2}{t^2}} + c \cdot \arcsin\left(\frac{c}{at}\right)$ for some constant $a > 0$, prove that this surface has constant zero Gaussian curvature. An example is illustrated in Fig. 4.15 (right).

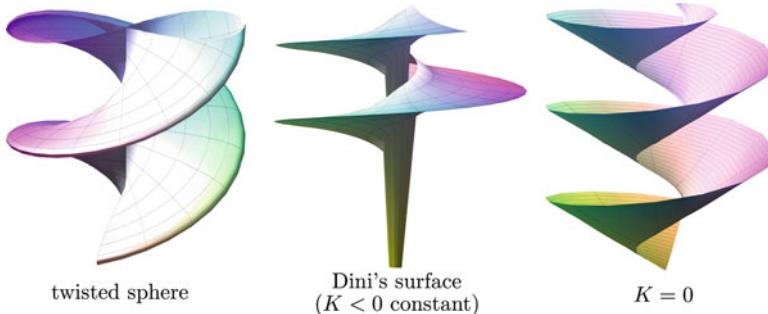


FIGURE 4.15. Generalized helicoids

□

6. Minimal Surfaces (Optional)

The goal of this section is to provide a geometric interpretation of mean curvature and to construct a few examples of minimal surfaces (surfaces whose mean curvature is everywhere zero). The significance of mean curvature stems from its appearance in the following formula, which describes how the area of a polygonal region is affected when S is deformed in the normal direction through a distance that changes from point to point according to an arbitrary smooth function φ :

PROPOSITION 4.22.

Let S be a regular surface, N an orientation of S , and $\varphi : S \rightarrow \mathbb{R}$ a smooth function. For $t \in \mathbb{R}$, define

$$S_t = \{p + t\varphi(p)N(p) \mid p \in S\}.$$

Let $R \subset S$ be a polygonal region. Setting $R_t = \{p + t\varphi(p)N(p) \mid p \in R\}$ and $A(t) = \text{Area}(R_t)$, we have

$$A'(0) = - \iint_R 2\varphi H \, dA.$$

This proposition is illustrated in Fig. 4.16. Recall from Exercise 3.68 (page 159) that S_t is a regular surface for values of t in a sufficiently small interval I about 0, provided that φ has compact support. In this case, the family $\{S_t \mid t \in I\}$ is a **deformation** of S , which means a smooth one-parameter family of regular surfaces beginning at $S_0 = S$. Even if φ does not have compact support, the proposition remains valid; it doesn't matter whether all of S_t is a regular surface, since R_t alone equals a polygonal region of a regular surface for sufficiently small t , so its area is a meaningful measurement.

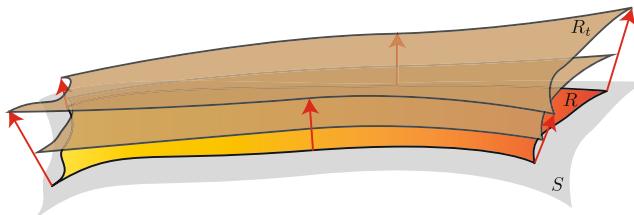


FIGURE 4.16. R_t is obtained by moving each point $p \in R$ a distance $t\varphi(p)$ in the normal direction

PROOF. Let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch that covers R . Denote the first and second fundamental forms of σ by

$$\mathcal{F}_1 = E \, du^2 + 2F \, du \, dv + G \, dv^2 \quad \text{and} \quad \mathcal{F}_2 = e \, du^2 + 2f \, du \, dv + g \, dv^2.$$

Set $T = \sigma^{-1}(R) \subset U$. Recall that the area of R equals

$$\text{Area}(R) = \iint_R 1 \, dA = \iint_T \sqrt{EG - F^2} \, dA.$$

As usual, we will sometimes denote by the same name a function on S and its composition with σ , whenever this does not risk causing confusion; in particular, we will sometimes regard N , φ , and H as functions on U . For $q \in T$ and $t \in \mathbb{R}$, we define

$$(4.19) \quad \sigma_t(q) = \sigma(q) + t\varphi(q)N(q),$$

so that σ_t is a surface patch for (part of) S_t , and $R_t = \sigma_t(T)$.

Notice that

$$(\sigma_t)_u = \sigma_u + t\varphi N_u + t\varphi_u N \quad \text{and} \quad (\sigma_t)_v = \sigma_v + t\varphi N_v + t\varphi_v N.$$

Therefore, the coefficients of the first fundamental form of σ_t are

$$\begin{aligned} E_t &= E + 2t\varphi \underbrace{\langle \sigma_u, N_u \rangle}_{-e} + t^2\varphi^2 \langle N_u, N_u \rangle + t^2\varphi_u\varphi_u, \\ F_t &= F + 2t\varphi \underbrace{(\langle \sigma_u, N_v \rangle + \langle \sigma_v, N_u \rangle)}_{-2f} + t^2\varphi^2 \langle N_u, N_v \rangle + t^2\varphi_u\varphi_v, \\ G_t &= G + 2t\varphi \underbrace{\langle \sigma_v, N_v \rangle}_{-g} + t^2\varphi^2 \langle N_v, N_v \rangle + t^2\varphi_v\varphi_v. \end{aligned}$$

Thus, using the fact that $H = \frac{eG - 2fF + gE}{2(EG - F^2)}$ from Eq. 4.13 (page 220), we have

$$\begin{aligned} E_t G_t - (F_t)^2 &= EG - F^2 - 2t\varphi(eG - 2fF + gE) + \mathcal{E}(t) \\ &= (EG - F^2) \left((1 - 4t\varphi) \left(\frac{eG - 2fF + gE}{2(EG - F^2)} \right) \right) + \mathcal{E}(t) \\ &= (EG - F^2)(1 - 4t\varphi H) + \mathcal{E}(t), \end{aligned}$$

where $\mathcal{E}(t)$ is an error term with $\lim_{t \rightarrow 0} \frac{\mathcal{E}(t)}{t} = 0$. Thus,

$$\begin{aligned} A'(0) &= \frac{d}{dt} \Big|_{t=0} \iint_T \sqrt{E_t G_t - (F_t)^2} \, dA \\ &= \iint_T \left(\frac{d}{dt} \Big|_{t=0} \sqrt{(EG - F^2)(1 - 4t\varphi H) + \mathcal{E}(t)} \right) \, dA \\ &= \iint_T (-2\varphi H \sqrt{EG - F^2}) \, dA = - \iint_R 2\varphi H \, dA. \end{aligned}$$

□

COROLLARY 4.23.

If $p \in S$ satisfies $H(p) \neq 0$, then there exists an area-decreasing deformation of S near p .

More precisely, for every sufficiently small polygonal region R of S containing p , there exists a one-parameter family of regular surfaces formed by replacing R with polygonal regions R_t that have the same boundary as R but less area than R ; see Fig. 4.17.

PROOF. Let $R \subset V$ be a polygonal region containing p that is small enough that H has the same sign on all of R . It is possible to choose a smooth function $\varphi : S \rightarrow \mathbb{R}$ that equals 0 on $S - R$ but is nonzero on at least some points of R . We can further choose φ to have the same sign as H at the points of R where it is nonzero. For the deformation of S described in Proposition 4.22, $A'(0) < 0$, so R_t has less area than R for small t . □

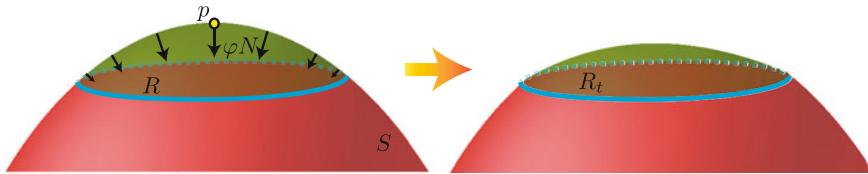


FIGURE 4.17. If $H(p) \neq 0$, then a sufficiently small polygonal region R containing p can be smoothly replaced by a smaller-area polygonal region R_t that has the same boundary

It is often convenient to consider the following vector-valued version of mean curvature:

DEFINITION 4.24.

The **mean curvature field** of an oriented regular surface S is the normal field $\mathbf{H} : S \rightarrow \mathbb{R}^3$ defined such that for all $p \in S$,

$$\mathbf{H}(p) = H(p)N(p).$$

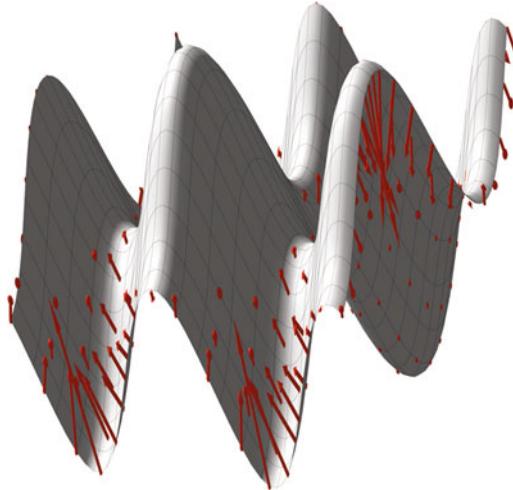


FIGURE 4.18. The mean curvature field for the graph of $z = \sin x + \cos y$

One advantage of this definition is that \mathbf{H} is independent of the choice of orientation of S . The other orientation would reverse the sign of both H and N , and therefore leave \mathbf{H} unchanged. Thus, \mathbf{H} is a well-defined normal field on even a nonorientable surface. Geometrically, if $\mathbf{H}(p) \neq \mathbf{0}$, then it points in the direction in which one should deform the surface near p in order to decrease area, as is apparent in Fig. 4.18. In fact, this observation helps

remove the apparent dichotomy between the two cases ($H > 0$ and $H < 0$) in the above proof of Corollary 4.23. In both cases, the area-decreasing deformation of S near p moves in the direction of a positive scalar multiple of \mathbf{H} , as can be seen by rewriting Eq. 4.19 as

$$\sigma_t = \sigma + t\varphi N = \sigma + t \underbrace{\frac{\varphi}{H} \mathbf{H}}_{\geq 0}$$

Corollary 4.23 helps us understand the bubble surfaces that are formed when metal frames are dipped into soap solution, as in Fig. 4.19. What optimization problem is nature solving here? Soap film acts like elastic stretched taut. It wants to get smaller; that is, it wants to decrease its surface area. It is forced to assume a surface whose boundary is the metal frame, because molecular bonds prevent it from letting go of the frame. Given this constraint, surface tension tends to cause the soap solution to quickly snap into a shape that has less area than other possible surfaces with this same boundary. Although there is no a priori guarantee that the soap solution will find a *global* minimum of this optimization problem, one does expect it to find at least a *local* minimum—a surface that does not have any deformations (of the type discussed above) with $A'(0) < 0$. By Corollary 4.23, such a surface must have mean curvature everywhere zero:

DEFINITION 4.25.

*A (regular or parametrized) surface is called **minimal** if its mean curvature is everywhere zero.*



FIGURE 4.19. Soap solution minimal surfaces

Minimal surfaces represent one of the most extensively studied topics within differential geometry, with deep connections to physics, complex analysis, topology, and differential equations. Some of the research in this area has been motivated by **Plateau's problem**, which roughly asks whether for every simple closed space curve C , there exists an area-minimizing surface with C as its boundary. Experiments dipping curved wire loops into soap solution, like the middle image of Fig. 4.19, might lead one to hypothesize an

affirmative answer. Although there are technicalities involved in formulating this problem with sufficient precision, the answer did essentially turn out to be yes for a large class of simple closed space curves.

Some of the more recent literature is related to modeling how soap solution quickly “snaps” into the shape of a minimal surface after a wire frame is pulled out of a bucket of soap solution. This process can be modeled as an area-decreasing deformation of the initial soap surface called **mean curvature flow**. Surface tension acts in such a way that at each time, each droplet of the soap solution moves in the normal direction at a speed proportional to the mean curvature of the current surface shape at that point. One usually takes the constant of proportionality to be 1 for simplicity. In other words, the droplet’s velocity vector equals the mean curvature vector of the current surface shape. This is *initially* the same as setting $\varphi = H$ in Proposition 4.22, but then it then evolves in a manner dependent on the mean curvature fields of the later surfaces rather than being fully determined by the initial surface. Mean curvature flow is the subject of a large body of modern differential geometry research.

A plane is obviously a minimal surface. Our next goal is to identify a few others. The following general observation is useful for confirming that certain other surfaces are minimal:

LEMMA 4.26.

A conformal parametrized surface $\sigma : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is minimal if and only if $\sigma_{uu} + \sigma_{vv} = 0$.

PROOF. Since σ is conformal, we have $F = 0$ and $E = G$ (Exercise 3.108 on page 188). By Eq. 4.13 (on page 220), $2H = \frac{e+g}{G}$, which vanishes if and only if

$$0 = e + g = \langle \sigma_{uu} + \sigma_{vv}, N \rangle.$$

To prove the lemma, it will therefore suffice to establish that $\sigma_{uu} + \sigma_{vv}$ is a normal field, that is, that it has no tangential component. For this, rewrite the equations $E = G$ and $F = 0$ as

$$\langle \sigma_u, \sigma_u \rangle = \langle \sigma_v, \sigma_v \rangle \quad \text{and} \quad \langle \sigma_u, \sigma_v \rangle = 0.$$

By taking partial derivatives of these equation with respect to u and v , we learn that

$$\langle \sigma_{uu}, \sigma_u \rangle = \langle \sigma_{vv}, \sigma_v \rangle = -\langle \sigma_u, \sigma_{vv} \rangle,$$

so $\langle \sigma_{uu} + \sigma_{vv}, \sigma_u \rangle = 0$, and similarly, $\langle \sigma_{uu} + \sigma_{vv}, \sigma_v \rangle = 0$. \square

EXAMPLE 4.27 (The Catenoid Is Minimal). *The catenoid is the surface obtained by revolving about the z -axis the trace of $\gamma(t) = (c \cosh(t), 0, ct)$, where $c > 0$ is a constant; see Fig. 4.20. According to Example 3.25 (on page 131), this surface can be parametrized as*

$$(4.20) \quad \sigma(\theta, t) = (c \cosh t \cos \theta, c \cosh t \sin \theta, ct).$$

The first-order partial derivatives are

$$\begin{aligned}\sigma_\theta &= (-c \cosh t \sin \theta, c \cosh t \cos \theta, 0), \\ \sigma_t &= (c \sinh t \cos \theta, c \sinh t \sin \theta, c).\end{aligned}$$

Therefore, $E = G = c^2 \cosh^2 t$ and $F = 0$, which means that σ is conformal. The second-order partial derivatives are

$$\begin{aligned}\sigma_{\theta\theta} &= (-c \cosh t \cos \theta, -c \cosh t \sin \theta, 0), \\ \sigma_{tt} &= (c \cosh t \cos \theta, c \cosh t \sin \theta, 0).\end{aligned}$$

Since $\sigma_{\theta\theta} = -\sigma_{tt}$, Lemma 4.26 implies that the catenoid is a minimal surface.

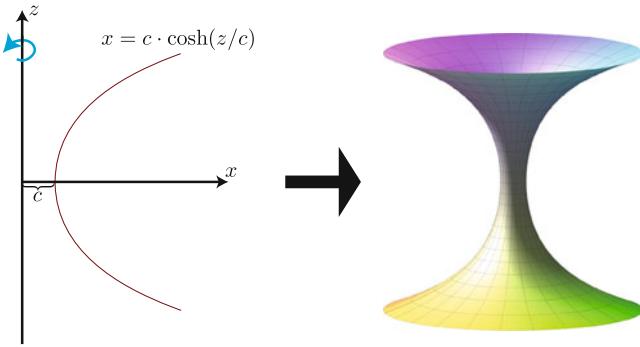


FIGURE 4.20. The catenoid is a minimal surface

EXAMPLE 4.28 (The Helicoid Is Minimal). The helicoid was defined in Exercise 3.103 (on page 186) as the regular surface covered by the single surface patch $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as

$$\sigma(\theta, \tilde{t}) = (\tilde{t} \cos \theta, \tilde{t} \sin \theta, c\theta),$$

where $c > 0$ is a constant. This surface patch is not conformal, but it can be made conformal by the substitution $\tilde{t} = c \sinh t$. In other words, the following alternative surface patch for the helicoid is conformal:

$$(4.21) \quad \sigma(\theta, t) = (c \sinh t \cos \theta, c \sinh t \sin \theta, c\theta).$$

It qualifies as a surface patch because $\sinh : \mathbb{R} \rightarrow \mathbb{R}$ is a smooth function with everywhere positive derivative. The first-order partial derivatives are

$$\begin{aligned}\sigma_\theta &= (-c \sinh t \sin \theta, c \sinh t \cos \theta, c), \\ \sigma_t &= (c \cosh t \cos \theta, c \cosh t \sin \theta, 0).\end{aligned}$$

Therefore, $E = G = c^2 \cosh^2 t$ and $F = 0$, which confirms that σ is conformal, as claimed. The second-order partial derivatives are

$$\begin{aligned}\sigma_{\theta\theta} &= (-c \sinh t \cos \theta, -c \sinh t \sin \theta, 0), \\ \sigma_{tt} &= (c \sinh t \cos \theta, c \sinh t \sin \theta, 0).\end{aligned}$$

Since $\sigma_{\theta\theta} = -\sigma_{tt}$, Lemma 4.26 implies that the helicoid is a minimal surface.

The red coloring above calls attention to the fact that our surface patches for the catenoid and helicoid have the same first fundamental form. By Exercise 3.102 (on page 186), this implies that an open subset of the helicoid is isometric to an open subset of the catenoid. An explanation of this surprising fact is discussed in Exercises 4.56 and 4.57.

A few other examples of minimal surfaces are described in the exercises, but compared to the vast amount of literature on the subject, our treatment of minimal surfaces in this section is admittedly—well—minimal. To further explore minimal surfaces, [5] and [2] are good self-contained introductions to this beautiful area of mathematics.

EXERCISES

EXERCISE 4.47. Explain why a minimal surface satisfies $K \leq 0$. Using Exercise 4.16 (on page 211), conclude that a minimal surface cannot be compact.

EXERCISE 4.48. If a connected minimal surface has constant zero Gaussian curvature, prove that it is a subset of a plane.

EXERCISE 4.49. Let S be the catenoid from Example 4.27 with $c = 1$. For $\lambda > 0$, define $S_\lambda = \{(x, y, z) \in S \mid |z| < \lambda\}$. Notice that the boundary of S_λ consists of two circles, namely the circles along which S intersects the planes $z = \lambda$ and $z = -\lambda$. There is another minimal surface, which we'll call \tilde{S}_λ , with the same boundary as S_λ , namely the two disks that fill in these two circles. Prove that for sufficiently large values of λ , $\text{Area}(\tilde{S}_\lambda) < \text{Area}(S_\lambda)$. Thus, for sufficiently large λ , S_λ is an example of a minimal surface that is not globally area-minimizing compared to other regular surfaces with the same boundary.

EXERCISE 4.50. For $\lambda > 0$, let $B_\lambda = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1 \text{ and } |z| = \lambda\}$. That is, B_λ consists of the circles of radius 1 in the planes $z = \lambda$ and $z = -\lambda$.

- (1) Prove that for sufficiently small $\lambda > 0$, the constant $c > 0$ in Example 4.27 can be chosen such that B_λ equals the intersection of the catenoid with the planes $z = \lambda$ and $z = -\lambda$.
- (2) Prove that for sufficiently large $\lambda > 0$, there is no such constant c . What is the cutoff value?
- (3) How is this related to the maximal height to which the bubble shown on page 193 can be raised before it pinches off in the middle and pops?

EXERCISE 4.51. The function $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ defined as

$$\sigma(u, v) = \left(u - \frac{u^3}{3} + uv^2, v - \frac{v^3}{3} + vu^2, u^2 - v^2 \right)$$

is called **Enneper's surface**. Exercise 3.27 (on page 139) verified that σ is a parametrized surface, and Exercise 3.98 (on page 182) proved that σ is conformal. Verify that this parametrized surface is minimal.

EXERCISE 4.52 (Scherk's Surface). Describe the largest possible domain on which the function $f(x, y) = \ln\left(\frac{\cos y}{\cos x}\right)$ is defined. On this domain, prove that the graph of f is a minimal surface. Use a computer graphing application to plot the graph.

EXERCISE 4.53 (Catalan's Surface). Show that the following parametrized surface $\sigma : \mathbb{R}^2 \rightarrow \mathbb{R}^3$ is minimal:

$$\sigma(u, v) = \left(u - \sin u \cosh v, 1 - \cos u \cosh v, -4 \sin \frac{u}{2} \sinh \frac{v}{2} \right).$$

EXERCISE 4.54. Describe the effect of mean curvature flow on the sphere S^2 . For how long does the flow exist (in other words, during what time interval does the flow exist through regular surfaces)?

EXERCISE 4.55. (Requires differential equations) Prove that every minimal surface of revolution (of the type considered in Example 3.25 on page 131) is a portion of a catenoid.

EXERCISE 4.56 (A Deformation of the Helicoid to the Catenoid). Let $U = \{(\theta, t) \in \mathbb{R}^2 \mid -\pi < \theta < \pi\}$. For each $s \in \mathbb{R}$, define $\sigma_s : U \rightarrow \mathbb{R}^3$ as

$$\sigma_s(\theta, t) = (\cos s) \underbrace{(\cosh t \cos \theta, \cosh t \sin \theta, ct)}_{\text{catenoid}} + (\sin s) \underbrace{(\sinh t \cos \theta, \sinh t \sin \theta, c\theta)}_{\text{helicoid}}.$$

Comparing to Equations 4.20 and 4.21, we see that σ_0 and $\sigma_{\pi/2}$ are the surface patches given in this section for the catenoid and the helicoid respectively (with domain restricted to U); see Fig. 4.21.

- (1) Prove that σ_s is a minimal parametrized surface for all $s \in \mathbb{R}$.
- (2) Verify that the first fundamental form of σ_s does not depend on s .
- (3) Conclude that the open subsets of the helicoid and the catenoid that correspond to the domain U are isometric to each other.
- (4) What happens for $s \in [\pi/2, \pi]$?
- (5) Use a computer graphing application to create an animation of the deformation of the catenoid into the helicoid (such an animation can also be easily found on YouTube).

EXERCISE 4.57 (Conjugate Minimal Surfaces). A pair $\sigma, \tilde{\sigma} : U \subset \mathbb{R}^2 \rightarrow \mathbb{R}^3$ of conformal parametrized surfaces with the same domain U is called *conjugate* if $\sigma_u = \tilde{\sigma}_v$ and $\sigma_v = -\tilde{\sigma}_u$. When this is the case, prove that:

- (1) Both σ and $\tilde{\sigma}$ satisfy the hypothesis of Lemma 4.26, and are therefore both minimal.
- (2) For every $s \in \mathbb{R}$, the function $\sigma_s : U \rightarrow \mathbb{R}^3$ defined as

$$\sigma_s(u, v) = (\cos s)\sigma(u, v) + (\sin s)\tilde{\sigma}(u, v)$$

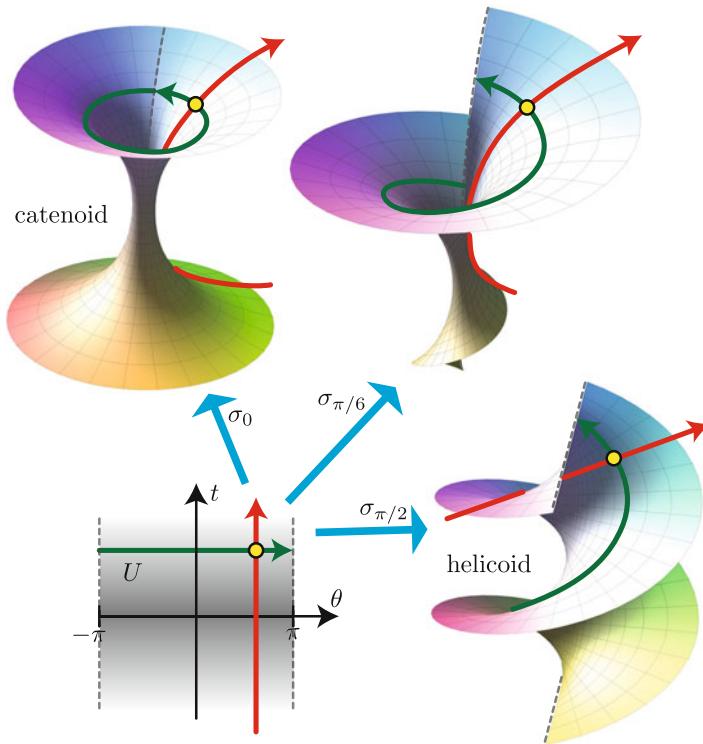


FIGURE 4.21. The helicoid can be deformed into the catenoid through a one-parameter family of minimal surfaces

is a conformal parametrized minimal surface.

- (3) The first fundamental form of σ_s is independent of s .

Verify that the surface patches given in this section for the catenoid (Eq. 4.20) and the helicoid (Eq. 4.21) are conjugate.

EXERCISE 4.58. Consider the deformation in Proposition 4.22 with $\varphi = 1$. Let k_1 and k_2 denote the principal curvatures of S . Let K and H denote the Gaussian and mean curvatures of S . With notation as in the proof of Proposition 4.22:

- (1) Prove that

$$(\sigma_t)_u \times (\sigma_t)_v = (1 - 2tH + t^2K)(\sigma_u \times \sigma_v) = (1 - tk_1)(1 - tk_2)(\sigma_u \times \sigma_v).$$

Thus, if $f_t : S \rightarrow S_t$ denotes the natural map $f_t(p) = p + tN(p)$, then

$$\|d(f_t)_p\| = |(1 - tk_1(p))(1 - tk_2(p))|.$$

- (2) In terms of the maximal principal curvature of S , what is the largest time interval, I , for which all σ_t are parametrized surfaces?

- (3) If t is in the above-described time interval I , $R \subset S$ is a polygonal region, and S is a minimal surface, show that $\text{Area}(R) \leq \text{Area}(R_t)$, where $R_t = f_t(R)$. *COMMENT: this helps justify the term “minimal surface”—at least S has minimal area compared to the other members of this natural one-parameter family of surfaces.*
- (4) For values of t in the above-described time interval I , show that the principal curvatures of S_t are $\frac{k_1}{1-tk_1}$ and $\frac{k_2}{1-tk_2}$, and the principal vectors of S_t are the images under $d(f_t)$ of the principal vectors of S . Show that the Gaussian curvature and mean curvature of S_t are given by

$$K_t = \frac{K}{1 - 2tH + t^2K} \quad \text{and} \quad H_t = \frac{H - tK}{1 - 2tH + t^2K}.$$

- (5) With R_t and K_t defined as in parts (3) and (4), prove that $\int_{R_t} K_t dA$ is independent of t .
- (6) If S has constant Gaussian curvature $\frac{1}{r^2}$, prove that S_r has constant mean curvature $\frac{1}{2r}$. If S has constant mean curvature $\frac{1}{2r}$, prove that S_r has constant Gaussian curvature $\frac{1}{r^2}$.

EXERCISE 4.59. Let S be a minimal surface with negative Gaussian curvature K that is oriented by the unit normal field N .

- (1) Show that for all $p \in S$ and all $x, y \in T_p S$,

$$\langle dN_p(x), dN_p(y) \rangle = -K(p) \langle x, y \rangle.$$

In particular, the Gauss map is conformal.

- (2) Prove that S admits an atlas of conformal surface patches. *HINT: compose the Gauss map with stereographic projection.*



7. The Fary–Milnor Theorem (Optional)

This section is devoted to proving the Fary–Milnor theorem, a global result about space curves that we didn’t include in Chap. 2 because the proof requires techniques from Chap. 4; specifically, it requires measuring the Gaussian curvature of a tubular neighborhood of a space curve.

Throughout this section, $\gamma : [a, b] \rightarrow \mathbb{R}^3$ will denote a unit-speed simple closed space curve whose curvature function, κ , is nowhere zero, and C will denote its trace. We learned in Chap. 2 (page 79) the following:

PROPOSITION 4.29 (Limited Version of Fenchel’s Theorem 2.15).
The total curvature of γ is $\geq 2\pi$.

The goal of this section is to strengthen Fenchel’s theorem in the case that γ is *knotted*. Intuitively, γ (visualized as a loop of string) is called *unknotted* if it can be fiddled into a circle without cutting it. It is called *knotted* if it cannot; see Fig. 4.22. There is a precise definition that captures this “fiddled

into a circle” intuition,¹ but we will instead give an equivalent definition that aligns with the needs of the forthcoming proofs:

DEFINITION 4.30.

A simple closed space curve is called **unknotted** if there exists a set $S \subset \mathbb{R}^3$ (called a **spanning disk**) that is homeomorphic to the disk $D = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$ such that the boundary of S equals the trace of the curve; otherwise, it is called **knotted**.

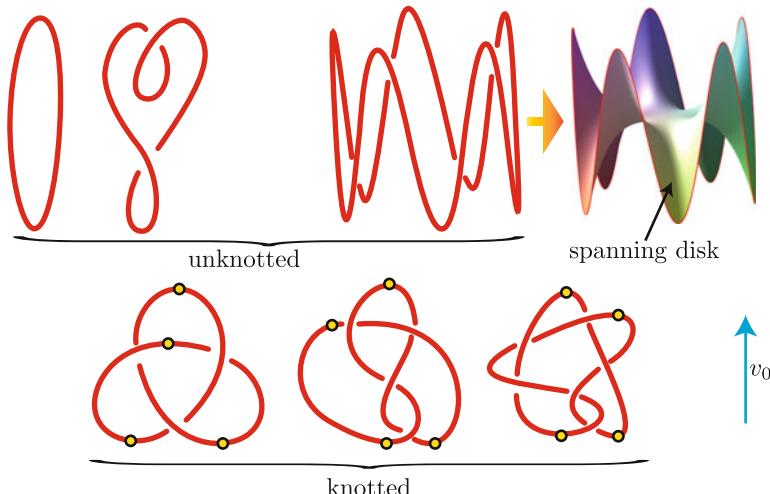


FIGURE 4.22. An unknotted curve has a spanning disk; a knotted curve does not

Since γ is closed, it returns to where it started, which requires some curving; in fact, Fenchel’s theorem says that it requires at least 2π total curvature. The goal of this section is to prove that twice this much total curvature is required in order for γ to tie itself into a knot before returning to where it started:

THEOREM 4.31 (The Fary–Milnor Theorem).

If γ is knotted, then its total curvature is $\geq 4\pi$.

The main technique will involve studying the Gaussian curvature of an auxiliary surface that surrounds C , the way plastic coating surrounds a copper wire. Specifically, the **tubular neighborhood** of radius ϵ about C is defined as follows:

$$S = \{p \in \mathbb{R}^3 \mid \text{dist}(p, C) = \epsilon\},$$

¹We recommend [1] for an elementary excursion into the theory of knots.

where “ $\text{dist}(p, C)$ ” denotes the distance from p to the point of C closest to p .

Exercise 3.12 (on page 124) implies that S is a regular surface for sufficiently small $\epsilon > 0$. We henceforth fix such a choice of ϵ . Denote the portion of S with nonnegative Gaussian curvature by $S_+ = \{p \in S \mid K(p) \geq 0\}$. The integral over S_+ of the Gaussian curvature turns out to equal twice the total curvature of γ :

PROPOSITION 4.32.

$$\iint_{S_+} K \, dA = 2 \int_a^b \kappa(t) \, dt.$$

In particular, the proof will show that S_+ is the type of region over which integrals were defined in Definition 3.55 (on page 161).

PROOF. Exercise 3.12 (on page 124) implies that S can be parametrized as

$$(4.22) \quad \sigma(s, t) = \gamma(t) + \epsilon (\cos(s)\mathbf{n}(t) + \sin(s)\mathbf{b}(t)),$$

where $\{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}$ denotes the Frenet frame at $\gamma(t)$. We will take the domain of this surface patch σ to be $U = \{(s, t) \in \mathbb{R}^2 \mid t \in (a, b), s \in (0, 2\pi)\}$, so that the closure of the image of σ equals all of S (because this image misses only the parameter curves $s = 0$ and $t = a$).

Using the Frenet equations (on page 45), it is straightforward to compute σ_s and σ_t , and to conclude that

$$\sigma_s \times \sigma_t = \epsilon (1 - \epsilon(\cos s)\kappa) \underbrace{((\cos s)\mathbf{n} + (\sin s)\mathbf{b})}_N.$$

In particular, $N = ((\cos s)\mathbf{n} + (\sin s)\mathbf{b})$ is the outward-pointing normal field, and

$$(4.23) \quad \|d\sigma\| = |\sigma_s \times \sigma_t| = \epsilon (1 - \epsilon(\cos s)\kappa).$$

It is then straightforward to calculate E, F, G, e, f, g and use Eq. 4.13 (on page 220) to learn that

$$(4.24) \quad K = \frac{-(\cos s)\kappa}{\epsilon (1 - \epsilon(\cos s)\kappa)}.$$

Equation 4.24 implies that K has the same sign as $-(\cos s)$, so exactly half of the s -values correspond to points of positive curvature, namely the values $s \in (\frac{\pi}{2}, \frac{3\pi}{2})$, as shown in Fig. 4.23. This generalizes the torus of revolution example of Exercise 4.23 on page 212. Using Eq. 4.23 and 4.24, we have

$$\begin{aligned} \iint_{S_+} K \, dA &= \int_a^b \int_{\pi/2}^{3\pi/2} K \|d\sigma\| \, ds \, dt = \int_a^b \int_{\pi/2}^{3\pi/2} -\kappa(t)(\cos s) \, ds \, dt \\ &= - \int_a^b \kappa(t) \int_{\pi/2}^{3\pi/2} (\cos s) \, ds \, dt = 2 \int_a^b \kappa(t) \, dt. \end{aligned}$$

□

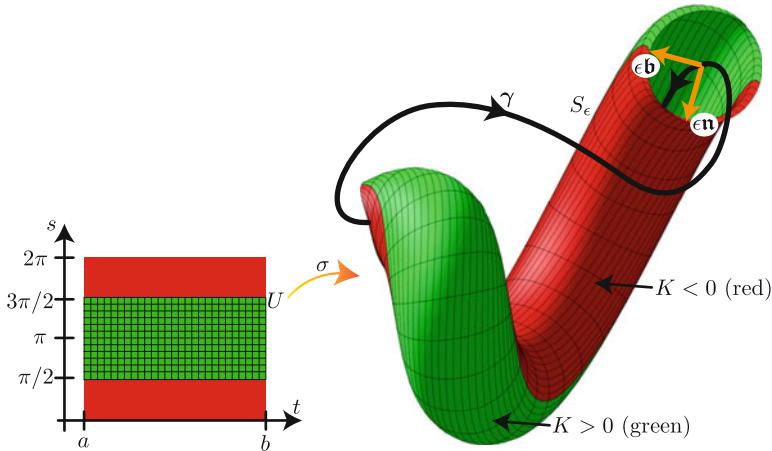


FIGURE 4.23. Regions of positive (green) and negative (red) Gaussian curvature on a tubular neighborhood of a space curve (shown cut away for clarity)

In light of Proposition 4.32, all that remains to re-prove the above limited version of Fenchel's theorem, and more importantly to prove the Fary–Milnor theorem (4.31) is to demonstrate that

$$\iint_{S_+} K \, dA \geq \begin{cases} 4\pi & \text{generally,} \\ 8\pi & \text{if } \gamma \text{ is knotted.} \end{cases}$$

That is exactly what we will prove with the next two propositions.

PROPOSITION 4.33.

- (1) $N : S_+ \rightarrow S^2$ (the restriction of the Gauss map to S_+) is surjective.
- (2) $\iint_{S_+} K \, dA \geq 4\pi$.

For a unit vector $v_0 \in \mathbb{R}^3$, we will henceforth denote by $H_{v_0} : \mathbb{R}^3 \rightarrow \mathbb{R}$ the corresponding **height function**, defined as $H_{v_0}(w) = \langle v_0, w \rangle$. For example, if $v_0 = (0, 0, 1)$, then $H_{v_0}(x, y, z) = z$, which is why it's called a “height function.” Even for an arbitrary choice of v_0 , the term is appropriate if you imagine tilting your head so that v_0 becomes the “up” direction (or more precisely, apply a rigid motion sending v_0 to $(0, 0, 1)$).

PROOF OF PROPOSITION 4.33. For part (1), let $v_0 \in \mathbb{R}^3$ be an arbitrary unit vector (that is, an arbitrary element of S^2). Let $p \in S$ be the global maximum of the restriction of H_{v_0} to S ; that is, p is the highest point of S . Notice that S lies entirely on one (closed) side of the plane normal to v_0 passing through p (because this plane is a level set of H_{v_0}). It follows (by Exercise 3.35 on page 145) that this plane is the tangent plane to S at p_0 ,

and also (by Proposition 4.14 on page 216) that $K(p_0) \geq 0$. Thus, $p_0 \in S_+$ and $N(p_0) = v_0$.

For part (2), Fig. 4.23 illustrates how to decompose S_+ into the union of a finite collection $\{R_i\}$ of arbitrarily small polygonal regions intersecting only along boundaries. For this, just define the sets $\{R_i\}$ to be the images under σ of the cells of a grid of rectangles in the green portion of the closure of U . We compute

$$\begin{aligned} \iint_{S_+} K \, dA &= \iint_{S_+} |K| \, dA = \sum_i \iint_{R_i} |K| \, dA = \sum_i \underbrace{\iint_{R_i} \|dN\| \, dA}_{\text{Equation 4.11, page 216}} \\ &= \underbrace{\sum_i \text{Area}(N(R_i))}_{\text{Proposition 3.60, page 164}} \geq \text{Area}(N(S_+)) = \underbrace{\text{Area}(S^2)}_{\text{part (1)}} = 4\pi. \end{aligned}$$

We cheated a bit in the fourth equality above. To apply Proposition 3.60 honestly would require N to be a diffeomorphism on each R_i , which is not quite the case for the green cells bordering red cells (which have edges along which K equals zero). This minor problem can be fixed by defining $S_+^\delta \subset S_+$ to equal the portion of S_+ corresponding to values $s \in [\pi/2 + \delta, 3\pi/2 - \delta]$. For every fixed small $\delta > 0$, S_+^δ can be decomposed into a grid of polygonal regions R_i such that N is a diffeomorphism on each R_i . Thus, $\sum_i \text{Area}(N(R_i)) \geq \text{Area}(N(S_+^\delta))$, which as $\delta \rightarrow 0$ becomes arbitrarily close to $\text{Area}(N(S_+)) = \text{Area}(S^2)$. \square

Part (1) of Proposition 4.33 is more generally true if S is an arbitrary compact regular surface and S_+ denotes the portion of S with nonnegative Gaussian curvature (Exercise 4.26 on page 216). In fact, part (2) is also true in this generality, although to interpret and prove this assertion would require generalizing our definition of integration (Definition 3.55 on page 161) to cover more than just finite unions of polygonal regions.

The Fary–Milnor theorem follows immediately from Proposition 4.32 together with part (2) of the following refinement of Proposition 4.33:

PROPOSITION 4.34.

If γ is knotted, then

- (1) $N : S_+ \rightarrow S^2$ is at least 2-to-1; that is, every point of S^2 is the image under the Gauss map of at least two points of S_+ .
- (2) $\iint_{S_+} K \, dA \geq 8\pi$.

We will prove the contrapositive of (1); more precisely, if the Gauss map is not at least two-to-one, then the following lemma will help us to conclude that γ is unkotted:

LEMMA 4.35.

If there exists a unit vector $v_0 \in \mathbb{R}^3$ such that $H_{v_0} \circ \gamma : [a, b] \rightarrow \mathbb{R}$ (the

corresponding height function composed with γ) has only two local extrema, then γ is unknotted.

In counting local extrema, a and b are considered a single element of the domain. For example, $\{a, b\}$ would count as a single local minimum of $H_{v_0} \circ \gamma$ if there exists $\delta > 0$ such that $H_{v_0}(\gamma(t)) \geq H_{v_0}(\gamma(a)) = H_{v_0}(\gamma(b))$ for all $t \in (a, a + \delta)$ and for all $t \in (b - \delta, b)$. This convention ensures that the extremum count is independent of parametrization.

To better understand the lemma's content, look at any of the knotted curves in Fig. 4.22. The height function $H_{v_0} \circ \gamma$ has four local extrema corresponding to the four illustrated yellow dots where the curve (viewed from this angle) seems to switch between going up and going down. The lemma says that this angle isn't special—even after performing an arbitrary rotation, there must still be at least three such points, and we'll see in the proof that this actually forces there to be at least four such points.

PROOF OF LEMMA 4.35. Suppose that $v_0 \in \mathbb{R}^3$ is a unit vector such that $H_{v_0} \circ \gamma$ has only two local extrema, which by compactness must correspond to the global minimum and maximum of $H_{v_0} \circ \gamma$. After reparametrizing γ , we can assume that $H_{v_0} \circ \gamma$ is monotonically increasing (and in particular one-to-one) on $[a, c]$ and is monotonically decreasing (and therefore one-to-one) on $[c, b]$ for some $c \in (a, b)$. Set $z_a = H_{v_0}(\gamma(a))$ and $z_c = H_{v_0}(\gamma(c))$, so that the image of $H_{v_0} \circ \gamma$ is the interval $[z_a, z_c]$.

For fixed $z \in (z_a, z_c)$, there is exactly one value $t_z \in (a, c)$ with $H_{v_0}(\gamma(t_z)) = z$ and exactly one value $\tilde{t}_z \in (c, b)$ with $H_{v_0}(\gamma(\tilde{t}_z)) = z$. Visually, $\gamma(t_z)$ and $\gamma(\tilde{t}_z)$ are the two points of intersection of the plane $P_z = \{w \in \mathbb{R}^3 \mid \langle w, v_0 \rangle = z\}$ with the trace of γ ; see Fig. 4.24. Let $f_z : [0, 1] \rightarrow \mathbb{R}^3$ denote the constant-speed parametrization of the straight line from $f_z(0) = \gamma(t_z)$ to $f_z(1) = \gamma(\tilde{t}_z)$. Let $U = \{(z, s) \in \mathbb{R}^2 \mid z \in (z_a, z_c), s \in (0, 1)\}$ and define $\sigma : U \rightarrow \mathbb{R}^3$ as $\sigma(z, s) = f_z(s)$. It is straightforward to check that σ is a homeomorphism onto its image (Exercise 4.63). Since U is homeomorphic to the disk D , the image $\sigma(U)$ is a spanning disk. \square

PROOF OF PROPOSITION 4.34. For part (1), let $v_0 \in \mathbb{R}^3$ be an arbitrary unit vector (that is, an arbitrary element of S^2). Since γ is knotted, Lemma 4.35 implies that $H_{v_0} \circ \gamma$ has at least three local extrema, which actually forces it to have at least two distinct maxima. Why? If not all extrema are isolated (for example if $H_{v_0} \circ \gamma$ is constant on an interval), then there are infinitely many maxima. On the other hand, if all extrema are isolated, then local minima and maxima alternate around the curve, so there are equal numbers of each, forcing the total count to be even. In either case, there are at least two distinct maxima.

We will show that each of these two maxima determines a point of S_+ that is mapped by N to v_0 . These two points of S_+ will be distinct, because their t -values (as in Eq. 4.22) will be distinct—they will be the time parameters at which the two maxima occur.

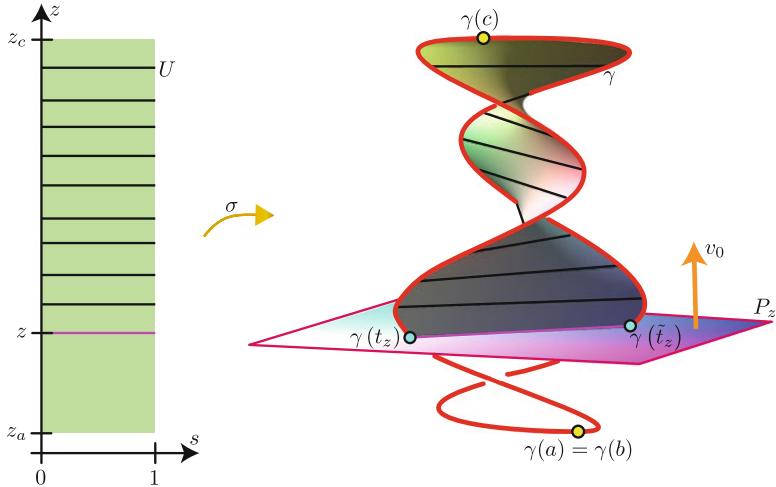


FIGURE 4.24. If $H_{v_0} \circ \gamma$ has only two extrema, then a spanning disk can be constructed by connecting pairs of points at which planes normal to v_0 intersect the trace of γ

For this, let $t_0 \in [a, b]$ be the time parameter at which one such local maximum occurs. Notice that $\gamma'(t_0)$ is orthogonal to v_0 , because

$$0 = (H_{v_0} \circ \gamma)'(t_0) = \langle \gamma'(t_0), v_0 \rangle.$$

Therefore, $v_0 \in \text{span}\{\mathbf{n}(t_0), \mathbf{b}(t_0)\}$, so $v_0 = (\cos s_0)\mathbf{n}(t_0) + (\sin s_0)\mathbf{b}(t_0)$ for some $s_0 \in [0, 2\pi]$. The restriction of H_{v_0} to S is locally maximized at the point

$$p_0 = \gamma(t_0) + \epsilon v_0 = \sigma(t_0, s_0) \in S,$$

because $H_{v_0} \circ \gamma$ is locally maximized at t_0 , and because the value of H_{v_0} can't change by more than ϵ between a point of the trace of γ and a point of its corresponding s -parameter curve in S . As in the proof of Proposition 4.33, this implies that $p_0 \in S_+$ and that $N(p_0) = v_0$.

Part (2) is proven just like the corresponding statement from Proposition 4.33. \square

The above proof idea can be visualized with a loop of plastic-coated copper wire. After arbitrarily rotating the wire loop into any position, notice that the highest point of the plastic coating lies on the s -parameter circle around the highest point of the inner copper wire. You could visualize its tangent plane at this highest point by lowering a horizontal sheet of cardboard until it first touches the plastic. At this point, the plastic coating has upward normal vector and has nonnegative Gaussian curvature. The same claims can be made of locally highest points. Think about the conditions under which

a (locally) highest point has strictly positive Gaussian curvature, and then compare your visual intuition with Exercise 4.62.

EXERCISES

EXERCISE 4.60. Verify Eq. 4.24.

EXERCISE 4.61. Prove that the area of S equals $2\pi\epsilon l$, and the volume enclosed in S equals $\pi\epsilon^2 l$, where $l = b - a$ is the length of γ .

EXERCISE 4.62. Let $v_0 \in \mathbb{R}^3$ be a unit vector that is not in the image of the unit binormal function; that is, $v_0 \neq \mathbf{b}(t)$ for all $t \in [a, b]$.

- (1) Prove that every critical point of $H_{v_0} \circ \gamma$ is either a local minimum or a local maximum.
- (2) If $p \in S$ is a local extremum of the restriction of H_{v_0} to S , prove that $K(p) > 0$.

EXERCISE 4.63. Prove the claim in the proof of Lemma 4.35 that σ is a homeomorphism onto its image. If we additionally assume that $H_{v_0} \circ \gamma$ has no critical points on (a, b) other than c , use Proposition 3.29 (on page 135) to prove that σ is a diffeomorphism onto its image. *COMMENT:* *The previous exercise provides a condition under which the “no other critical points” hypotheses is guaranteed to hold.*

EXERCISE 4.64. Notice that the interior of S_+ is $(S_+)^{\circ} = \{p \in S \mid K > 0\}$.

- (1) If γ is a convex curve contained in a plane, prove that $N : (S_+)^{\circ} \rightarrow S^2$ is one-to-one, and conclude that the total curvature of γ equals exactly 2π .
- (2) If the total curvature of γ is greater than 2π , prove that $N : (S_+)^{\circ} \rightarrow S^2$ is not one-to-one, and conclude that γ is not a convex plane curve.

EXERCISE 4.65. In the Fary–Milnor theorem, replace the inequality with *strict* inequality. That is, prove that the total curvature of a knotted simple closed space curve with nowhere-vanishing curvature is greater than 4π .

EXERCISE 4.66. Let S be the torus of revolution described in Exercise 3.23 (on page 138), which was parametrized as

$$\sigma(\theta, t) = ((2 + \cos t) \cos \theta, (2 + \cos t) \sin \theta, \sin t), \quad \theta, t \in [0, 2\pi].$$

For any pair p, q of relatively prime positive integers, the (p, q) -torus knot is defined as

$$\gamma(s) = \sigma(p \cdot s, q \cdot s), \quad s \in [0, 2\pi];$$

see Fig. 4.25.

- (1) If $p = 1$ or $q = 1$, prove that γ is unknotted.
- (2) Prove that the total curvature of the (p, q) -torus knot grows without bound as $p, q \rightarrow \infty$.

COMMENT: It can be shown that γ is knotted if neither p nor q equals 1.

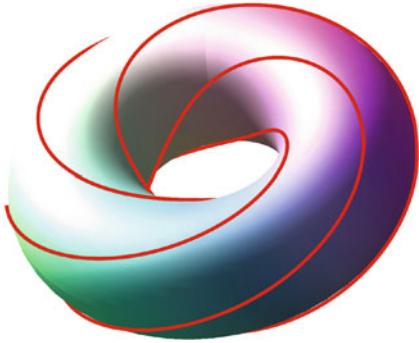


FIGURE 4.25. The $(4, 5)$ -torus knot



This re-creation of Foucault's pendulum tips over pegs as its swing-plane rotates throughout the day. The phenomenon is best modeled using fundamental concepts presented in this chapter: geodesics and parallel vector fields.

Geodesics

The most fundamental concept for studying the geometry of \mathbb{R}^2 is a straight line. The goal of this chapter is to generalize this fundamental notion from \mathbb{R}^2 to arbitrary regular surfaces. Although most surfaces curve in such a way that they don't contain any straight lines, they do contain curves called *geodesics*, which will turn out to share many important characterizing properties of straight lines.

In order to focus the beginning sections of this chapter on geometry, we will postpone (until Sect. 6) the proofs of two necessary algebraic results, namely, (1) the existence and uniqueness of geodesics and (2) a formula for Gaussian curvature in normal polar coordinates.

1. Definition and Examples of Geodesics

If a small motorized toy car is set on a flat surface, it will travel forward in a straight line. What if it is set on the curved surface of the earth (modeled as a perfect sphere that is very large compared to the size of the car)? The car will remain on the sphere due to gravity and friction, so it won't travel in a straight line, because the sphere doesn't contain any straight lines. What sort of curve, γ , will it traverse? It will maintain a constant speed, which

means that $\gamma''(t)$ will not have any component in the direction of $\gamma'(t)$. It will not turn left or right, which means that $\gamma''(t)$ will not have a component in the direction of $R_{90}(\gamma'(t))$. The only possible direction left for $\gamma''(t)$ is the direction normal to the surface. Thus, we expect the car to follow a *geodesic*—a curve that turns only in the normal direction:

DEFINITION 5.1.

A regular curve in a regular surface S , denoted by $\gamma : I \rightarrow S$, is called a **geodesic** if for every $t \in I$, the acceleration vector $\gamma''(t)$ is a normal vector to S at $\gamma(t)$.

In this definition, the domain I is allowed to be any type of interval. We'll use the term **geodesic segment** for a geodesic whose domain is a *compact* interval $[a, b]$. The following is immediate from the definition:

LEMMA 5.2.

Let S be a regular surface and let $\gamma : I \rightarrow S$ be a regular curve.

- (1) If γ is a geodesic, then γ has constant speed (because $\gamma''(t) \perp \gamma'(t)$).
- (2) If γ is of unit speed, then γ is a geodesic if and only if $\kappa_g(t) = 0$ for all $t \in I$.

Recall from Sect. 3 of Chap. 4 that κ_g denotes the **geodesic curvature** function, which roughly measures how sharply γ turns to the left (as discussed in Exercise 4.24 on page 213). Although the sign of κ_g depends on the choice of (local) orientation, the question whether it equals zero does not, so the “ $\kappa_g = 0$ ” condition is well defined even for a nonorientable surface. This condition roughly means that the acceleration vector has no left/right component.

If the toy car is placed anywhere on the surface facing in any direction, it will go in that direction at least for a little while, say until it reaches the boundary of the surface. Its trajectory is completely determined by its starting position and direction and varies smoothly with this initial data. Here is the precise statement:

PROPOSITION 5.3 (Existence and uniqueness of geodesics).

Let S be a regular surface, $p \in S$, and $v \in T_p S$ with $r = |v| \neq 0$. There exists $\epsilon = \epsilon(p, r) > 0$ (depending smoothly on p and r) such that:

- (1) There exists a geodesic $\gamma_v : (-\epsilon, \epsilon) \rightarrow S$ satisfying the initial conditions $\gamma_v(0) = p$ and $\gamma'_v(0) = v$.
- (2) Any two geodesics with this domain satisfying these initial conditions must be equal.

Furthermore, $\gamma_v(t)$ depends smoothly on p, v, t .

This proposition says that a unique geodesic passes through a given point p in a given direction v . The smoothness claim is a bit technical; it means that

$$(p, v, t) \mapsto \begin{cases} p & \text{if } |v| = 0, \\ \gamma_v(t) & \text{otherwise,} \end{cases}$$

is smooth on the domain

$$\{(p, v, t) \in \mathbb{R}^3 \times \mathbb{R}^3 \times \mathbb{R} \mid p \in S, v \in T_p S, |t| < \epsilon(p, |v|)\}.$$

We will postpone the proof of this proposition until Sect. 6, in order to focus now on examples and geometric applications.

First notice that it is enough to understand the *unit-speed* geodesics, because all other geodesics are just reparametrizations of them. More specifically, if $u \in T_p S$ is of unit length and $r \neq 0$ is sufficiently small, then

$$(5.1) \quad \gamma_{ru}(t) = \gamma_u(rt).$$

That is, the geodesic through p in the direction ru is just the reparametrization with speed r of the geodesic through p in the direction u . In particular, the dependence of $\epsilon(p, r)$ on r is straightforward: $\epsilon(p, r) = \frac{\epsilon(p, 1)}{r}$.

EXAMPLE 5.4 (Geodesics in the plane). A regular curve γ in the plane \mathbb{R}^2 is a geodesic if and only if it is (all or a portion of) a straight line parametrized at constant speed. This is because $\gamma''(t) \in \mathbb{R}^2$, so $\gamma''(t)$ is normal to \mathbb{R}^2 if and only if it vanishes.

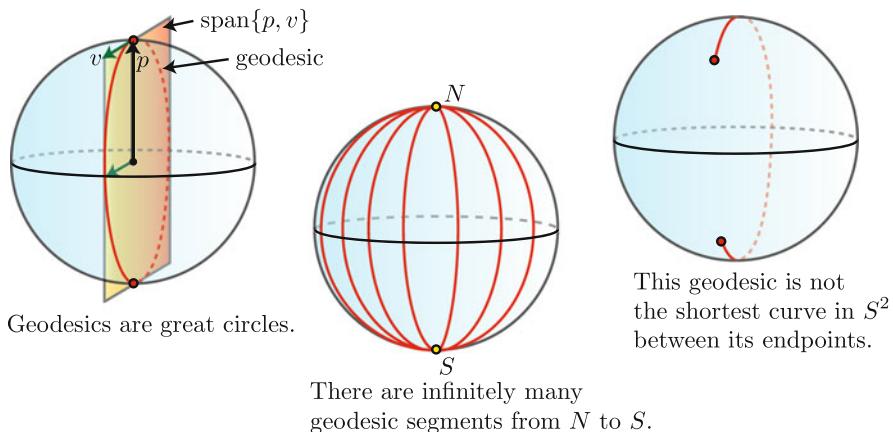
EXAMPLE 5.5 (Geodesics in the punctured plane). If the origin is removed from the plane, the resulting set, denoted by $S = \mathbb{R}^2 - \{(0, 0)\}$, is a regular surface. As in the previous example, a regular curve in S is a geodesic if and only if it is (all or a portion of) a straight line parameterized at constant speed. This exemplifies two interesting phenomena:

- (1) There does not exist a geodesic in S between $(-1, 0)$ and $(1, 0)$ (because the unique line connecting these points in \mathbb{R}^2 passes through the missing origin).
 - (2) The largest possible choice of ϵ in Proposition 5.3 for this surface is $\epsilon(p, 1) = |p|$. The limiting factor at each point is the direction pointing toward the missing origin. Notice that $\epsilon(p, 1) \rightarrow 0$ as $p \rightarrow (0, 0)$.
-

EXAMPLE 5.6 (Geodesics on the sphere). Recall that the outward-pointing normal field on S^2 is described as $N(p) = p$. Let $p \in S^2$ and let $v \in T_p S^2$ be of unit length. The curve $\gamma: \mathbb{R} \rightarrow S^2$ defined as

$$\gamma(t) = (\cos t)p + (\sin t)v$$

is a geodesic, because $\gamma''(t) = -\gamma(t) = -N(\gamma(t))$. Notice that the trace of γ is a **great circle**, which means the intersection of S^2 with a plane through the origin (in this case, the plane spanned by p and v). Proposition 5.3(2) now implies that every unit-speed geodesic in S^2 is (all or part of) a great circle. Some implications are illustrated in Fig. 5.1. For example, the north and south poles have infinitely many geodesic segments between them. Also,

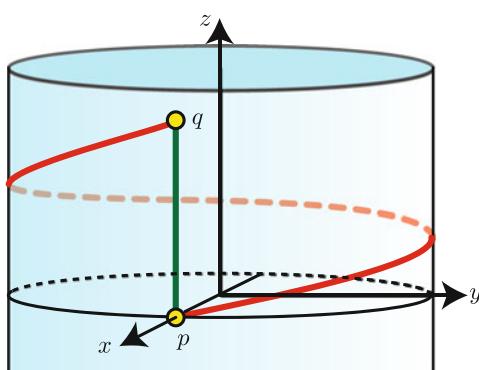
FIGURE 5.1. Geodesics in S^2

a geodesic segment of length greater than π is not a shortest curve in S^2 between its endpoints.

EXAMPLE 5.7 (Geodesics on a cylinder). Consider the cylinder

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 = 1\}.$$

Its outward-pointing normal field can be described as $N(x, y, z) = (x, y, 0)$. For every $c \in \mathbb{R}$, the **helix** $\gamma(t) = (\cos t, \sin t, ct)$ is a geodesic, because $\gamma''(t) = -N(\gamma(t))$. Its initial position is $p = \gamma(0) = (1, 0, 0)$, and its initial velocity is $\gamma'(0) = (0, 1, c)$. After possibly reversing the direction in which γ is traversed, we can choose c to make the initial velocity vector of γ point in the direction of any unit vector in $T_p C$ except $(0, 0, \pm 1)$. The geodesic in either omitted direction is the vertical line $\gamma(t) = (1, 0, \pm t)$. As illustrated in Fig. 5.2, a sufficiently large segment of a helix is not the shortest curve in C between its endpoints.

FIGURE 5.2. Two geodesic segments with different lengths between p and q

The above examples help dispel some incorrect guesses about geodesics. For example, there is *not* necessarily a unique geodesic segment joining each pair of points of a surface; in fact, there could be none or there could be infinitely many. Further, even though geodesics are analogous to straight lines in many ways, a geodesic segment is not necessarily the

shortest possible curve in the surface between its endpoints (but we will learn in the next section that a *sufficiently short* geodesic segment is guaranteed to have this property).

For surfaces more complicated than the above examples, it is often impossible to describe all geodesics explicitly. Nevertheless, there is a large family of surfaces, namely surfaces of revolution, for which the behavior of geodesics can be qualitatively understood in a marvelously uniform way:

THEOREM 5.8 (Clairaut's Theorem).

Let S be a surface of revolution (Example 3.25 on page 131). Let $\beta : I \rightarrow S$ be a unit-speed curve in S . For every $s \in I$, let $\rho(s)$ denote the distance from $\beta(s)$ to the axis of rotation (the z -axis) and let $\psi(s) \in [0, \pi]$ denote the angle between $\beta'(s)$ and the longitudinal curve through $\beta(s)$, as in Fig. 5.3 (left).

- (1) If β is a geodesic, then $\rho(s) \sin(\psi(s))$ is constant on I .
- (2) If $\rho(s) \sin(\psi(s))$ is constant on I , then β is a geodesic, provided no subsegment of β equals a subsegment of a latitudinal curve.

The value of $\sin \psi$ is unaffected by the decision whether to interpret ψ as the angle between β' and the up-pointing or the down-pointing vector tangent to the longitudinal curve. Either choice is fine, or even better, define ψ to be the smaller possibility, so that $\psi(s) \in [0, \pi/2]$ for all $s \in I$. With this convention, Clairaut's theorem says that ρ must increase as ψ decreases (and vice versa) along a geodesic. Thus, a geodesic will be more nearly vertical (longitudinal) when ρ is large, and more nearly horizontal (latitudinal) when ρ is small, as illustrated in Fig. 5.3 (left).

Every longitudinal curve is a geodesic (when parametrized at constant speed). This follows from Clairaut's theorem (because $\psi = 0$) or from elementary arguments (since a longitudinal curve lies in a plane, its acceleration vector must also lie in this plane). On the other hand, the fine print prevents Clairaut's theorem from telling us anything about which latitudinal curves are geodesics. Along any latitudinal curve, the value $\rho \sin \psi = \rho$ is constant, but it turns out (Exercise 5.4) that the latitude $t = t_0$ is the trace of a geodesic if and only if $x'(t_0) = 0$, where $\gamma(t) = (x(t), 0, z(t))$ denotes the generating curve. Notice that $\rho = x$ along the generating curve, so this really says the latitudes at critical values of ρ are geodesics; see Fig. 5.3 (right).

To physicists, Clairaut's theorem is just the law of conservation of angular momentum for a unit-speed particle constrained to S that is acted on only by a force field normal to S at every point. Angular momentum is conserved in this situation, because such a force field has no moment about the axis of revolution. The particle's angular momentum is proportional to $\rho v_\theta = \rho \sin \psi$, where v_θ is the component of velocity in the latitudinal direction; the conservation of this quantity is exactly Clairaut's theorem. Thus, the latitudinal

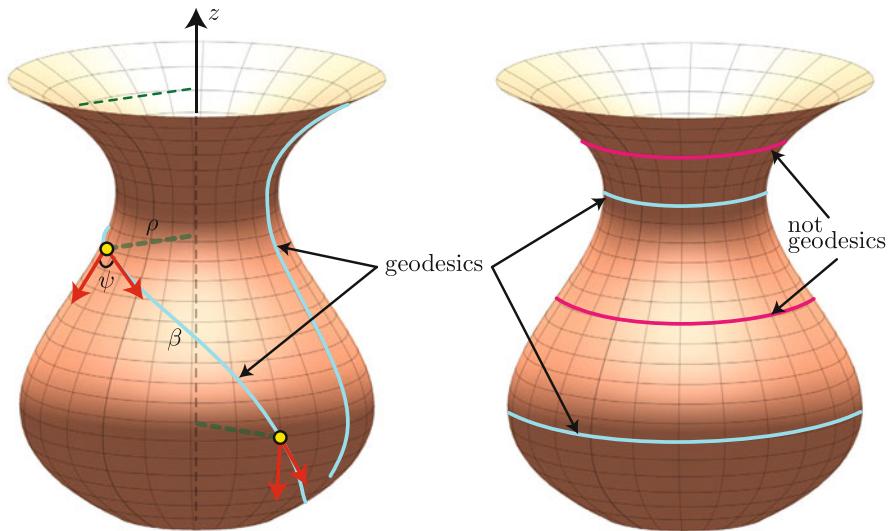


FIGURE 5.3. All longitudes and all critical latitudes are geodesics. Along other geodesics, $\rho \sin \psi$ is constant

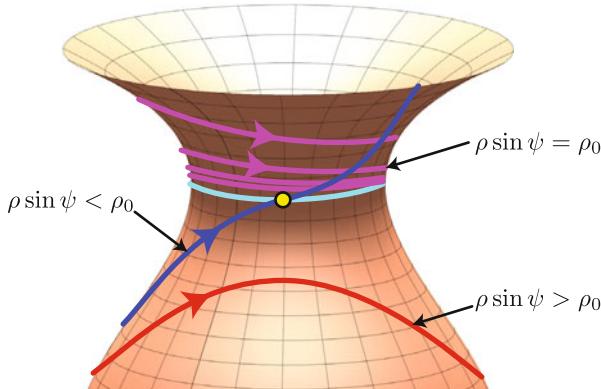


FIGURE 5.4. The behavior of geodesics approaching a “thin waist” (a latitude at which a local minimum value of $\rho = \rho_0$ occurs)

component of its speed (the part that contributes to angular momentum) is smaller when the particle is farther from the axis. This is reminiscent of the way in which an ice skater can slow her spin by extending her hands farther from her body.

Clairaut’s theorem can provide good qualitative information about the behavior of geodesics in a surface of revolution. For example, Fig. 5.4 shows

three geodesics that are initially approaching a “thin waist” (a latitude at which a local minimum value of $\rho = \rho_0$ occurs). The red geodesic has too much angular momentum to reach this waist without contradicting the assumption that it has unit speed, so it turns around at a wider latitude. Imagine that the wider latitude represents the farthest that our ice skater can pull her hand into her body without her hand going too fast. The angular momentum of the purple geodesic equals that of the waist, namely ρ_0 , so it spirals forever toward the waist. Only the angular momentum of the blue geodesic is small enough to cross the waist.

PROOF OF CLAIRAUT’S THEOREM.

The natural surface patch for S was defined in Example 3.25 as

$$\sigma(\theta, t) = (x(t) \cos \theta, x(t) \sin \theta, z(t)),$$

where $\gamma(t) = (x(t), 0, z(t))$ denotes the generating curve. According to Exercise 3.105 (on page 186), the first fundamental form of σ is

$$(5.2) \quad \underbrace{(x'(t)^2 + z'(t)^2)}_E dt^2 + \underbrace{x(t)^2}_G d\theta^2.$$

A regular curve $\beta : I \rightarrow S$ on the portion of S covered by this surface patch will have the form $\beta(s) = \sigma(\theta(s), t(s))$, for some smooth functions $\theta, t : I \rightarrow \mathbb{R}$. Using the prime symbol $'$ to denote derivatives with respect to s , and assuming that everything is evaluated at an arbitrary $s \in I$ or at $(\theta(s), t(s))$ as appropriate, the chain rule gives

$$\beta' = \theta' \sigma_\theta + t' \sigma_t \quad \text{and} \quad \beta'' = \theta'' \sigma_\theta + \theta' (\theta' \sigma_{\theta\theta} + t' \sigma_{\theta t}) + t'' \sigma_t + t' (\theta' \sigma_{t\theta} + t' \sigma_{tt}).$$

One can confirm (with visual reasoning or a calculation) that the vector σ_θ is orthogonal to each of these vectors: σ_t , σ_{tt} and $\sigma_{\theta\theta}$. Therefore,

$$(5.3) \quad \langle \beta'', \sigma_\theta \rangle = \langle \theta'' \sigma_\theta + 2\theta' t' \sigma_{\theta t}, \sigma_\theta \rangle = \underbrace{\theta'' \langle \sigma_\theta, \sigma_\theta \rangle + 2\theta' t' \langle \sigma_{\theta t}, \sigma_\theta \rangle}_{\text{because } (G)' = t' G_t + \theta' G_\theta = t' G_t = 2t' \langle \sigma_\theta, \sigma_{\theta t} \rangle} = (\theta' G)',$$

where $G = \langle \sigma_\theta, \sigma_\theta \rangle = x^2 = \rho^2$, as in Eq. 5.2. In summary,

$$(5.4) \quad \langle \beta'', \sigma_\theta \rangle = (\theta' G)' = (\theta' \rho^2)'.$$

If β is a unit-speed geodesic, then $\beta'' \perp \sigma_\theta$, so $(\theta' \rho^2)' = 0$, which means that $\theta' \rho^2$ is constant. So to justify the first claim of Clairaut’s theorem, it remains to show that $\theta' \rho^2 = \rho \sin \psi$, or equivalently that $\theta' \rho = \sin \psi$. But this just follows from the fact that $\sin \psi$ is the length of the projection of β' onto the latitudinal direction.

The second claim of the theorem is left to the reader in Exercise 5.9. \square

We saw in Figs. 5.1 and 5.2 that a geodesic segment is not necessarily the shortest curve in the surface between its endpoints. But the geodesic segments from those examples were just too long. We will learn in the next

section that a sufficiently short geodesic segment *is* the shortest curve in the surface between its endpoints. In anticipation of this, we end the current section with some vocabulary for discussing shortest curves in surfaces.

DEFINITION 5.9.

Let S be a connected regular surface. The *intrinsic distance function* of S , denoted by $d : S \times S \rightarrow \mathbb{R}$, is defined so that

$$d(p, q) = \inf\{\text{length}(\gamma) \mid \gamma \text{ is a piecewise regular curve in } S \text{ from } p \text{ to } q\}$$

for all $p, q \in S$.

The value $d(p, q)$ is finite, because there exist piecewise regular curves between every pair of points of S (Exercise 5.13). Furthermore, we have the following:

LEMMA 5.10.

Let S be a connected regular surface. For all $p, q, r \in S$,

- (1) $d(p, q) \geq 0$, with equality if and only if $p = q$.
- (2) $d(p, q) = d(q, p)$
- (3) $d(p, q) \leq d(p, r) + d(r, q)$ (*the triangle inequality*)

PROOF. Exercise 5.14. □

The term “intrinsic distance function” is consistent with our previous definition of “intrinsic” in Sect. 7 of Chap. 3. That is, the measurements on S that are intrinsic are exactly those that depend only on the intrinsic distance function, since this is the same as depending only on the first fundamental form (Exercise 5.18). For example, the restriction to S of the distance function from \mathbb{R}^3 ($\text{dist}(p, q) = |p - q|$) is *not* intrinsic (Exercise 5.19).

EXAMPLE 5.11. Let S be the punctured plane of Example 5.5, $p = (-1, 0)$, and $q = (1, 0)$. Then $d(p, q) = 2$, but there does not exist a curve in S from p to q with length exactly 2, because the straight line in \mathbb{R}^2 between these points passes through the omitted origin. Thus, the infimum in Definition 5.9 is not always attained.

The following definition provides a vocabulary for discussing situations in which the infimum in Definition 5.9 *is* obtained, or even is obtained by a unique curve:

DEFINITION 5.12.

Let S be a regular surface and $\gamma : I \rightarrow S$ a piecewise regular curve in S .

- (1) When $I = [a, b]$ is compact, γ is called **minimizing** if $\text{length}(\gamma) = d(\gamma(a), \gamma(b))$, and γ is called **uniquely minimizing** if additionally, every other piecewise regular curve in S from $\gamma(a)$ to $\gamma(b)$ with the same length as γ must have the same trace as γ .
- (2) For general I , γ is called **minimizing** (or **uniquely minimizing**) if every restriction of γ to a compact subinterval of I has this property, as defined in part (1).
- (3) For general I , γ is called **locally minimizing** if for every $t_0 \in \text{interior}(I)$, there exists a compact subinterval of I whose interior contains t_0 , restricted to which γ is minimizing, as defined in part (1).

When I is compact, you might worry that definition (2) requires more than definition (1), but you'll see in Exercise 5.16 that the definitions are equivalent in this case. So there is no inconsistency, although (2) is really intended to be applied only to noncompact intervals such as (a, b) , $[a, \infty)$, and $(-\infty, \infty)$.

A primary goal of the next section is to prove that a constant-speed curve is locally minimizing if and only if it is a geodesic. This fact will obviate the need for the term "locally minimizing curve"—we'll instead just say "geodesic."

EXERCISES

EXERCISE 5.1. Explain why the rulings of every ruled surface (Exercise 3.29 on page 139) are geodesics.

EXERCISE 5.2. Let C be the cylinder from Example 5.7. Let $p, q \in C$ be points with different heights (different z -coordinates). Prove that there are infinitely many distinct geodesic segments in C between p and q , where "distinct" means having different traces.

EXERCISE 5.3. If a tubular neighborhood of a space curve is parametrized as in Eq. 4.22 on page 239, prove that the θ -parameter curves are geodesics.

EXERCISE 5.4. Let S be the surface of revolution generated by the curve $\gamma(t) = (x(t), 0, z(t))$. Prove that the latitude $t = t_0$ is a geodesic if and only if $x'(t_0) = 0$.

EXERCISE 5.5. Describe all geodesics in the cone

$$C = \{(x, y, z) \in \mathbb{R}^3 \mid z^2 = x^2 + y^2, z > 0\}.$$

Are there any closed geodesics?

EXERCISE 5.6. If two geodesics in a regular surface have the same domain I and the same position and velocity at a single point $t_0 \in I$, prove that they must be equal. *You may use Proposition 5.3(2), which contains a weaker version of this claim.*

EXERCISE 5.7. For a surface of revolution, prove that every segment of a longitudinal curve is uniquely minimizing.

EXERCISE 5.8. On the paraboloid $P = \{(x, y, z) \in \mathbb{R}^3 \mid z = x^2 + y^2\}$, define $\gamma : \mathbb{R} \rightarrow P$ as $\gamma(t) = (t, 0, t^2)$. Verify that γ is a geodesic. Prove that γ is not minimizing (although it follows from the previous exercise that the restriction of γ to $[0, \infty)$ is minimizing).

EXERCISE 5.9. Prove part (2) of Clairaut's theorem.

EXERCISE 5.10. Describe the behavior of the geodesics with the initial positions and velocities shown in Fig. 5.5.

EXERCISE 5.11. Prove that the hyperboloid of one sheet, $S = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 - z^2 = 1\}$, has exactly one closed geodesic (up to reparametrization).

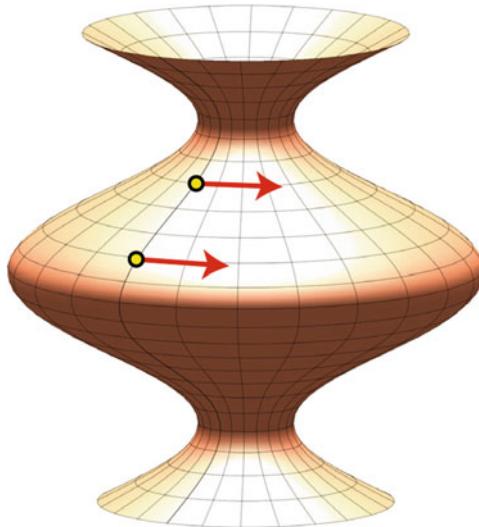


FIGURE 5.5. Where do the geodesics in these directions go?

EXERCISE 5.12. Let S be the surface of revolution obtained by revolving the bell curve (the graph of $z = e^{-x^2}$ in the xz -plane) about the z -axis. Prove that the trace of every geodesic in S is bounded.

EXERCISE 5.13. Let S be a connected regular surface. Prove that there exists a piecewise regular curve in S between an arbitrary pair of points of S .

HINT: By Exercise 3.15 on page 136, S is path-connected, so there is a continuous curve γ between the points. By compactness, γ is covered by finitely many coordinate charts. The segment of γ covered by any single coordinate chart can be replaced by a regular curve.

EXERCISE 5.14. Prove Lemma 5.10.

EXERCISE 5.15. Let S be a connected regular surface and let $p_0 \in S$. Prove that the function $d_{p_0} : S \rightarrow \mathbb{R}$ defined as $d_{p_0}(q) = d(p_0, q)$ is continuous. Is it smooth?

EXERCISE 5.16. In Definition 5.12, When I is compact, prove that (1) and (2) are equivalent. That is, if $\gamma : [a, b] \rightarrow S$ (uniquely) minimizes the distance between its endpoints, then it also does so between every pair of intermediate points that it visits.

EXERCISE 5.17. Let S be a connected regular surface.

- (1) If a geodesic γ in S is contained in a plane in \mathbb{R}^3 , prove that it is a line of curvature (Exercise 4.18 on page 211).
- (2) If every geodesic in S is contained in a plane in \mathbb{R}^3 , prove that S is contained in either a plane or a sphere. *HINT: use Exercise 4.19 on page 212.*

EXERCISE 5.18. Prove that a diffeomorphism $f : S \rightarrow \tilde{S}$ between connected regular surfaces is an isometry if and only if $\tilde{d}(f(p), f(q)) = d(p, q)$ for all $p, q \in S$, where d and \tilde{d} denote their intrinsic distance functions.

EXERCISE 5.19. Give an example of an isometry $f : S \rightarrow \tilde{S}$ between connected regular surfaces that does *not* satisfy the property $|f(p) - f(q)| = |p - q|$ for all $p, q \in S$. In other words, describe an isometry that does not preserve the *extrinsic* distance function inherited from the ambient \mathbb{R}^3 .

EXERCISE 5.20. Let S be a surface of revolution with Gaussian curvature $K < 0$. Prove that S has at most one simple closed geodesic (up to reparametrization). *HINT: Use Eq. 4.15 on page 221 and Clairaut's theorem.*

2. The Exponential Map

In this section, we will prove that every sufficiently short geodesic segment is minimizing. The main tool for proving this and other fundamental results is the *exponential map*, which describes the behavior (at least for a small amount of time) of *all* geodesics beginning at a single point of a surface.

DEFINITION 5.13.

Let S be a regular surface and $p \in S$. Choose the largest possible $\epsilon = \epsilon(p, 1)$ (possibly infinite) guaranteed by Proposition 5.3. Set $B_\epsilon = \{v \in T_p S \mid |v| < \epsilon\}$ (interpreted as $B_\epsilon = T_p S$ if $\epsilon = \infty$). The **exponential map** of S at p is the function $\exp_p : B_\epsilon \rightarrow S$, defined as

$$\exp_p(v) = \begin{cases} p & \text{if } v = \mathbf{0} \text{ (the zero vector),} \\ \gamma_v(1) & \text{for all other } v \in B_\epsilon, \end{cases}$$

where γ_v denotes the geodesic in S with $\gamma(0) = p$ and $\gamma'(0) = v$.

Equation 5.1 (on page 249) provides a simpler alternative description of the exponential map; namely, for every unit-length vector $u \in T_p S$ and $0 < r < \epsilon$, we have

$$\exp_p(ru) = \gamma_u(r).$$

Thus, $\exp_p(ru)$ is the point of S reached by traveling for time r (and hence traveling a distance r) along the *unit-speed* geodesic away from p in the direction u . In particular, the line $t \mapsto tu$ in $T_p S$ is mapped by \exp_p to the geodesic $t \mapsto \gamma_u(t)$ in S . In summary, the composition of \exp_p with any unit-speed line through the origin in $T_p S$ is a unit-speed geodesic through p in S ; see Fig. 5.6.

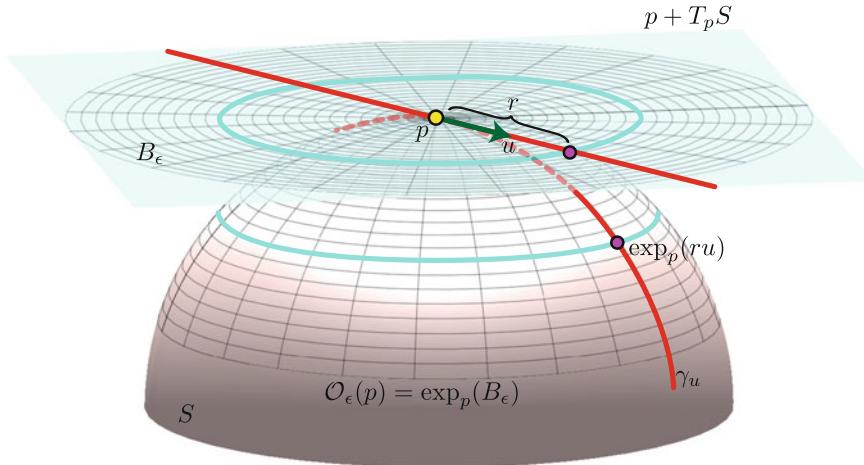


FIGURE 5.6. $\exp_p : B_\epsilon \rightarrow S$ maps radial lines to geodesics

Proposition 5.3 implies that \exp_p is smooth. Even better, we have the following:

PROPOSITION 5.14.

Let S be a regular surface and $p \in S$. There exists $\epsilon > 0$ (possibly smaller than the value from Definition 5.13) such that \exp_p is a diffeomorphism from B_ϵ onto its image $\mathcal{O}_\epsilon(p) = \exp_p(B_\epsilon)$.

This image $\mathcal{O}_\epsilon(p)$ is called a **normal neighborhood** of p in S with radius ϵ .

PROOF. According to the inverse function theorem for surfaces (on page 143), it will suffice to prove that its derivative at $\mathbf{0}$,

$$d(\exp_p)_\mathbf{0} : T_{\mathbf{0}}(T_p S) \rightarrow T_{\exp_p(\mathbf{0})} S,$$

is invertible. The domain and codomain here can be redescribed as

$$d(\exp_p)_\mathbf{0} : T_p S \rightarrow T_p S.$$

In fact, $d(\exp_p)_\mathbf{0}$ is just the identity map on $T_p S$, because it sends every $v \in T_p S$ to the initial velocity vector of the geodesic $t \mapsto \gamma_v(t)$, which is v . \square

EXAMPLE 5.15 (Maximal normal neighborhoods of the punctured plane, sphere, and cylinder). At a point p of the punctured plane in Example 5.5, one must choose $\epsilon \leq |p|$ in order for \exp_p to be defined on B_ϵ . Under this restriction, \exp_p is automatically a diffeomorphism from B_ϵ onto its image; in fact, it's an isometry.

The family of all geodesics on the sphere S^2 and the cylinder C were described in Examples 5.6 and 5.7 respectively. The domain of each geodesic was all of \mathbb{R} ; in other words, $\epsilon(p, 1) = \infty$ at each point p of either surface, so the domain of \exp_p is the entire tangent plane at p . But in order for the restriction of \exp_p to B_ϵ to be a diffeomorphism onto its image, one must choose $\epsilon \leq \pi$; every such choice serves as the radius of a valid normal neighborhood of every point of either surface. A value $\epsilon > \pi$ would not work, because \exp_p would not be one-to-one on B_ϵ . To visualize this on the cylinder, think of C as the surface of your finger, and of B_ϵ as a round Band-Aid that \exp_p wraps around your finger, overlapping on the other side if the Band-Aid's diameter is greater than your finger's circumference.

We will learn in Sect. 4 that \exp_p is defined on the entire tangent plane at every point p of every closed regular surface. So for closed surfaces, the only restriction on the radius of a normal neighborhood comes from the local diffeomorphism consideration.

If we use an orthonormal basis $\{e_1, e_2\}$ of $T_p S$ to identify $\mathbb{R}^2 \cong T_p S$ via the identification $(u, v) \leftrightarrow (ue_1 + ve_2)$, then we can consider \exp_p to be a surface patch. The corresponding local variables $\{u, v\}$ are called *normal*

coordinates at p . If this surface patch is composed with the diffeomorphism f from Example 3.6 (on page 117) that converts from polar to rectangular coordinates, then the local variables $\{r, \theta\}$ for the resulting surface patch are called *normal polar coordinates* at p . All of this is illustrated in Fig. 5.7 and summarized in the following definition:

DEFINITION 5.16.

Let S be a regular surface and $p \in S$. Let $\epsilon > 0$ be the radius of a normal neighborhood of p . Let $\{e_1, e_2\}$ be an orthonormal basis of $T_p S$.

- (1) **Normal coordinates** at p are the local variables $\{u, v\}$ corresponding the surface patch

$$\mu(u, v) = \exp_p(ue_1 + ve_2)$$

with domain $\{(u, v) \in \mathbb{R}^2 \mid u^2 + v^2 < \epsilon^2\}$.

- (2) **Normal polar coordinates** at p are the local variables $\{r, \theta\}$ corresponding the surface patch

$$\sigma(r, \theta) = \exp_p((r \cos \theta)e_1, (r \sin \theta)e_2)$$

with domain $\{(r, \theta) \in \mathbb{R}^2 \mid 0 < r < \epsilon, 0 < \theta < 2\pi\}$.

While normal coordinates parametrize a normal neighborhood of p in S , the surface patch for normal polar coordinates does not cover p or the geodesic segment in the direction e_1 (the dashed green line in Fig. 5.7). This is unfortunate, since p is the center of the story, but it will prove to be only a minor inconvenience.

In Fig. 5.7, it is not a coincidence that the vectors labeled σ_r and σ_θ appear to be orthogonal. Gauss proved that this will always happen:

LEMMA 5.17 (Gauss's Lemma).

Let S be a regular surface and $p \in S$. The first fundamental form of the surface patch σ for normal polar coordinates at p has the form

$$\mathcal{F}_1 = dr^2 + G d\theta^2.$$

In other words, $\langle \sigma_r, \sigma_r \rangle = 1$ and $\langle \sigma_r, \sigma_\theta \rangle = 0$ on the entire domain of σ .

PROOF. As usual, $\mathcal{F}_1 = E dr^2 + 2F dr d\theta + G d\theta^2$. But $E = 1$, because the r -parameter curves are unit-speed geodesics. It therefore remains to prove that $F = \langle \sigma_r, \sigma_\theta \rangle$ is everywhere zero. For this, observe that σ_{rr} is everywhere normal to S , because it is the acceleration vector of a geodesic. Thus

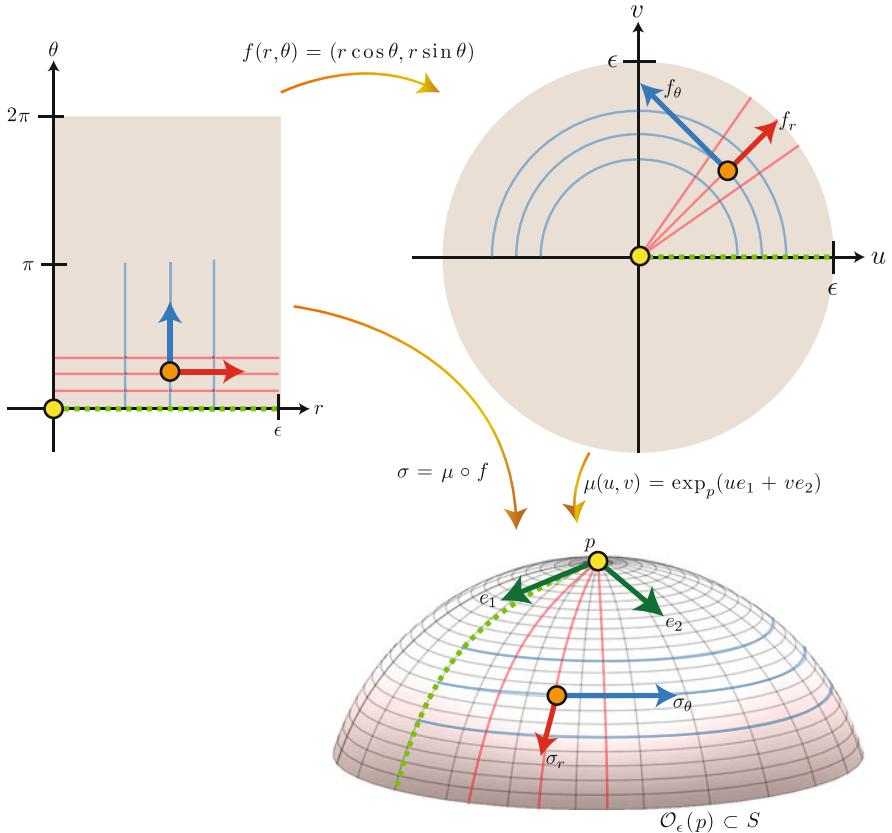


FIGURE 5.7. $\{u, v\}$ are normal coordinates, while $\{r, \theta\}$ are normal polar coordinates

$$\begin{aligned} F_r &= \frac{\partial}{\partial r} \langle \sigma_r, \sigma_\theta \rangle = \langle \sigma_{rr}, \sigma_\theta \rangle + \langle \sigma_r, \sigma_{\theta r} \rangle = 0 + \langle \sigma_r, \sigma_{\theta r} \rangle \\ &= \langle \sigma_r, \sigma_{r\theta} \rangle = (1/2) \frac{\partial}{\partial \theta} \langle \sigma_r, \sigma_r \rangle = (1/2) E_\theta = 0. \end{aligned}$$

For every fixed $\theta_0 \in (0, 2\pi)$, the function $r \mapsto F(r, \theta_0)$ is therefore constant. We wish to prove that its constant value equals zero. For this, write $\sigma = \mu \circ f$ as in Fig. 5.7. Since $\lim_{r \rightarrow 0} |f_\theta(r, \theta_0)| = 0$, and $d(\exp_p)_0$ is the identity, the chain rule gives that $\lim_{r \rightarrow 0} |\sigma_\theta(r, \theta_0)| = 0$. Therefore,

$$\lim_{r \rightarrow 0} F(r, \theta_0) = \lim_{r \rightarrow 0} \langle \sigma_\theta(r, \theta_0), \sigma_r(r, \theta_0) \rangle = 0.$$

□

Gauss's lemma is about normal *polar* coordinates, but it implies the following important fact about normal coordinates:

COROLLARY 5.18 (Signed curvature \leftrightarrow geodesic curvature).

Let S be a regular surface and $p \in S$. Let $\mu : U \rightarrow S$ denote the surface patch for normal coordinates at p . Let $\mathbf{0} \in U$ denote the origin.

- (1) The vectors $\mu_{uu}(\mathbf{0})$, $\mu_{vv}(\mathbf{0})$, and $\mu_{uv}(\mathbf{0})$ are all orthogonal to $T_p S$.
- (2) Assume that S is oriented, and that the basis $\{e_1, e_2\}$ (with respect to which μ was defined) is positively oriented. If $\hat{\gamma} : I \rightarrow U$ is a regular curve with $\hat{\gamma}(0) = \mathbf{0}$, and $\gamma = \mu \circ \hat{\gamma}$ has unit speed, then the signed curvature of $\hat{\gamma}$ at $t = 0$ equals the geodesic curvature of γ at $t = 0$, as in Fig. 5.8.

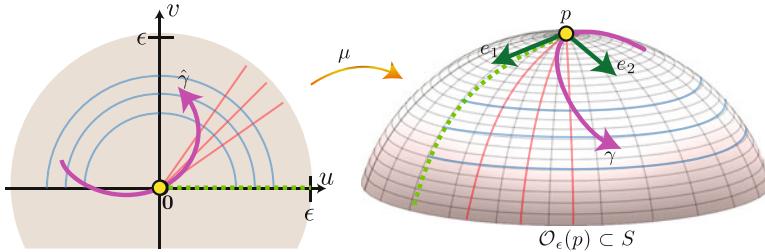


FIGURE 5.8. The signed curvature of $\hat{\gamma}$ at $\mathbf{0}$ equals the geodesic curvature of $\gamma = \mu \circ \hat{\gamma}$ at p

Part (2) is reminiscent of Exercise 4.15 (on page 211), but requires a different proof method.

PROOF. For part (1), $\mu_{vv}(\mathbf{0})$ is orthogonal to $T_p S$, because $v \mapsto \mu_{vv}(0, v)$ is the acceleration vector field of the geodesic $v \mapsto \mu(0, v)$. Similarly, $\mu_{uu}(\mathbf{0})$ is orthogonal to $T_p S$. The mixed partial is handled by observing that

$$\langle \mu_{uv}, \mu_v \rangle = \frac{\partial}{\partial v} \langle \mu_u, \mu_v \rangle - \langle \mu_u, \mu_{vv} \rangle.$$

The first term on the right side of this equation, $\frac{\partial}{\partial v} \langle \mu_u, \mu_v \rangle$, vanishes at $\mathbf{0}$, because Gauss's lemma says that μ_u and μ_v remain orthogonal along the geodesic $v \mapsto \mu(0, v)$. The second term vanishes at $\mathbf{0}$, because $\mu_{vv}(\mathbf{0})$ is orthogonal to $T_p S$. Thus, the left side of this equation, $\langle \mu_{uv}, \mu_v \rangle$, must also vanish at $\mathbf{0}$. By a similar argument $\langle \mu_{uv}, \mu_u \rangle$ also vanishes at $\mathbf{0}$. This proves that $\mu_{uv}(\mathbf{0})$ is orthogonal to $T_p S$.

For part (2), express the components of $\hat{\gamma}$ as $\hat{\gamma}(t) = (u(t), v(t))$, so that $\gamma(t) = \mu(u(t), v(t))$. Using the prime symbol $'$ to denote the derivative with respect to t , and assuming that everything is evaluated at an arbitrary $t \in I$ or at $(u(t), v(t))$ as appropriate, the chain rule gives

$$\gamma' = u' \mu_u + v' \mu_v, \quad \gamma'' = u'' \mu_u + u'(u' \mu_{uu} + v' \mu_{uv}) + v'' \mu_v + v'(u' \mu_{vu} + v' \mu_{vv}).$$

Combining this conclusion with part (1) gives

$$(\text{The projection of } \gamma''(0) \text{ onto } T_p S) = u''(0)\mu_u(\mathbf{0}) + v''(0)\mu_v(\mathbf{0}),$$

from which the result follows. \square

Gauss's lemma also finally allows us to keep our promise of proving that sufficiently short geodesic segments are minimizing:

THEOREM 5.19 (Minimizing within a normal neighborhood).

Let S be a regular surface, $p \in S$, and $\mathcal{O} = \mathcal{O}\epsilon(p)$ a normal neighborhood of p . Then every geodesic segment in \mathcal{O} beginning at p is uniquely minimizing.

Thus, every geodesic segment beginning at p is the unique shortest curve in S between p and every other point that it visits before leaving \mathcal{O} . The following proof roughly generalizes our proof on page 14 that straight lines are shortest paths in \mathbb{R}^n :

PROOF OF PROPOSITION 5.19.

Let $\sigma : U \subset \mathbb{R}^2 \rightarrow \mathcal{O} \subset S$ be the surface patch for normal polar coordinates at p . Let $q_0 \in \mathcal{O}$ with $q_0 \neq p$. After possibly choosing $\{e_1, e_2\}$ again, we can ensure that q_0 is covered by σ ; this means that q_0 is not on the trace, C , of the geodesic in the direction of e_1 , which is shown as a dashed green line in the figures. Thus, $q_0 = \sigma(r_0, \theta_0)$ for some $(r_0, \theta_0) \in U$. The curve defined as

$$\gamma(t) = \begin{cases} p & \text{if } t = 0, \\ \sigma(t, \theta_0) & \text{if } t \in (0, r_0], \end{cases}$$

is a unit-speed geodesic segment from p to q_0 with length equal to r_0 . We must prove that γ is the unique shortest curve in S from p to q_0 .

Let $\beta : [a, b] \rightarrow S$ be another curve starting at $\beta(a) = p$ and ending at $\beta(b) = q_0$; see Fig. 5.9. We assume for simplicity that β is regular; the case in which β is only piecewise regular is left for the reader in Exercise 5.23. Let L denote the arc length of β . Assume for now that the image of β (except for p) is covered by σ , so we can write $\beta(s) = \sigma(r(s), \theta(s))$, where $s \mapsto (r(s), \theta(s))$, $s \in (a, b]$, is a regular curve in U .

Lemma 3.77 (on page 183) describes arc length in terms of the first fundamental form, which in our case looks like $\mathcal{F}_1 = dr^2 + G d\theta^2$ according to Gauss's lemma. Interpret the following as improper integrals, since $r(t)$ and $\theta(t)$ are not defined at $t = a$:

$$\begin{aligned} L &= \int_a^b \sqrt{r'(t)^2 + G(r)\theta'(t)^2} dt \geq \int_a^b \sqrt{r'(t)^2 + 0} dt \\ &= \int_a^b |r'(t)| dt \geq \int_a^b r'(t) dt = r(b) - \lim_{t \rightarrow a} r(t) = r(b) = r_0, \end{aligned}$$

with equality if and only if $\theta' = 0$, which means that $\theta = \theta_0$ is constant. In other words, $L \geq r_0$, with equality if and only if β is a reparametrization of γ .

It remains to discuss the possibility that the image of β is not covered by σ . If β crosses C finitely many times, then partition $[a, b]$ into finitely many subintervals on the interiors of which β avoids C . The above argument shows that the length of β on each subinterval is $\geq \Delta r$ (the change in r between its endpoints), so $L \geq \sum \Delta r = r_0$. On the other hand, if β crosses C infinitely many times, no matter how the basis $\{e_1, e_2\}$ is chosen, then β must spiral infinitely many times around p as it approaches p , which contradicts the regularity of β at $t = a$. Finally, if β leaves \mathcal{O} , then the above argument shows that the length of the segment of β up until it first leaves \mathcal{O} is $\geq \epsilon > r_0$. \square

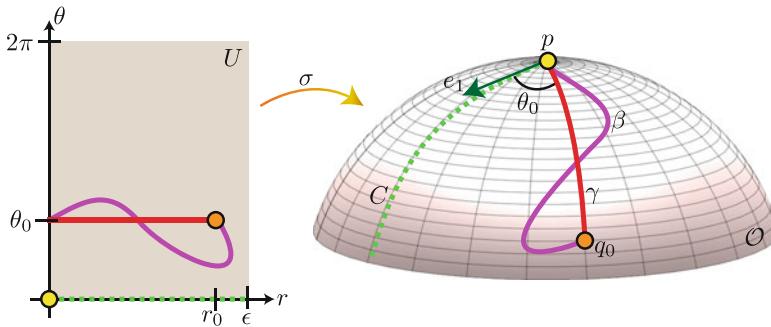


FIGURE 5.9. $\text{length}(\beta) \geq \text{length}(\gamma)$

Theorem 5.19 implies in particular that a normal neighborhood of p equals the set of points of S whose intrinsic distance to p is less than the radius:

$$\mathcal{O}_\epsilon(p) = \{q \in S \mid d(p, q) < \epsilon\}.$$

The theorem becomes even more powerful when the radius of the normal neighborhood is chosen small enough to guarantee an even stronger property:

PROPOSITION 5.20 (Better-than-normal neighborhoods).

If S is a regular surface and $p \in S$, then a normal neighborhood, \mathcal{O} , of p with sufficiently small radius has the added property that it is contained in a normal neighborhood of each of its points. In particular, this implies that every geodesic segment in \mathcal{O} (not necessarily beginning at p) is uniquely minimizing.

PROOF. The real work lies in solving Exercise 5.26, which says that there exist a neighborhood, W , of p in S and a single value $\delta > 0$ that serves as the radius of a valid normal neighborhood of each point of W .

Assuming this, choose $\epsilon > 0$ such that $\epsilon < \delta/2$ and such that the normal neighborhood $\mathcal{O}_\epsilon(p)$ is contained in W . By the triangle inequality, this choice ensures that $\mathcal{O}_\epsilon(p) \subset \mathcal{O}_\delta(q)$ for each $q \in \mathcal{O}_\epsilon(p)$; in other words, $\mathcal{O}_\epsilon(p)$ is contained in a normal neighborhood of each of its points; see Fig. 5.10. The second claim follows from Proposition 5.19. \square

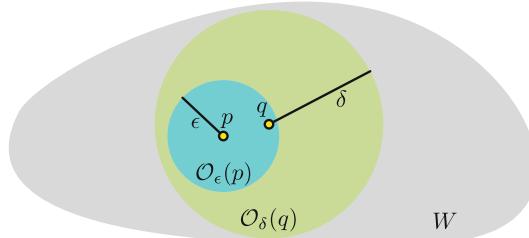


FIGURE 5.10. If $\epsilon < \delta/2$, then $\mathcal{O}_\epsilon(p) \subset \mathcal{O}_\delta(q)$ for all $q \in \mathcal{O}_\epsilon(p)$

Such a better-than-normal neighborhood, \mathcal{O} , has this property: between each pair of points in \mathcal{O} , there exists a unique minimizing geodesic β . But unfortunately, the trace of β is not guaranteed to lie in \mathcal{O} ; it is guaranteed to lie only in the encompassing normal neighborhoods of either endpoint. But it turns out to be possible to choose the radius of \mathcal{O} small enough to ensure that β remains in \mathcal{O} :

PROPOSITION 5.21 (Convex neighborhoods).

*If S is a regular surface and $p \in S$, then a normal neighborhood, \mathcal{O} , of p with sufficiently small radius has the added property that between each pair of points in \mathcal{O} there exists a unique minimizing geodesic, and the trace of this geodesic lies in \mathcal{O} . Such a neighborhood is called a **convex neighborhood**.*

PROOF. Exercise 5.28, with hints. \square

Our term “better-than-normal” is nonstandard. This property doesn’t really need to be named, because it is just a stepping-stone toward the stronger convexity property.

EXAMPLE 5.22. As discussed in Example 5.15, $\delta = \pi$ is the maximal radius of a normal neighborhood of every point p of either the sphere S^2 or the cylinder C . Since this choice does not depend on p , the above proof shows that half this amount, $\epsilon = \pi/2$, is a radius that ensures the better-than-normal property of Proposition 5.20. On S^2 , no larger choice of ϵ would work. What about on C ?

One consequence of Proposition 5.20 is that geodesics can be characterized by their locally minimizing property:

COROLLARY 5.23.

A constant-speed curve in S is a geodesic if and only if it is locally minimizing.

PROOF. Let $\gamma : I \rightarrow S$ be a constant-speed curve, $t_0 \in I$, and $\epsilon > 0$ any number smaller than the radius of a better-than-normal neighborhood of $\gamma(t_0)$ (as in Proposition 5.20). If γ is a geodesic, then the restriction of γ to $[t_0 - \epsilon, t_0 + \epsilon] \cap I$ is minimizing.¹

Conversely, if γ is locally minimizing, then the uniqueness assertion of Proposition 5.20 forces this restriction of γ to be a geodesic. \square

In light of this characterization of geodesics, Proposition 5.3 (on page 248) roughly says that at a given point of a regular surface, there is a unique path along which you can walk in any given direction that provides the shortest route between the breadcrumbs you drop along the way.

A significant consequence of the locally minimizing characterization of geodesics is the fact that isometries preserve geodesics:

COROLLARY 5.24 (Geodesics are intrinsic).

If $f : S \rightarrow \tilde{S}$ is an isometry between regular surfaces, and γ is a geodesic in S , then $f \circ \gamma$ is a geodesic in \tilde{S} .

This fact was not obvious from Definition 5.1 (on page 248), but its proof follows quickly from the locally minimizing characterization of geodesics, and is left to the reader in Exercise 5.21.

Inhabitants of a domain $U \subset \mathbb{R}^2$ who know that their world is identified with some surface can (in principle) determine which curves in their world correspond to geodesics in this surface, provided they have knowledge of the functions E, F, G . Explicit formulas for doing so will be given in Sect. 6 of this chapter. In fact, not only can these inhabitants determine which curves correspond to geodesics, they can also measure how much an arbitrary curve deviates from being a geodesic:

COROLLARY 5.25 (Geodesic curvature is intrinsic).

If $f : S \rightarrow \tilde{S}$ is an orientation-preserving isometry between oriented regular surfaces, then every unit speed curve $\gamma : I \rightarrow S$ has the same geodesic curvature function as the composition $\tilde{\gamma} = f \circ \gamma$ in \tilde{S} .

¹Better-than-normal really is required here; if it were only normal, then γ would be minimizing on $[t_0 - \epsilon, t_0]$ and on $[t_0, t_0 + \epsilon]$, but not necessarily on $[t_0 - \epsilon, t_0 + \epsilon]$; think about the example $S = S^2$ with $\epsilon = \pi$.

PROOF. Let $t_0 \in I$. Define $p_0 = \gamma(t_0)$ and $\tilde{p}_0 = f(p_0) = \tilde{\gamma}(t_0)$. Let $\mu : U \rightarrow S$ be the surface patch for normal coordinates at p , defined with respect to a positively oriented orthonormal basis $\{e_1, e_2\}$ of $T_p S$. Since f sends unit-speed geodesics to unit-speed geodesics (Corollary 5.24), $f \circ \mu$ equals the surface patch for normal coordinates of \tilde{S} at $f(p)$ (with respect to the positively oriented orthonormal basis $\{df_p(e_1), df_p(e_2)\}$). Proposition 5.18 now provides the same geometric interpretation of the geodesic curvature at time t_0 of both γ and $\tilde{\gamma}$, namely the signed curvature of $\mu^{-1} \circ \gamma$ at the origin of U . \square

EXERCISES

EXERCISE 5.21. Prove Corollary 5.24.

EXERCISE 5.22. Explicitly describe the surface patch for normal polar coordinates when $S = S^2$, $p = (0, 0, 1)$, $e_1 = (1, 0, 0)$, and $e_2 = (0, 1, 0)$.

EXERCISE 5.23. We simplified the proof of Proposition 5.19 by assuming that β is regular. How do you argue if β is only piecewise regular?

EXERCISE 5.24. Prove that every regular surface can be covered by an atlas of surface patches with $E = 1$ and $F = 0$. *HINT: Normal polar coordinates work, but you must deal with the issue that normal polar coordinates at p don't cover p .*

EXERCISE 5.25. In Fig. 5.4 on page 252, the purple geodesic is asymptotic to the light-blue latitudinal curve.

- (1) Prove that a geodesic on a surface of revolution could not be asymptotic to a latitudinal curve unless the latitudinal curve is itself a geodesic.
- (2) Rigorously justify the assertions in the discussion of Fig. 5.4.
- (3) Let $\gamma : \mathbb{R} \rightarrow S$ be a geodesic in the paraboloid

$$S = \{(x, y, z) \in \mathbb{R}^3 \mid z = x^2 + y^2\}.$$

Prove that the height function, $h(x, y, z) = z$, composed with γ has exactly one critical point t_0 ; it is decreasing on $(-\infty, t_0)$ and increasing on (t_0, ∞) .

EXERCISE 5.26. If S is a regular surface and $p \in S$, prove that there exist a neighborhood, W , of p in S and a single value $\delta > 0$ that serves as the radius of a valid normal neighborhood of each point of W .

HINT: Let $\epsilon > 0$ be any radius such that $\mathcal{O}_{2\epsilon}(p)$ is a normal neighborhood of p . Set $\mathcal{U} = \{(q, v) \mid q \in \mathcal{O}_\epsilon(p), v \in T_q S, |v| < \epsilon\}$ and define $F : \mathcal{U} \rightarrow \mathcal{O}_\epsilon(p) \times \mathcal{O}_{2\epsilon}(p)$ as $F(q, v) = (q, \exp_q(v))$. The smoothness assertion of Proposition 5.3 on page 248 implies that F is well defined and smooth on all of \mathcal{U} (after possibly shrinking ϵ). Next apply the inverse function theorem to F at $(p, 0)$.

EXERCISE 5.27. Prove that a locally minimizing piecewise-regular curve must be regular.

EXERCISE 5.28 (Existence of Convex Neighborhoods). Let S be a regular surface and $p \in S$. Let $\mathcal{O}_\epsilon(p)$ be a normal neighborhood of p . Let $f : \mathcal{O}_\epsilon(p) \rightarrow \mathbb{R}$ denote the “squared distance to p ” function; that is, for all $q \in \mathcal{O}_\epsilon(p)$, $f(q)$ is the squared length of the shortest curve in S from p to q .

Prove that if ϵ is sufficiently small, then $\mathcal{O}_\epsilon(p)$ has the following added properties:

- (1) For every geodesic, β , in $\mathcal{O}_\epsilon(p)$, the function $f \circ \beta$ is concave up.
- (2) Between every pair of points in $\mathcal{O}_\epsilon(p)$, there exists a unique minimizing geodesic in S , and its trace lies in $\mathcal{O}_\epsilon(p)$.

HINT: For part (1), if β is of unit speed and $\beta(0) = p$, then $(f \circ \beta)''(0) = 2$. Use continuity to conclude that $(f \circ \beta)''(0) > 0$ if β is of unit speed and begins sufficiently near p . For part (2), choose ϵ such that 2ϵ works for (1) and such that the neighborhood is better-than-normal.

EXERCISE 5.29. In Definition 5.12 on page 255, we could have also defined “locally uniquely minimizing” in the obvious way. Prove that this would have been equivalent to “locally minimizing.”

EXERCISE 5.30. Let $\gamma : [a, b] \rightarrow S$ be a unit-speed curve in a regular surface. If γ is one-to-one, prove that there exists a single surface patch that covers the trace of γ .

EXERCISE 5.31. Let C be the trace of a simple closed curve in a regular surface S . Let $p \in S$ be a point not contained in C . Assume that p is close enough to C that a normal ball about p intersects C . Let $q \in C$ be the point of C closest to p . Prove that a minimizing geodesic from p to q must intersect C orthogonally.

EXERCISE 5.32. If $p, q \in S^2$ (the sphere), explain why $d(p, q) = \cos^{-1}(\langle p, q \rangle)$.



3. Gauss's Remarkable Theorem

In this section, we will present a simple and powerful formula for the Gaussian curvature in normal polar coordinates (to be proven later in Sect. 6). This formula will provide our best-yet geometric interpretation of Gaussian curvature; namely, $K(p)$ represents the “infinitesimal spreading” of the geodesics through p . Unlike our previous characterizations of Gaussian curvature, this interpretation is purely in terms of intrinsic measurements and constructions (including geodesics). Hence, the Gaussian curvature itself is intrinsic! When Gauss originally discovered this fact, he named it the *Theorema Egregium* (Latin for “remarkable theorem”). It represents one of the fundamental breakthroughs in the field. It all begins with the following result:

THEOREM 5.26.

Let S be a regular surface and $p \in S$. The Gaussian curvature in normal polar coordinates at p is given by

$$K = -\frac{(\sqrt{G})_{rr}}{\sqrt{G}},$$

where $\mathcal{F}_1 = dr^2 + G d\theta^2$ is the first fundamental form in these coordinates, as in Lemma 5.17.

The proof will be postponed until Sect. 6, so that we can focus now on geometric interpretations and applications, the most famous of which is the following:

COROLLARY 5.27 (Gauss's Theorema egregium).

Gaussian curvature is intrinsic. In other words, if $f : S \rightarrow \tilde{S}$ is an isometry between regular surfaces, then $K(p) = K(f(p))$ for all $p \in S$.

PROOF. Let $p \in S$ and let $\sigma : U \rightarrow S$ be the surface patch for normal polar coordinates of S at p . Since f sends unit-speed geodesics to unit-speed geodesics (Corollary 5.24 on page 266), $f \circ \sigma$ equals the surface patch for normal polar coordinates of \tilde{S} at $f(p)$. Since f is an isometry, the first fundamental forms of σ and $f \circ \sigma$ agree (Exercise 3.101 on page 186). This common first fundamental form can be expressed as $\mathcal{F}_1 = dr^2 + G d\theta^2$ (Lemma 5.17 on page 260). Theorem 5.26 expresses the Gaussian curvature purely in terms of this common first fundamental form; hence $K(x) = K(f(x))$ for all x covered by σ . Even though σ does not cover the point p itself, it does cover points arbitrarily close to p , so the continuity of K implies that $K(p) = K(f(p))$. \square

Admittedly, such a short proof is possible only because we've postponed the proof of Theorem 5.26, which is where the hard work lies. In fact, Sect. 6 contains a local coordinate formula for K with respect to an *arbitrary* surface patch in terms of $\{E, F, G\}$, from which follows an even shorter proof of Gauss's Theorema Egregium than the one above.

Recall that the Gaussian curvature equals the product of the principal curvatures, which are not individually intrinsic (Example 4.10 on page 207). The theorem roughly means that inhabitants of a domain $U \subset \mathbb{R}^2$ who know that their world is identified with some surface can (in principle) determine the Gaussian curvature of that surface, provided they know how to measure lengths of curves. It is surprising that this is enough to measure how the surface curves within the ambient \mathbb{R}^3 , given that the two-dimensional inhabitants of U could hardly fathom \mathbb{R}^3 .

Gauss's theorem might seem less surprising after we discuss some geometric characterizations of the Gaussian curvature that more obviously depend only on intrinsic measurements. For this, we'll study the particular expression for K in terms of G provided by Theorem 5.26, and use it to prove the following:

PROPOSITION 5.28 (Geodesic spreading).

Let S be a regular surface and $p \in S$. Denote the first fundamental form in normal polar coordinates at p by $\mathcal{F}_1 = dr^2 + G d\theta^2$. Fix $\theta_0 \in (0, 2\pi)$, and for all $r \in (0, \epsilon)$, define

$$g(r) = \sqrt{G(r, \theta_0)} = |\sigma_\theta(r, \theta_0)|.$$

Extending the domain of g to include $r = 0$ by defining $g(0) = 0$ makes g become smooth on $[0, \epsilon]$, and its third-order Taylor polynomial at $r = 0$ is

$$g(r) \approx r - \frac{r^3}{6} K(p).$$

In other words, $g'(0) = 1$, $g''(0) = 0$ and $g'''(0) = -K(p)$.

PROOF. The assertion that g extends smoothly to $r = 0$ is justified by viewing σ as a composition, as in Fig. 5.7 on page 261, so for $r \in (0, \epsilon)$, the chain rule gives

$$\begin{aligned} g(r) &= |\sigma_\theta(r, \theta_0)| = |d\sigma_{(r, \theta_0)}(0, 1)| \\ &= |d\mu_{(r \cos \theta_0, r \sin \theta_0)}(df_{(r, \theta_0)}(0, 1))| \\ &= |d\mu_{(r \cos \theta_0, r \sin \theta_0)}(-r \sin \theta_0, r \cos \theta_0)| \\ &= r |d\mu_{(r \cos \theta_0, r \sin \theta_0)}(-\sin \theta_0, \cos \theta_0)|. \end{aligned}$$

This last formula for g is smooth on the domain $r \in (-\epsilon, \epsilon)$. We also learn from this that $g'(0) = 1$, since $d\mu_{(0,0)}$ is the identity map.

Define $K(r) = K(r, \theta_0)$, so Theorem 5.26 says that

$$(5.5) \quad g''(r) = -g(r)K(r)$$

for all $r \in (0, \epsilon)$, and thus also for $r = 0$ by continuity. In particular,

$$g''(0) = -g(0)K(0) = 0.$$

Finally, differentiating Eq. 5.5 gives

$$g'''(0) = -g'(0)K(0) - g(0)K'(0) = -g'(0)K(0) - 0 = -K(p).$$

□

In Proposition 5.28, r represents distance along the geodesic $\gamma_0(r) = \sigma(\theta_0, r)$. As illustrated in Fig. 5.11, the growth rate of $g(r)$ visually represents how quickly the other geodesics through p (the r -parameter curves of σ with

θ fixed at other values close to θ_0) spread away from γ_0 . If two friends walk away from p along two different unit-speed geodesics (forming a small angle $\Delta\theta$ at p), how quickly will their paths diverge? Their separation after a short time r will be approximately

$$(5.6) \quad \text{distance} \approx (2\pi\Delta\theta)g(r).$$

The first-order Taylor polynomial of g is $g(r) \approx r$, so their paths diverge as quickly as if they lived on a plane. This makes sense, because S is approximated to first order by its tangent plane. But the third-order Taylor polynomial of g detects the curvature of S . If $K(p) > 0$, then their paths initially diverge more slowly than if they lived on a plane. If $K(p) < 0$, then they separate more quickly.

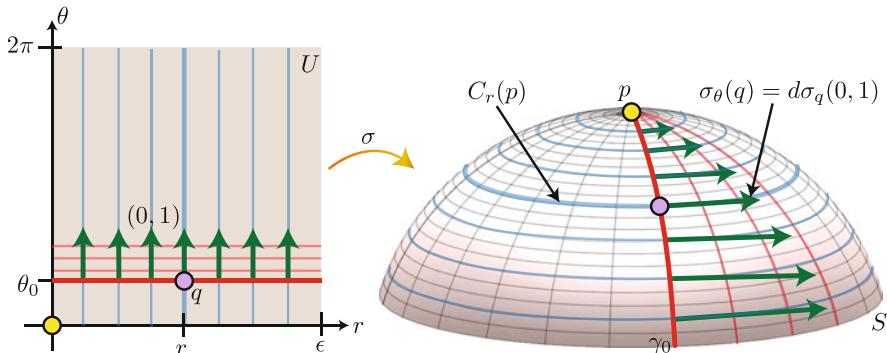


FIGURE 5.11. $K(p)$ represents the rate at WHICH unit-speed geodesics through p initially spread apart

Notice that the third-order Taylor polynomial of g depends only on $K(p)$ (although a higher-order polynomial might also depend on θ_0). Thus, the third-derivative prediction is that pairs of nearby geodesics in *every* direction spread apart at the same approximate initial rate. If 360 friends all walk away from p along unit-speed geodesics with angles regularly spaced at $\Delta\theta = 1^\circ$, they will remain roughly uniformly spread out along the circle they form a short time later. Such circles are colored light blue in Fig. 5.11 and are called *distance circles*:

DEFINITION 5.29.

The **distance circle** of radius $r > 0$ about a point p in a regular surface S is $C_r(p) = \{q \in S \mid d(p, q) = r\}$.

For small r , notice that $C_r(p)$ is the closure of the trace of the θ -parameter curve in normal polar coordinates at p with that fixed r -value. It would have been more precise in Eq. 5.6 to clarify that “distance” means “distance along $C_r(p)$.”

The interesting question is, how quickly do the circumferences of the distance circles grow? The infinitesimal answer is contained in the following proposition:

PROPOSITION 5.30 (Distance circle growth rate).

At every point p of a regular surface S , the third-order Taylor polynomial of $r \mapsto \text{length}(C_r(p))$ at $r = 0$ is

$$\text{length}(C_r(p)) \approx 2\pi r - \frac{\pi}{3}r^3 K(p).$$

PROOF. Proposition 5.28 says that for each fixed θ_0 , $g(r) = r - \frac{r^3}{6}K(p) + \epsilon(r, \theta_0)$ for some error term $\epsilon(r, \theta_0)$ that satisfies $\lim_{r \rightarrow 0} \frac{\epsilon(r, \theta_0)}{r^3} = 0$. Thus,

$$\begin{aligned} \text{length}(C_r(p)) &= \int_0^{2\pi} |\sigma_\theta(r, \theta)| d\theta = \int_0^{2\pi} \left(r - \frac{r^3}{6}K(p) + \epsilon(r, \theta) \right) d\theta \\ &= 2\pi \left(r - \frac{r^3}{6}K(p) \right) + \int_0^{2\pi} \epsilon(r, \theta) d\theta \approx 2\pi \left(r - \frac{r^3}{6}K(p) \right), \end{aligned}$$

because the first three derivatives of $r \mapsto \int_0^{2\pi} \epsilon(r, \theta) d\theta$ at $r = 0$ vanish. \square

The first-order Taylor polynomial is $\text{length}(C_r(p)) \approx 2\pi r$, which says that distance circles in S grow as they do in the plane. But the third-order Taylor polynomial detects the curvature of S . If $K(p) > 0$ (as in Fig. 5.11), then a distance circle of small radius about p is shorter than a circle of the same radius in the plane. If $K(p) < 0$, then it is longer.

Solving for $K(p)$ yields the following geometric characterization of Gaussian curvature:

$$(5.7) \quad K(p) = \lim_{r \rightarrow 0} \frac{3}{\pi} \frac{2\pi r - \text{length}(C_r(p))}{r^3}.$$

This characterization depends only on intrinsic measurements and constructions (including geodesics), so the Theorema Egregium would be obvious if this were taken as the definition of Gaussian curvature. Notice that the numerator equals the difference in length between $C_r(p)$ and a circle in the plane with the same radius.

We next wish to interpret the above results in the case of surfaces with *constant* Gaussian curvature. This is an extremely natural class of surfaces to study, in part because other surfaces can be understood by comparison to them. For example, in a sufficiently small neighborhood of a point p of a general surface, $K(p)$ is approximately constant, so geodesics will spread apart from p approximately as they would on a surface with the same constant Gaussian curvature. When K is constant, we can do better than just a Taylor polynomial for $g(r)$; we can solve Eq. 5.5 to obtain an explicit formula for $g(r)$:

LEMMA 5.31.

If $g : [0, \epsilon) \rightarrow \mathbb{R}$ is a smooth function satisfying

$$g(0) = 0 \quad \text{and} \quad g'(0) = 1 \quad \text{and} \quad g''(t) = -\lambda \cdot g(t) \quad \text{for all } t \in [0, \epsilon),$$

where $\lambda \in \mathbb{R}$ is a constant, then

$$g(t) = \begin{cases} \frac{1}{\sqrt{\lambda}} \sin(\sqrt{\lambda}t) & \text{if } \lambda > 0, \\ t & \text{if } \lambda = 0, \\ \frac{1}{\sqrt{|\lambda|}} \sinh(\sqrt{|\lambda|}t) & \text{if } \lambda < 0. \end{cases}$$

We omit the proof, which requires a differential equations background. The three possibilities are graphed in Fig. 5.12, together with illustrations of the geodesic spreading on surfaces with the corresponding Gaussian curvature sign.

Thus, if S has constant curvature $K = \lambda$ in Theorem 5.26, then G depends only on r (not on θ) and is completely determined by λ . This will allow us to prove the following:

PROPOSITION 5.32 (Homogeneity of constant-curvature surfaces).

If S and \tilde{S} are regular surfaces with the same constant Gaussian curvature $K = \lambda$, then they are **locally isometric**; that is, for every $p \in S$ and $\tilde{p} \in \tilde{S}$, there exist a neighborhood \mathcal{O} of p in S and a neighborhood $\tilde{\mathcal{O}}$ of \tilde{p} in \tilde{S} such that \mathcal{O} and $\tilde{\mathcal{O}}$ are isometric.

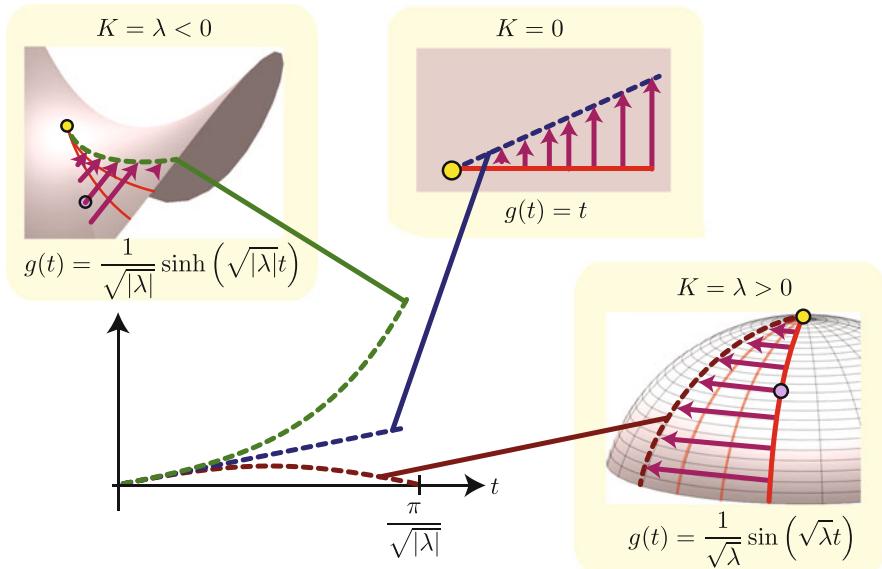


FIGURE 5.12. Geodesics spread apart according to the sign of the Gaussian curvature

PROOF. Let $p \in S$ and $\tilde{p} \in \tilde{S}$. Let $\sigma : U \rightarrow V \subset S$ and $\tilde{\sigma} : U \rightarrow \tilde{V} \subset \tilde{S}$ be the surface patches for normal polar coordinates of S at p and of \tilde{S} at \tilde{p} respectively. Notice that σ and $\tilde{\sigma}$ have the same domain, which is achieved by choosing a radius small enough to work at both points. Express their first fundamental forms as $\mathcal{F}_1 = dr^2 + G d\theta^2$ and $\tilde{\mathcal{F}}_1 = dr^2 + \tilde{G} d\theta^2$. Lemma 5.31 gives the same solution for G and \tilde{G} . In particular, $G = \tilde{G}$.

Consider the diffeomorphism $f = \tilde{\sigma} \circ \sigma^{-1} : V \rightarrow \tilde{V}$. Since $\tilde{\sigma} = f \circ \sigma$, Exercise 3.101 (on page 186) implies that f is an isometry. Since V and \tilde{V} are defined as the images of normal *polar* coordinate surface charts, they are not quite neighborhoods of p and \tilde{p} , but it is straightforward to extend f to the images, \mathcal{O} and $\tilde{\mathcal{O}}$, of the corresponding normal coordinate surface charts. \square

It can be seen from the above proof that there are actually many isometries between \mathcal{O} and $\tilde{\mathcal{O}}$, because the orthonormal bases $\{e_1, e_2\} \subset T_p S$ and $\{\tilde{e}_1, \tilde{e}_2\} \subset T_{\tilde{p}} \tilde{S}$ with respect to which σ and $\tilde{\sigma}$ were defined were arbitrary. They were so arbitrary that they were not even mentioned in the proof. If we let $h : T_p S \rightarrow T_{\tilde{p}} \tilde{S}$ be the linear transformation mapping $e_1 \mapsto \tilde{e}_1$ and $e_2 \mapsto \tilde{e}_2$, an alternative description of the isometry, f , constructed in the above proof is $f = (\exp_{\tilde{p}}) \circ h \circ (\exp_p)^{-1}$, as illustrated in this diagram:

$$\begin{array}{ccc} T_p S & \xrightarrow{h=df_p} & T_{\tilde{p}} \tilde{S} \\ \exp_p \downarrow & & \downarrow \exp_{\tilde{p}} \\ \mathcal{O} \subset S & \xrightarrow{f} & \tilde{\mathcal{O}} \subset \tilde{S} \end{array}$$

Notice that $h = df_p$. Thus, an isometry $f : \mathcal{O} \rightarrow \tilde{\mathcal{O}}$ can be constructed such that $df_p : T_p S \rightarrow T_{\tilde{p}} \tilde{S}$ becomes an *arbitrary* orthogonal linear transformation.

Proposition 5.32 implies that a cylinder is locally isometric to a plane, which we already knew from Example 3.63 on page 167. More surprisingly, it implies that fake spheres (Example 4.20 on page 222) are locally isometric to S^2 , which was not previously obvious.

EXERCISES

EXERCISE 5.33. Prove that \mathbb{R}^2 and S^2 are not locally isometric.

COMMENT: According to Proposition 3.75 on page 179, this means that it is impossible for a flat map of any region of the earth to be both equiareal and conformal.

EXERCISE 5.34. With $C_r(p)$ defined as in Definition 5.29, prove that

$$K(p) = \lim_{r \rightarrow 0} \frac{12}{\pi} \frac{\pi r^2 - \text{Area}(C_r(p))}{r^4},$$

where $\text{Area}(C_r(p))$ denotes the enclosed area.

EXERCISE 5.35. For $S = S^2$, find explicit formulas for $\text{length}(C_r(p))$ and $\text{Area}(C_r(p))$, and verify Proposition 5.30 and Exercise 5.34.

EXERCISE 5.36. What is the first nonvanishing term in the Taylor series for $r \mapsto \text{Length}(C_r(p))^2 - 4\pi \text{Area}(C_r(p))$? Compare to the isoperimetric inequality from Sect. 5 of Chap. 2.

EXERCISE 5.37. Classify all isometries of the helicoid. *HINT: Use Exercise 4.40 on page 225 together with the fact that isometries preserve K .*

EXERCISE 5.38. Classify all isometries of the catenoid (Example 4.27 on page 232).

EXERCISE 5.39. Let $\mathcal{F}_1 = dr^2 + G d\theta^2$ be the first fundamental form of a regular surface S in normal polar coordinates at $p \in S$. If G depends only on r (equivalently, if $G_\theta = 0$), prove that every distance circle $C_r(p)$ within the surface patch has constant geodesic curvature. *HINT: Construct an isometry of the image of the surface patch sending any point on the distance circle to any other point.*

EXERCISE 5.40. The converse to the Theorema Egregium is false in the sense that a Gaussian-curvature-preserving diffeomorphism between surfaces need not be an isometry. In particular, use Exercise 4.45 (on page 226) to verify that a rotation by an angle θ about the z -axis is a Gaussian-curvature-preserving diffeomorphism from the monkey saddle to itself for every choice of θ , but is an isometry only for certain choices of θ .

COMMENT: Another such example will be described in Exercise 5.64.

4. Complete Surfaces

All of the definitions and theorems from the past few chapters about a general regular surface S apply equally well to every set $V \subset S$ that is open in S (because such a set V is itself a regular surface, according to Exercise 3.16 on page 136). Such results are categorized as *local* differential geometry. As we shift our focus toward *global* differential geometry, we will often require our surfaces to be *complete*, which we now define.

DEFINITION 5.33.

*Let S be a connected regular surface. A sequence $\{p_1, p_2, \dots\}$ in S is called **d-Cauchy** if for every $\epsilon > 0$, there exists an integer N such that $d(p_i, p_j) < \epsilon$ for all $i, j > N$.*

This is the same as the definition of a **Cauchy sequence** from page 349 of the appendix, but with the distance function “ dist ” in \mathbb{R}^n replaced with the intrinsic distance function, d , of S that was defined in Definition 5.9 (on page 254). Since $d(p_i, p_j) \geq \text{dist}(p_i - p_j)$, every d -Cauchy sequence in S is a Cauchy sequence.

PROPOSITION AND DEFINITION 5.34.

Let S be a connected regular surface. The following are equivalent characterizations of what it means for S to be **complete**:

- (1) Every d -Cauchy sequence of points of S converges to a point of S .
- (2) Every geodesic $\gamma : I \rightarrow S$ can be extended to a geodesic on the domain \mathbb{R} .
- (3) For every $p \in S$, the domain of \exp_p is all of $T_p S$.

PROOF. $(1) \implies (2)$ (by contradiction) Assume that (1) is true and (2) is false. If a unit-speed geodesic γ is defined on an interval of the form $(a, b]$, then it is extendible to $(a, b + \epsilon)$, where ϵ is the radius of any normal neighborhood of $\gamma(b)$. Thus, the only way that (2) could be false is because there is a unit-speed geodesic γ defined on an interval of the form (a, b) but not extendible to b . In this case, let $\{t_1, t_2, \dots\}$ be a sequence in (a, b) converging to b . Since the sequence is convergent, it is Cauchy. Since $d(\gamma(t_i), \gamma(t_j)) \leq |t_i - t_j|$, the sequence $\{\gamma(t_1), \gamma(t_2), \dots\}$ is d -Cauchy, so by hypothesis it converges to a point $p \in S$. Let \mathcal{O} be a normal neighborhood of p . Within \mathcal{O} , notice that γ is a geodesic whose boundary is p , so it must agree with a geodesic through p , and therefore be extendible beyond p , which is a contradiction.

$(2) \iff (3)$ Obvious.

$(3) \implies (1)$ Let $\{p_1, p_2, \dots\}$ be a d -Cauchy sequence in S . There exists an integer N such that for all $i > N$, $d(p_N, p_i) \leq 1$. In other words, the entire tail of this sequence is contained in the ball

$$B = \{q \in S \mid d(q, p_N) \leq 1\}.$$

Using hypothesis (3), B is compact because it is the image under the continuous function \exp_{p_N} of the compact set $\{v \in T_{p_N} S \mid |v| \leq 1\}$. Every sequence in a compact set subconverges to a point of that set. A subconvergent Cauchy sequence is convergent. Thus, this sequence converges to a point of B . \square

The punctured plane (Example 5.5 on page 249) is not complete. A sufficiently small normal neighborhood of any point of any surface is not complete. The issue is that these examples are missing limit points, which might lead one to guess that “complete” is related to “closed.” In fact, we have the following:

PROPOSITION 5.35.

Every closed connected regular surface S is complete. In particular, every compact connected regular surface is complete.

PROOF. Let $\{p_1, p_2, \dots\}$ be a d -Cauchy sequence in S . As previously mentioned, this sequence is Cauchy, so it converges to a point $p \in \mathbb{R}^3$. Since S is closed, we have $p \in S$. \square

Figure 5.13 shows (part of) a connected surface that is complete but not closed, namely, a generalized cylinder whose generating curve is a spiral asymptotic to a circle. It can be explicitly described as

$$S = \{((1 + e^{-s}) \cos s, (1 + e^{-s}) \sin s, t) \mid s, t \in \mathbb{R} \}.$$

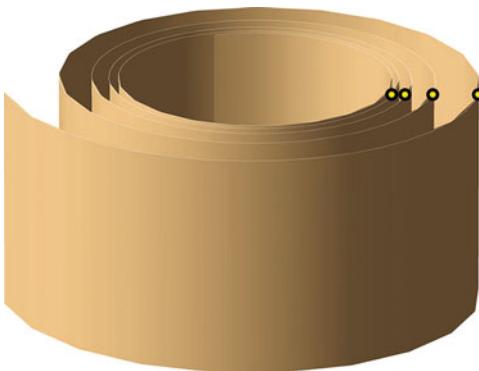


FIGURE 5.13. Complete but not closed

It is possible to construct a sequence of points of this surface (beginning like the illustrated yellow points) that is Cauchy but not d -Cauchy. Such a sequence verifies that the surface is not closed (it converges to a missing limit point), but it does not contradict completeness. In spite of such exotic examples, “complete” is *roughly* the same thing as “closed.” In fact, the “ d -Cauchy” definition of

“complete” is the most natural way to formulate the idea of “containing all limit points” intrinsically—as detectable by inhabitants of the domain of a surface patch who do not know about the ambient \mathbb{R}^3 or its distance function, but rather know only the first fundamental form.

The most important fact about complete surfaces is contained in the following theorem:

THEOREM 5.36 (The Hopf–Rinow Theorem).

If S is a connected complete regular surface, then there exists a minimizing geodesic segment between any given pair of points of S .

PROOF. Let $p, q \in S$ and define $r = d(p, q)$. Let $C = C_\epsilon(p)$ be a distance circle about p (as in Definition 5.29 on page 271) whose radius, ϵ , is smaller than r and is small enough to ensure that C is contained in a normal neighborhood of p . Notice that C is compact, because C is the image under the continuous function \exp_p of a circle in $T_p S$. The “distance to q ” function is continuous on S (Exercise 5.15 on page 257), so its restriction to C attains a minimum at some point x ; that is, x is a point of C closest to q . Let $v \in T_p S$ be the unit-length vector tangent to the unique minimizing geodesic from p to x . Thus, the unit-speed geodesic $\gamma : [0, \infty) \rightarrow S$ defined as $\gamma(t) = \exp_p(tv)$

satisfies $\gamma(0) = p$ and $\gamma(\epsilon) = x$. Our goal is to prove that $\gamma(r) = q$, and thus that the restriction of γ to $[0, r]$ is a minimizing geodesic from p to q .

For this, the idea is roughly to prove that someone traversing γ makes steady progress toward q ; every step along γ is a step closer to q . We make this notion precise by considering the set of times during which steady progress has been made:

$$U = \{t \in [\epsilon, r] \mid d(\gamma(t), q) = r - t\}.$$

It will suffice to show that $U = [\epsilon, r]$, since this implies that $\gamma(r) = q$, as desired. We will achieve this by showing that $\epsilon \in U$ (so U is nonempty) and that U is both open in $[\epsilon, r]$ and closed in $[\epsilon, r]$.

The assertion that U is closed in $[\epsilon, r]$ is straightforward and left to the reader. The assertion that $\epsilon \in U$ follows from the observation that every curve in S between p and q must intersect C , so that

$$r = d(p, q) = \inf\{d(p, y) + d(y, q) \mid y \in C\} = \inf\{\epsilon + d(y, q) \mid y \in C\} = \epsilon + d(x, q).$$

It remains to prove that U is open in $[\epsilon, r]$. For this, suppose that $t_0 \in U$ with $t_0 < r$. Let $\tilde{C} = C_\delta(\gamma(t_0))$ be a distance circle about $\gamma(t_0)$ whose radius δ is smaller than $r - t_0$ and is small enough to ensure that \tilde{C} is contained in a normal neighborhood of $\gamma(t_0)$. As before, compactness implies that some point y of \tilde{C} is closest to q , and we claim that $y = \gamma(t_0 + \delta)$. To justify this claim, suppose that $y \neq \gamma(t_0 + \delta)$, as shown in Fig. 5.14. The argument we previously used to prove that $\epsilon \in U$ applies in the current situation to show that $d(y, q) = r - (t_0 + \delta)$. The triangle inequality gives

$$d(p, y) \geq d(p, q) - d(y, q) = r - (r - (t_0 + \delta)) = t_0 + \delta.$$

The distance between p and y is therefore realized by a nonsmooth curve from p to y , namely the curve that follows γ from p to $\gamma(t_0)$ and then follows the minimizing geodesic from $\gamma(t_0)$ to y . This contradicts the regularity of minimizing curves (Exercise 5.27 on page 267).

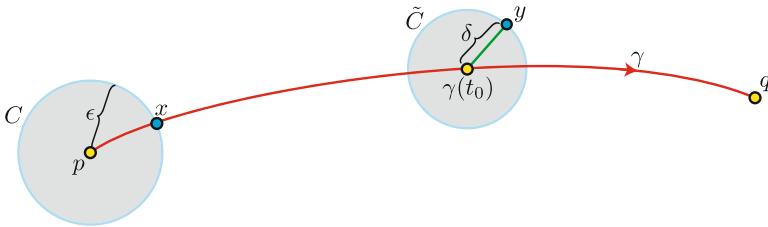


FIGURE 5.14. The proof of the Hopf–Rinow theorem

Thus, $y = \gamma(t_0 + \delta)$, and its distance to q is $r - (t_0 + \delta)$. It follows that $t_0 + \delta \in U$. Every number between ϵ and an element of U is an element of U , so this concludes our proof that U is open in $[\epsilon, r]$. \square

EXERCISES

EXERCISE 5.41. Let S be a connected regular surface. The surface S is called **nonextendible** if it is not a proper subset of a larger connected regular surface.

- (1) Prove or disprove: If S is nonextendible, then S is complete.
- (2) Prove or disprove: If S is complete, then S is nonextendible.

HINT: See Example 3.20 on page 126.

EXERCISE 5.42. Let S be a connected complete regular surface. The **diameter** of S is defined as

$$\text{diam}(S) = \sup\{d(p, q) \mid p, q \in S\}.$$

If the diameter of S is finite, prove that S is compact.

EXERCISE 5.43. Let S be a connected regular surface. Prove that S is complete if and only if every closed d -bounded subset $A \subset S$ is compact, where “ d -bounded” means that $\sup\{d(p, q) \mid p, q \in A\}$ is finite.

EXERCISE 5.44. Let S be a noncompact complete connected regular surface. Prove that for every $p \in S$, there exists a minimizing geodesic $\gamma : [0, \infty) \rightarrow S$ with $\gamma(0) = p$. Such a geodesic is called a **ray**.

HINT: Consider the limit of a sequence of unit-length vectors in $T_p S$ tangent to minimizing geodesics from p to points farther and farther away from p .

EXERCISE 5.45 (An isometry is determined by its derivative at a point).

- (1) Let S be a connected complete regular surface and $p \in S$. Let $f : S \rightarrow S$ be an isometry with $f(p) = p$ and $df_p = \text{identity}$. Prove that f is the identity map.
- (2) Let $f, g : S \rightarrow \tilde{S}$ be a pair of isometries between connected complete regular surfaces. If there exists $p \in S$ such that $f(p) = g(p)$ and $df_p = dg_p$, prove that $f = g$.

EXERCISE 5.46. Let S be a connected complete regular surface. Let $f : S \rightarrow S$ be an isometry of S that is not the identity. Let γ be a unit-speed curve in S . Suppose that the trace, C , of γ is fixed by f ; that is, $f(p) = p$ for all $p \in C$. Prove that γ is a geodesic.

HINT: Let $p = \gamma(t_0)$. For t near t_0 , let $v_t \in T_p S$ be the unit tangent vector to the minimizing geodesic from p to $\gamma(t)$. All such v_t must be mutually parallel to avoid contradicting Exercise 5.45.

EXERCISE 5.47. Verify that the pseudosphere (Exercise 4.44 on page 225) is not complete. Using your solution to Exercise 4.35 (on page 224), prove that there does *not* exist a complete surface of revolution with constant Gaussian curvature $K = -1$.

COMMENT: Hilbert proved the much stronger result that there does not exist a complete regular surface with constant Gaussian curvature $K = -1$.



5. Parallel Transport and the Covariant Derivative

The plane \mathbb{R}^2 is the prototype regular surface, but there are still measurements and constructions for the plane that we don't yet know how to generalize to arbitrary regular surfaces. For example, we can differentiate every vector field along every regular curve in \mathbb{R}^2 . Here is how we will generalize this idea to arbitrary surfaces:

DEFINITION 5.37.

Let S be a regular surface and let $\gamma : I \rightarrow S$ be a regular curve in S . Let v be a **vector field along γ** , which means a smooth function $v : I \rightarrow \mathbb{R}^3$ such that $v(t) \in T_{\gamma(t)}S$ for all $t \in I$.

- (1) The **covariant derivative** of v (denoted by $\frac{Dv}{dt}$, $\frac{D}{dt}v$, or v^θ) is the vector field along γ defined such that for all $t_0 \in I$,

$$v^\theta(t_0) = \left. \frac{Dv}{dt} \right|_{t=t_0} = \text{the projection of } v'(t_0) \text{ onto } T_{\gamma(t_0)}S.$$

- (2) v is called **parallel** if $v^\theta(t_0) = \mathbf{0}$ for all $t_0 \in I$.

The term “projection” was precisely defined in Sect. 2 of Chap. 1. Notice that the usual derivative of v (Definition 1.6 on page 3) is denoted by $\frac{dv}{dt}$ or v' , while the covariant derivative of v is denoted by $\frac{Dv}{dt}$ or v^θ . The latter option is nonstandard, but we find it a convenient way to avoid the clunkiness of “ $\frac{Dv}{dt}|_{t=t_0}$ ” when the input needs to be specified.

EXAMPLE 5.38 ($S = \mathbb{R}^2$). A vector field along $\gamma : I \rightarrow \mathbb{R}^2$ just means an arbitrary smooth function $v : I \rightarrow \mathbb{R}^2$. In this situation, $\frac{Dv}{dt} = \frac{dv}{dt}$; that is, the covariant derivative is the usual derivative. Thus, v is parallel if and only if it is constant.

This example indicates that our definition of a *parallel* vector field along a curve in an arbitrary surface generalizes the notion of a *constant* vector field along a curve in \mathbb{R}^2 . To nearsighted inhabitants of a general surface who believe that their world is flat, a parallel vector field will appear constant, because it turns in only the undetectable normal direction.

LEMMA 5.39 (Algebraic properties).

Let S be a regular surface, and $\gamma : I \rightarrow S$ a regular curve in S . Let v and w be vector fields along γ and let $a, b : I \rightarrow \mathbb{R}$ be smooth functions.

- (1) $(v + w)^\theta = v^\theta + w^\theta$.

In particular, if v and w are parallel, then $v + w$ is parallel.

- (2) $(av)^\theta = a'v + av^\theta$.

In particular, if v is parallel, then av is parallel if and only if a is constant.

- (3) $\frac{d}{dt} \langle v, w \rangle = \langle v^\theta, w \rangle + \langle v, w^\theta \rangle$.

In particular, a parallel field has constant norm, while two parallel fields maintain a constant angle.

- (4) (**Reparametrization**) If $\phi : \tilde{I} \rightarrow I$ is smooth, then $\tilde{v} = v \circ \phi$ is a vector field along $\tilde{\gamma} = \gamma \circ \phi$, and $\tilde{v}^\theta(t) = \phi'(t)v^\theta(\phi(t))$ for all $t \in \tilde{I}$. In particular, \tilde{v} is parallel along $\tilde{\gamma}$ if and only if v is parallel along γ .

PROOF. All claims follow immediately from the corresponding properties of the usual derivative. The blue assertions come from considering the special case in which the vector fields are parallel. \square

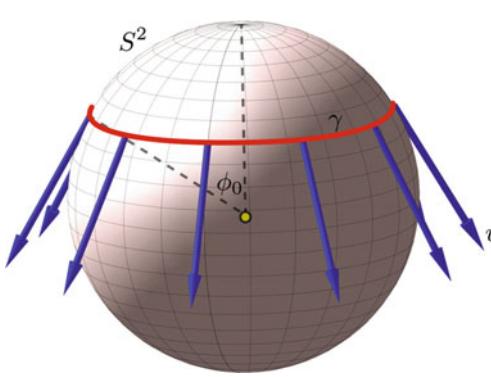


FIGURE 5.15. Not a parallel vector field

multiple of γ' . A calculation shows that $v^\theta = (-\cot \phi_0)\gamma'$. In particular, v is parallel only if $\phi_0 = \frac{\pi}{2}$. For other values of ϕ_0 , we will determine the parallel vector fields along γ after we first prove generally that they are guaranteed to exist.

EXAMPLE 5.40. Let σ be the spherical-coordinates surface patch for S^2 given in Example 3.24 on page 129. For fixed $\phi_0 \in (0, \pi)$, let γ be a unit-speed parametrization of the ϕ_0 -latitudinal curve; that is $\gamma(t) = \sigma(\frac{t}{\delta}, \phi_0)$, $t \in [0, 2\pi\delta]$, where $\delta = \sin \phi_0$ is chosen to make γ of unit speed. Let v be the restriction of σ_ϕ to γ , which means $v(t) = \sigma_\phi(\frac{t}{\delta}, \phi_0)$; see Fig. 5.15. Since v is of unit length, Lemma 5.39(3) implies $v^\theta \perp v$, so v^θ must be a multiple of γ' .

THEOREM 5.41 (Existence and uniqueness of parallel vector fields).

Let S be a regular surface and $\gamma : I \rightarrow S$ a regular curve. Let $t_0 \in I$. For every $w_0 \in T_{\gamma(t_0)}S$, there exists a unique parallel vector field, w , along γ satisfying the initial condition $w(t_0) = w_0$.

PROOF. Due to Lemma 5.39, we can assume without loss of generality that γ is of unit speed and that w_0 is of unit length. For simplicity, we also assume for now that S is oriented, and we let κ_g denote the geodesic curvature function of γ . Let $v = \gamma'$, which is a unit-length vector field along γ . Notice that $n = R_{90} \circ v$ is another unit-length vector field along γ , where R_{90} denotes the 90° rotation that is counterclockwise with respect to the orientation.

We will first establish the following Frenet-like equations:

$$(5.8) \quad \boxed{v^\theta = \kappa_g n \quad \text{and} \quad n^\theta = -\kappa_g v}.$$

The left equation is immediate from our original definition of κ_g given in Eq. 4.8 on page 209. To justify the right equation, since n is of unit length, its covariant derivative is orthogonal to itself, and is therefore a multiple of v , and this multiple is

$$\langle n^\theta, v \rangle = \langle n, v' \rangle - \langle n, v^\theta \rangle = 0 - \kappa_g.$$

We now wish to construct a parallel vector field, w , along γ with $w(t_0) = w_0$. Such a vector field must be of unit length, so the most natural strategy is to find a smooth function $\theta : I \rightarrow \mathbb{R}$ such that the following works:

$$w(t) = (\cos \theta(t))v(t) + (\sin \theta(t))n(t).$$

Equation 5.8 together with Lemma 5.39 gives

$$w^\theta(t) = -(\sin \theta(t))(\theta'(t) + \kappa_g(t))v(t) + (\cos \theta(t))(\theta'(t) + \kappa_g(t))n(t).$$

Therefore, w is parallel if and only if $\theta' = -\kappa_g$. A solution is

$$\theta(t) = \theta_0 - \int_{t_0}^t \kappa_g(s) \, ds,$$

where $\theta_0 \in \mathbb{R}$ is chosen such that $w(t_0) = w_0$.

The uniqueness of this solution can be seen from the above construction, but we will instead prove uniqueness as follows. Let w be the parallel field along γ constructed above with $w(t_0) = w_0$. Notice that $R_{90} \circ w$ is also a parallel field along γ (obtained by adding $\frac{\pi}{2}$ to the θ function). Any other vector field along γ will have the form $aw + b(R_{90} \circ w)$ for some smooth functions $a, b : I \rightarrow \mathbb{R}$. Lemma 5.39 shows that this is parallel if and only if a and b are constant functions. The solution is therefore unique.

The case in which S is not orientable is handled by applying the above argument to portions of γ covered by a single surface patch (with the induced local orientation). The details are left to the reader in Exercise 5.55. \square

In the above proof, the function θ described a parallel field in terms of γ' . But we will define the *angle function* of γ to be the negative of this function, which does the opposite: it describes γ' in terms of a parallel field, which is often more convenient. After implementing this minor change, the key observation of the proof can be summarized as follows:

PROPOSITION AND DEFINITION 5.42.

Let S be an oriented regular surface, $\gamma : I \rightarrow S$ a unit-speed curve, and $\kappa_g : I \rightarrow \mathbb{R}$ its geodesic curvature function. Let $\theta : I \rightarrow \mathbb{R}$ be an antiderivative of κ_g . Then θ is called an **angle function** of γ . It is unique up to an additive constant, and has this property: for every unit-length parallel vector field w along γ , there exists $\theta_0 \in [0, 2\pi)$ such that

$$\gamma'(t) = \cos(\theta(t) + \theta_0)w(t) + \sin(\theta(t) + \theta_0)R_{90}(w(t))$$

for all $t \in I$.

The fact that $\kappa_g = \theta'$ means this: the geodesic curvature of γ measures the rate of turning of its velocity vector, reckoned with respect to a parallel vector field along γ .

EXAMPLE 5.43. If γ is a geodesic in Proposition 5.42, then $\kappa_g = 0$, so θ is a constant function. Thus, a unit-length vector field along γ is parallel if and only if it maintains a constant angle with γ' . In particular, γ' is parallel, which is also obvious directly from the definition of a geodesic.

EXAMPLE 5.44 ($S = \mathbb{R}^2$). Along every curve γ in \mathbb{R}^2 , a vector field is parallel if and only if it is constant. If the additive constant in Proposition 5.42 is chosen to make $\theta_0 = 0$ correspond to the parallel field $w(t) = (1, 0)$, then θ is exactly the angle function that was defined in Sect. 6 of Chap. 1. The geodesic curvature of γ (with respect to the orientation $N = (0, 0, 1)$) equals the signed curvature of γ , so Proposition 5.42 generalizes the main result of Sect. 6 of Chap. 1.

DEFINITION 5.45.

Let S be a regular surface and $\gamma : [a, b] \rightarrow S$ a regular curve in S .

- (1) **Parallel transport** along γ is the function $P_\gamma : T_{\gamma(a)}S \rightarrow T_{\gamma(b)}S$ defined to send each $w_0 \in T_{\gamma(a)}S$ to the value at $t = b$ of the parallel vector field along γ whose value at $t = a$ equals w_0 .
- (2) If γ is closed (so that $\gamma(a) = \gamma(b) = p$), then $P_\gamma : T_pS \rightarrow T_pS$ is also called the **holonomy** around γ .
- (3) If S is oriented and γ is of unit speed, then the **angle displacement** along γ is $\Delta\theta = \theta(b) - \theta(a)$, where θ is an angle function of γ .

By Lemma 5.39, parallel transport along γ is an orthogonal linear transformation from $T_{\gamma(a)}S$ to $T_{\gamma(b)}S$ that would be unchanged by an orientation-preserving reparametrization of γ . When S is oriented and γ is of unit speed, P_γ equals a clockwise rotation by $\Delta\theta$, but this makes sense only after one

identifies both tangent planes with \mathbb{R}^2 via the basis $\{\gamma', R_{90}(\gamma')\}$. If additionally γ is closed, then the following makes sense without even specifying any identification:

PROPOSITION 5.46.

If S is an oriented regular surface and $\gamma : [a, b] \rightarrow S$ is a unit-speed closed curve in S , then the holonomy around γ is equal to a clockwise rotation of $T_{\gamma(a)}S$ by $\Delta\theta$ (the angle displacement along γ).

PROOF. Since the basis $\{\gamma'(t), R_{90}(\gamma'(t))\}$ is the same at $t = a$ and $t = b$, the claim follows from Proposition 5.42. \square

But notice that $\Delta\theta$ might be more than 2π , so it contains more information than does P_γ . For example, when $S = \mathbb{R}^2$ and γ is a closed loop, P_γ is the identity map, while $\Delta\theta$ equals 2π times the rotation index of γ . In this example, P_γ contains no information, while $\Delta\theta$ encodes the rotation index.

The next example shows that the angle displacement around a closed loop in a general oriented surface might not be a multiple of 2π , so the notion of *rotation index* does not generalize to this context.

EXAMPLE 5.47 (Parallel transport around a latitude of the sphere). Let γ be the unit-speed parametrization of the ϕ_0 -latitudinal curve in S^2 , as in Example 5.40. According to Exercise 4.12 on page 211, the geodesic curvature of γ is constant at $\kappa_g = \cot \phi_0$ (with respect to the outward-pointing orientation of S^2). The length of γ is $l = 2\pi \sin \phi_0$. Therefore, the angle displacement along γ equals

$$(5.9) \quad \Delta\theta = \int_0^l \theta'(s) ds = \int_0^l \kappa_g(s) ds = (2\pi \sin \phi_0) \cot \phi_0 = 2\pi \cos \phi_0.$$

Parallel vector fields along three latitudinal curves are illustrated in Fig. 5.16. The latitude $\phi_0 = 90^\circ$ is a geodesic (the equator), so it has angle displacement $\Delta\theta = 0$. At the other extreme, if $\phi_0 \approx 0^\circ$, then $\Delta\theta \approx 360^\circ$. A sufficiently small neighborhood of the north pole looks like the plane \mathbb{R}^2 , so parallel fields along small loops circling the north pole are approximately constant, and Hopf's Umlaufsatz is approximately valid. The other latitudes between $\phi_0 \approx 0^\circ$ and $\phi_0 = 90^\circ$ interpolate between these two extremes.

In Fig. 5.16, the middle latitude corresponds to $\phi_0 = 41^\circ$ and has angle displacement $\Delta\theta = 272^\circ$. This particular latitude of the Earth passes through Paris, where a famous experiment was performed in 1851:

EXAMPLE 5.48 (Foucault's pendulum). Jean Foucault in 1851 famously demonstrated the rotation of the Earth using a pendulum made from a heavy iron ball attached by a long wire to the dome of the Pantheon in Paris; see

Fig. 5.17. The Coriolis force caused by the Earth's rotation makes the swing-plane of Foucault's pendulum rotate as time passes at a speed depending on the latitude where the pendulum sits. In Paris, this rotation is about 11° per hour, or 272° per day, the same number as above because the changing swing-plane as the Earth rotates equals the parallel transport of the initial swing plane along the latitude of the globe.

Physical descriptions of the Coriolis force are easily found online, together with animations of this phenomenon. We will content ourselves here with only an intuitive discussion of why one might expect the pendulum's swing-plane to follow a parallel vector field. If ϕ_0 is very small, so the pendulum is near the North Pole, then continuity should force its behavior to be close to the behavior of a pendulum positioned exactly at the North Pole. Figure 5.17 (right) shows a North-Pole pendulum. It is suspended from a wooden frame that rests on the Earth, but this is irrelevant; since it swivels freely at the attachment point, it might as well be suspended from the stars. Its

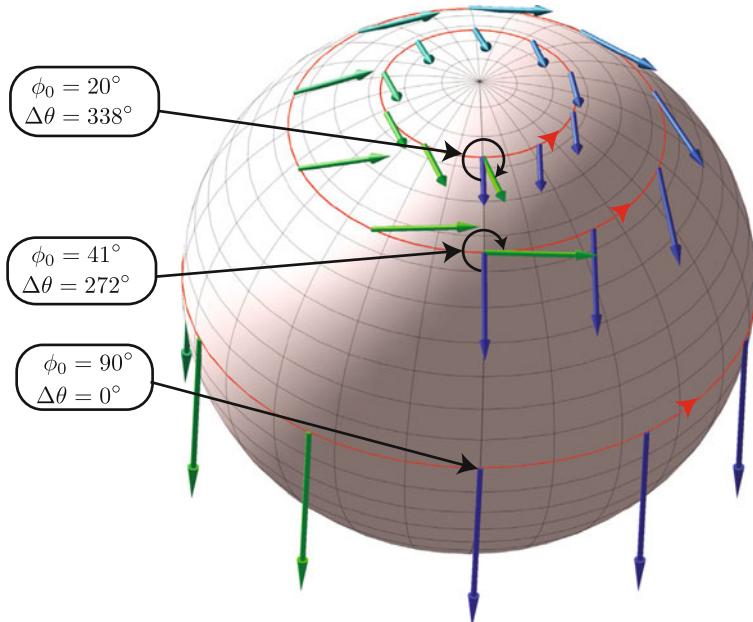


FIGURE 5.16. Parallel vector fields along three latitudinal curves

swing-plane remains constant relative to the stars, but the Earth will rotate beneath it, causing an observer on Earth (who is rotating with the Earth) to measure that its swing-plane rotates clockwise by 360° per day. Similarly, a pendulum at the South Pole would be observed to rotate counterclockwise at 360° per day. A pendulum was actually built at the South Pole in 2001 to verify this. Along the equator, a pendulum that begins swinging north-south will maintain a north-south swing-plane. All of this behavior is consistent

with Eq. 5.9 when ϕ_0 is close to 0° or equal to 90° or close to 180° . So it is perhaps unsurprising that this equation correctly describes pendulums at intermediate latitudes as well.

Re-creations of Foucault's pendulum allow visitors at science museums every day to “see” a parallel vector field along a latitude of the spinning Earth. Even better, there is a mechanical device that allows us to see a parallel vector field along any curve on any surface, namely the south-pointing chariot that we previously discussed in Sect. 6 of Chap. 1. Recall that its gearing causes the statue's clockwise angular speed (relative to the chariot) to be proportional the right wheel's speed minus the left wheel's speed. We will prove in Sect. 8 that if the chariot is driven along a curve in an arbitrary oriented surface, and if the chariot is small compared to the features of the surface, then the statue will point approximately in the direction of a parallel vector field along the curve. In other words, the statue's total clockwise rotation relative to the chariot will be an angle function for the curve. For example, if the chariot is driven once eastward along the latitude of the Earth through Paris, its right wheel will move faster than its left, causing an observer driving the chariot to see the statue turn approximately 272° clockwise during the journey, the same number that appears in Fig. 5.16. We will justify this claim in Sect. 8; for now, just notice that the chariot's behavior around northern latitudes must somehow interpolate between the extreme cases in which ϕ_0 is small (the chariot traverses a small loop about the North Pole, so the statue will rotate approximately 360° clockwise by Hopf's Umlaufsatz) and where $\phi_0 = \pi/2$ (the chariot traverses the equator, so its wheels move at equal speeds, and the statue doesn't rotate).

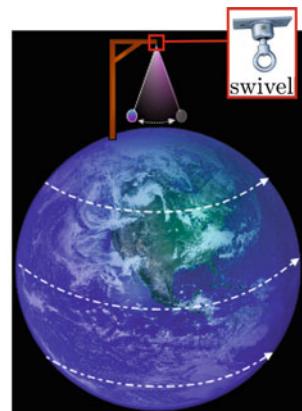


FIGURE 5.17. *Left:* a re-creation of Foucault's pendulum tips over pegs as the swing-plane rotates throughout the day. *Right:* the Earth would rotate underneath a pendulum at the North Pole

We end this section by proving that its key structures are preserved by isometries:

PROPOSITION 5.49 (Covariant differentiation is intrinsic).

Let $f : S \rightarrow \tilde{S}$ be an isometry between regular surfaces. Let $\gamma : I \rightarrow S$ be a regular curve in S and let $\tilde{\gamma} = f \circ \gamma$. Let v be a vector field along γ in S , and let \tilde{v} be the corresponding vector field along $\tilde{\gamma}$ in \tilde{S} , which is defined as $\tilde{v}(t) = df_{\gamma(t)}(v(t))$ for all $t \in I$.

- (1) For all $t \in I$, $\tilde{v}^\theta(t) = df_{\gamma(t)}(v^\theta(t))$.
- (2) v is parallel along γ if and only if \tilde{v} is parallel along $\tilde{\gamma}$.

PROOF. We will assume that γ is of unit speed, S is oriented, and f is orientation-preserving. We leave it to the reader to explain how the general case follows from this special case. Although (2) follows immediately from (1), we will prove these claims in the opposite order.

For part (2), Corollary 5.25 (on page 266) says that γ and $\tilde{\gamma}$ have the same geodesic curvature function κ_g . By Proposition 5.42, they share a common angle function θ . Since df_p preserves orientation and angles for all $p \in S$, the result now follows from the manner in which the global angle function determines the parallel vector fields.

For part (1), let w_1 be a unit-length parallel vector field along γ . As previously mentioned, $w_2 = R_{90} \circ w_1$ is another unit-length parallel vector field along γ that is everywhere orthogonal to w_1 . The arbitrary vector field v can be expressed as $v = aw_1 + bw_2$ for some smooth functions $a, b : I \rightarrow \mathbb{R}$. Lemma 5.39 gives $v^\theta = a'w_1 + b'w_2$. By part (1), the vector fields along $\tilde{\gamma}$ defined as $\tilde{w}_1(t) = df_{\gamma(t)}(w_1(t))$ and $\tilde{w}_2(t) = df_{\gamma(t)}(w_2(t))$ are parallel, so

$$\tilde{v}^\theta = (a\tilde{w}_1 + b\tilde{w}_2)^\theta = a'\tilde{w}_1 + b'\tilde{w}_2 = df(v^\theta).$$

□

EXERCISES

EXERCISE 5.48. Let S be the torus of revolution illustrated and described in Exercise 4.23 on page 212. Calculate the angle displacement (with respect to the outward-pointing orientation) around the two circles of zero Gaussian curvature (the boundaries between red and green). Calculate the angle displacement around an arbitrary latitudinal curve between these two circles.

EXERCISE 5.49. Suppose that $\gamma : I \rightarrow \mathbb{R}^3$ is a space curve whose trace lies in two regular surfaces, S_1 and S_2 . Suppose further that the surfaces share the same tangent planes along γ ; that is, $T_{\gamma(t)}S_1 = T_{\gamma(t)}S_2$ for all $t \in I$. Explain why covariant derivatives and parallel transports along γ are the same with respect to S_1 and S_2 .

EXERCISE 5.50. For the cone in Fig. 5.18, compute (in terms of α) the angle displacement around a latitudinal curve γ (shown in red). Solve this problem in three ways:

- (1) By computing κ_g .
- (2) By applying Proposition 5.49 to an isometry between (all but a line of) C and a sector, U , in \mathbb{R}^2 .
- (3) By applying Exercise 5.49 to a unit-radius sphere resting in C like a scoop of ice cream in an ice cream cone.

HINT: In the figure, how are ϕ and $\tilde{\alpha}$ determined from α ?

EXERCISE 5.51. Prove that parallel transport around the central loop of the Möbius strip (shown red in Fig. 3.25 on page 157) is an improper rigid motion of the tangent plane.

EXERCISE 5.52. Let γ be the ϕ_0 -latitudinal curve of S^2 (as in Example 5.47) and let R be the spherical cap enclosed by γ (as in Exercise 3.75 on page 165). Verify that the angle displacement along γ is

$$\Delta\theta = 2\pi - \text{Area}(R).$$

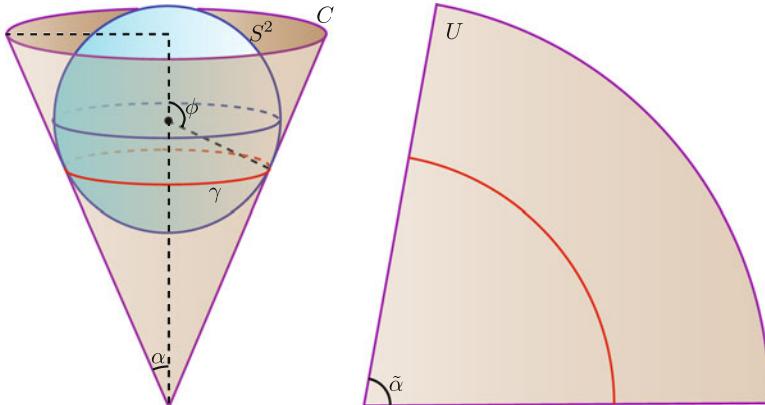


FIGURE 5.18. C , S^2 , and U have related angle displacements

EXERCISE 5.53 (Covariant differentiation is determined by parallel transport). Let γ be a regular curve in a regular surface S . Let $P_t : T_{\gamma(0)}S \rightarrow T_{\gamma(t)}S$ denote the parallel transport along the restriction of γ to $[0, t]$. Prove that for every vector field w along γ ,

$$w^\theta(0) = \frac{d}{dt} \Big|_{t=0} (P_t)^{-1}(w(t)).$$

EXERCISE 5.54 (Discrete parallel transport). Let $\gamma : [a, b] \rightarrow S$ be a unit-speed curve in a regular surface. Let $v \in T_{\gamma(a)}S$. Let $a = t_0 < t_1 < \dots < t_n = b$ be a regular partition of $[a, b]$, which means that $t_i = a + i\Delta t$, where $\Delta t = \frac{b-a}{n}$. Set $v_0 = v$, then define v_1 as the projection of v_0 onto

$T_{\gamma(t_1)}S$, then define v_2 as the projection of v_1 onto $T_{\gamma(t_2)}S$, and so on. Prove that $\lim_{n \rightarrow \infty} v_n$ equals the parallel transport of v along γ .

EXERCISE 5.55. In the proof of Theorem 5.41, how do you handle the case in which S is not oriented?

□

6. Geodesics in Local Coordinates

In this section, we introduce local-coordinate formulas for computing geodesics and covariant derivatives. Among other things, we will use these formulas to prove Proposition 5.3 on page 248 (the existence and uniqueness of geodesics) and Theorem 5.26 on page 269 (the formula for K in normal polar coordinates), as promised.

Throughout this section, let S be an oriented regular surface with orientation N , and let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch compatible with N , whose first and second fundamental forms are denoted respectively by

$$\mathcal{F}_1 = E du^2 + 2F du dv + G dv^2 \quad \text{and} \quad \mathcal{F}_2 = e du^2 + 2f du dv + g dv^2.$$

Recall from Eq. 4.12 (on page 218) that $e = \langle \sigma_{uu}, N \rangle$, $f = \langle \sigma_{uv}, N \rangle$, and $g = \langle \sigma_{vv}, N \rangle$, but what about the components of σ_{uu} , σ_{uv} , and σ_{vv} tangent to S ? To express σ_{uu} , σ_{uv} , and σ_{vv} at an arbitrary point of U as a linear combination of the basis $\{\sigma_u, \sigma_v, N\}$ will require several additional functions from U to \mathbb{R} (in addition to e , f , and g). These extra functions are the Christoffel symbols:

DEFINITION 5.50.

The **Christoffel symbols**, $\Gamma_{ij}^k : U \rightarrow \mathbb{R}$, for $i, j, k \in \{1, 2\}$, are defined such that the following hold on all of U :

$$\begin{aligned}\sigma_{uu} &= \Gamma_{11}^1 \sigma_u + \Gamma_{11}^2 \sigma_v + eN = \frac{D}{du} \sigma_u + eN, \\ \sigma_{uv} &= \Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v + fN = \frac{D}{du} \sigma_v + fN, \\ \sigma_{vu} &= \Gamma_{21}^1 \sigma_u + \Gamma_{21}^2 \sigma_v + fN = \frac{D}{dv} \sigma_u + fN, \\ \sigma_{vv} &= \Gamma_{22}^1 \sigma_u + \Gamma_{22}^2 \sigma_v + gN = \frac{D}{dv} \sigma_v + gN.\end{aligned}$$

To interpret this definition, think of “1 = u ” and “2 = v .“ Although there are eight Christoffel symbols, the fact that $\sigma_{uv} = \sigma_{vu}$ implies that $\Gamma_{21}^1 = \Gamma_{12}^1$ and $\Gamma_{21}^2 = \Gamma_{12}^2$, so really there are only six different Christoffel symbols.

The “ $\frac{D}{du}$ ” and “ $\frac{D}{dv}$ ” expressions in Definition 5.50 represent the covariant derivatives of σ_u and σ_v along the u - and v -parameter curves. For example, $\frac{D}{du} \sigma_v : U \rightarrow \mathbb{R}^3$ is defined such that for all $q \in U$, $\frac{D}{du} \sigma_v(q)$ equals the covariant

derivative at q of the restriction of σ_v to the u -parameter curve through q , and the other expressions are defined analogously. In short, the Christoffel symbols encode the covariant derivatives of the coordinate vector fields in the coordinate directions. So it's not surprising that covariant derivatives of arbitrary vector fields along arbitrary curves can be expressed in terms of them:

LEMMA 5.51 (Covariant differentiation in local coordinates).

Let $\gamma : I \rightarrow V$ be an arbitrary regular curve in V , which can be expressed as $\gamma(t) = \sigma(u(t), v(t))$ for some smooth functions $u, v : I \rightarrow \mathbb{R}$. Let w be an arbitrary vector field along γ , which can be expressed as

$$w(t) = a(t)\sigma_u(u(t), v(t)) + b(t)\sigma_v(u(t), v(t))$$

for some smooth functions $a, b : I \rightarrow \mathbb{R}$. Then

$$\begin{aligned} w' &= (a' + au'\Gamma_{11}^1 + (av' + bu')\Gamma_{12}^1 + bv'\Gamma_{22}^1) \sigma_u \\ &\quad + (b' + au'\Gamma_{11}^2 + (av' + bu')\Gamma_{12}^2 + bv'\Gamma_{22}^2) \sigma_v \end{aligned}$$

on all of I . In particular, w is parallel if and only if the coefficients of σ_u and σ_v in this expression are both zero.

PROOF. This follows from Definition 5.50 after the chain rule has been used to write

$$w' = a'\sigma_u + a(u'\sigma_{uu} + v'\sigma_{uv}) + b'\sigma_v + b(u'\sigma_{vu} + v'\sigma_{vv}).$$

□

The special case of this lemma with $w = \gamma'$ is particularly important:

LEMMA 5.52 (The geodesic equations).

The regular curve $\gamma(t) = \sigma(u(t), v(t))$ in V is a geodesic if and only if

$$\begin{aligned} u'' + (u')^2\Gamma_{11}^1 + 2(u'v')\Gamma_{12}^1 + (v')^2\Gamma_{22}^1 &= 0 \\ \text{and } v'' + (u')^2\Gamma_{11}^2 + 2(u'v')\Gamma_{12}^2 + (v')^2\Gamma_{22}^2 &= 0. \end{aligned}$$

PROOF. Since γ is a geodesic if and only if γ' is parallel, this is the special case of Lemma 5.51 in which $w = \gamma'$. □

Readers familiar with the theory of differential equations will observe that the system in Lemma 5.52 can be solved uniquely for $\{u, v\}$ (given any

prescriptions for u, u', v, v' at a single time $t_0 \in I$), which provides a proof of Theorem 5.3 on page 248 (the existence and uniqueness of geodesics).

Lemma 5.51 has another important consequence. Only the *first* derivatives of u and v appear in this Lemma's expression for w^θ . In other words, the dependence of $w^\theta(t)$ on the curve γ involves only $\gamma'(t)$. This means, for example, that we could have been more careless in defining the expression “ $\frac{D}{du}\sigma_v(q)$ ” above. We could have simply stated that it equals the value at q of the restriction of σ_v to *any* regular curve passing through q with initial velocity σ_u . More generally, if w is a tangent field on *all* of S , then the covariant derivative of its restriction to a regular curve in S can be computed at a point in terms only of the velocity vector of the curve at that point (and of w in a neighborhood of this point).

Gauss discovered that the Christoffel symbols can be expressed purely in terms of the coefficients of the first fundamental form and their partial derivatives:

PROPOSITION 5.53.

Setting $\alpha = 2(EG - F^2)$, we have

- (a) $\alpha \Gamma_{11}^1 = GE_u - 2FF_u + FE_v, \quad \alpha \Gamma_{11}^2 = 2EF_u - EE_v - FE_u,$
- (b) $\alpha \Gamma_{12}^1 = GE_v - FG_u, \quad \alpha \Gamma_{12}^2 = EG_u - FE_v,$
- (c) $\alpha \Gamma_{22}^1 = 2GF_v - GG_u - FG_v, \quad \alpha \Gamma_{22}^2 = EG_v - 2FF_v + FG_u.$

PROOF. The inner product of one of $\{\sigma_{uu}, \sigma_{uv}, \sigma_{vv}\}$ with one of $\{\sigma_u, \sigma_v\}$ can be computed in two ways: (1) using the product rule for inner products, and (2) using Definition 5.50 by which the Christoffel symbols are defined. Equating these approaches gives

- (a) $\Gamma_{11}^1 E + \Gamma_{11}^2 F = \langle \sigma_{uu}, \sigma_u \rangle = \frac{1}{2}E_u, \quad \Gamma_{11}^1 F + \Gamma_{11}^2 G = \langle \sigma_{uu}, \sigma_v \rangle = F_u - \frac{1}{2}E_v,$
- (b) $\Gamma_{12}^1 E + \Gamma_{12}^2 F = \langle \sigma_{uv}, \sigma_u \rangle = \frac{1}{2}E_v, \quad \Gamma_{12}^1 F + \Gamma_{12}^2 G = \langle \sigma_{uv}, \sigma_v \rangle = \frac{1}{2}G_u,$
- (c) $\Gamma_{22}^1 E + \Gamma_{22}^2 F = \langle \sigma_{vv}, \sigma_u \rangle = F_v - \frac{1}{2}G_u, \quad \Gamma_{22}^1 F + \Gamma_{22}^2 G = \langle \sigma_{vv}, \sigma_v \rangle = \frac{1}{2}G_v.$

Each of the above three lines, (a), (b), and (c), is a linear system of two equations involving two Christoffel symbols. Solving these linear systems for the Christoffel symbols yields the desired equations. \square

Sometimes these particular formulas are useful, but often one needs to know only that there exist *some* formulas for the Christoffel symbols purely in terms of the coefficients of the first fundamental form and their partial derivatives. In order to establish that curvature is intrinsic, Gauss went on to derive the following relationships between the functions $\{E, F, G, e, f, g\}$:

PROPOSITION 5.54 (The equations of compatibility).

The following relations hold on all of U :

$$(1u) \quad FK = (\Gamma_{12}^1)_u - (\Gamma_{11}^1)_v + \Gamma_{12}^2 \Gamma_{12}^1 - \Gamma_{11}^2 \Gamma_{22}^1,$$

$$(1v) \quad EK = (\Gamma_{11}^2)_v - (\Gamma_{12}^2)_u + \Gamma_{11}^1 \Gamma_{12}^2 + \Gamma_{11}^2 \Gamma_{22}^2 - \Gamma_{12}^1 \Gamma_{11}^2 - (\Gamma_{12}^2)^2$$

$$(1N) \quad e_v - f_u = e\Gamma_{12}^1 + f(\Gamma_{12}^2 - \Gamma_{11}^1) - g\Gamma_{11}^2,$$

$$(2u) \quad GK = (\Gamma_{22}^1)_u - (\Gamma_{12}^1)_v + \Gamma_{22}^2 \Gamma_{12}^1 + \Gamma_{22}^1 \Gamma_{11}^1 - \Gamma_{12}^2 \Gamma_{22}^1 - (\Gamma_{12}^1)^2$$

$$(2v) \quad FK = (\Gamma_{12}^2)_v - (\Gamma_{22}^2)_u + \Gamma_{12}^1 \Gamma_{12}^2 - \Gamma_{22}^1 \Gamma_{11}^2,$$

$$(2N) \quad g_u - f_v = g\Gamma_{12}^2 + f(\Gamma_{12}^1 - \Gamma_{22}^1) - e\Gamma_{22}^1.$$

PROOF. All six equations come from the following relationships:

$$(1) \quad (\sigma_{uu})_v = (\sigma_{uv})_u, \quad (2) \quad (\sigma_{vv})_u = (\sigma_{uv})_v.$$

More specifically, the three red equations come from equating the coefficients of the basis $\{\sigma_u, \sigma_v, N\}$ in the left and right of (1), while the three blue equations similarly derive from (2).

We will now outline the steps of deriving the three red equations. For this, (1) becomes

$$(5.10) \quad (\Gamma_{11}^1 \sigma_u + \Gamma_{11}^2 \sigma_v + eN)_v = (\Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v + fN)_u.$$

The left side of Eq. 5.10 expands as follows:

$$\begin{aligned} & (\Gamma_{11}^1)_v \sigma_u + \Gamma_{11}^1 \underbrace{(\Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v + fN)}_{\sigma_{uv}} \\ & + (\Gamma_{11}^2)_v \sigma_v + \Gamma_{11}^2 \underbrace{(\Gamma_{22}^1 \sigma_u + \Gamma_{22}^2 \sigma_v + gN)}_{\sigma_{vv}} + e_v N + e \underbrace{(-w_{12} \sigma_u - w_{22} \sigma_v)}_{N_v = -\mathcal{W}(\sigma_v)}, \end{aligned}$$

where the matrix $(\begin{smallmatrix} w_{11} & w_{12} \\ w_{21} & w_{22} \end{smallmatrix})$ was defined in Proposition 4.17 (on page 219) to represent the Weingarten map \mathcal{W} in the basis $\{\sigma_u, \sigma_v\}$. Recall from this proposition that

$$w_{12} = \frac{fG - gF}{EG - F^2}, \quad w_{22} = \frac{gE - fF}{EG - F^2}.$$

Making these substitutions completes the work of expanding the left side of Eq. 5.10. The right side expands in an analogous manner. The three red equations come from equating the resulting coefficients of σ_u , σ_v , and N between the left and the right sides of this expanded version of Eq. 5.10.

The blue equations can be derived analogously. But it is quicker to observe that u and v play symmetric roles, so the blue equations come from the red equations by swapping the following symbols: $u \leftrightarrow v$, $1 \leftrightarrow 2$, $E \leftrightarrow G$, $e \leftrightarrow g$. \square

Since the Christoffel symbols are expressible in terms of the first fundamental form, Gauss's Theorema Egregium follows immediately from any one

of (1u), (1v), (2u), and (2v). More explicitly, substituting the expressions for the Christoffel symbols from Proposition 5.53 into any one of these four equations yields the following formula for K :

$$K = \frac{\begin{vmatrix} -\frac{1}{2}E_{vv} + F_{uv} - \frac{1}{2}G_{uu} & \frac{1}{2}E_u & F_u - \frac{1}{2}E_v \\ F_v - \frac{1}{2}G_u & E & F \\ \frac{1}{2}G_v & F & G \end{vmatrix} - \begin{vmatrix} 0 & \frac{1}{2}E_v & \frac{1}{2}G_u \\ \frac{1}{2}E_v & E & F \\ \frac{1}{2}G_u & F & G \end{vmatrix}}{(EG - F^2)^2}, \quad (5.11)$$

where vertical bars denote determinants. The algebra required to derive this general formula is messy, but is not too bad when performed under the following added hypothesis:

COROLLARY 5.55.

If $E = 1$ and $F = 0$, then

$$K = -\frac{(\sqrt{G})_{uu}}{\sqrt{G}}.$$

This corollary is a generalization of Theorem 5.26 on page 269 (which expressed K in normal polar coordinates). It also generalizes Eq. 4.15 on page 221 (which expressed K for a surface of revolution). In fact, it formalizes what these two situations have in common.

We will end this section with an algebraic characterization of Gaussian curvature: it measures the failure of the familiar “ $\frac{\partial}{\partial v} \frac{\partial}{\partial u} = \frac{\partial}{\partial u} \frac{\partial}{\partial v}$ ” rule to carry over to covariant differentiation:

PROPOSITION 5.56.

If w is a tangent vector field on S , then

$$\underbrace{\frac{D}{dv} \frac{D}{du} w}_{\text{denoted } (\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv})w} - \underbrace{\frac{D}{du} \frac{D}{dv} w}_{= K \cdot |\sigma_u \times \sigma_v| \cdot R_{90}(w)} = K \cdot |\sigma_u \times \sigma_v| \cdot R_{90}(w).$$

Recall from our discussion of Definition 5.50 that if X is any tangent vector field on S , then $\frac{D}{dv}(X)$ means the tangent vector field on S whose value at $p \in S$ equals the covariant derivative at p of the restriction of X to the v -parameter curve through p . The expression “ $\frac{D}{du}$ ” is interpreted analogously. Also recall that $|\sigma_u \times \sigma_v|$ equals the infinitesimal area distortion of σ , while $R_{90}(w)$ denotes the 90° counterclockwise rotation of w .

Each side of the equation in Proposition 5.56 is a tangent vector field on S . The value of the right side at $p \in S$ clearly depends only on the value of w only at p (not at nearby points). The proposition implies that the same must be true of the left side, which is actually part of what must be proven to establish the proposition.

The dependence on σ is also noteworthy. The value of the right side at $p \in S$ would be the same if σ were replaced by any other coordinate

chart covering p that induced the same orientation and had the same area distortion at p . The proposition implies that the same must be true for the left side, which might seem surprising.

PROOF. We will first prove the result when $w = \sigma_u$. The definition of Christoffel symbols together with Lemma 5.39 (on page 280) gives

$$\begin{aligned} \frac{D}{dv} \frac{D}{du} \sigma_u &= \frac{D}{dv} \underbrace{\left(\Gamma_{11}^1 \sigma_u + \Gamma_{11}^2 \sigma_v \right)}_{\frac{D}{du} \sigma_u} \\ &= (\Gamma_{11}^1)_v(\sigma_u) + \Gamma_{11}^1 \underbrace{\left(\Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v \right)}_{\frac{D}{dv} \sigma_u} + (\Gamma_{11}^2)_v(\sigma_v) + \Gamma_{11}^2 \underbrace{\left(\Gamma_{22}^1 \sigma_u + \Gamma_{22}^2 \sigma_v \right)}_{\frac{D}{dv} \sigma_v} \\ &= ((\Gamma_{11}^1)_v + \Gamma_{11}^1 \Gamma_{12}^1 + \Gamma_{11}^2 \Gamma_{22}^1) \sigma_u + ((\Gamma_{11}^2)_v + \Gamma_{11}^1 \Gamma_{12}^2 + \Gamma_{11}^2 \Gamma_{22}^2) \sigma_v. \end{aligned}$$

Similarly,

$$\begin{aligned} \frac{D}{du} \frac{D}{dv} \sigma_u &= \frac{D}{du} \underbrace{\left(\Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v \right)}_{\frac{D}{dv} \sigma_u} \\ &= (\Gamma_{12}^1)_u(\sigma_u) + \Gamma_{12}^1 \underbrace{\left(\Gamma_{11}^1 \sigma_u + \Gamma_{11}^2 \sigma_v \right)}_{\frac{D}{du} \sigma_u} + (\Gamma_{12}^2)_u(\sigma_v) + \Gamma_{12}^2 \underbrace{\left(\Gamma_{12}^1 \sigma_u + \Gamma_{12}^2 \sigma_v \right)}_{\frac{D}{du} \sigma_v} \\ &= ((\Gamma_{12}^1)_u + \Gamma_{11}^1 \Gamma_{12}^1 + \Gamma_{12}^1 \Gamma_{12}^2) \sigma_u + ((\Gamma_{12}^2)_u + \Gamma_{12}^1 \Gamma_{11}^2 + \Gamma_{12}^2 \Gamma_{12}^2) \sigma_v. \end{aligned}$$

Subtracting gives

$$\begin{aligned} \frac{D}{dv} \frac{D}{du} \sigma_u - \frac{D}{du} \frac{D}{dv} \sigma_u &= ((\Gamma_{11}^1)_v - (\Gamma_{12}^1)_u + \Gamma_{11}^2 \Gamma_{22}^1 - \Gamma_{12}^1 \Gamma_{12}^2) \sigma_u \\ &\quad + ((\Gamma_{11}^2)_v - (\Gamma_{12}^2)_u + \Gamma_{11}^1 \Gamma_{12}^2 + \Gamma_{11}^2 \Gamma_{22}^2 - \Gamma_{12}^1 \Gamma_{11}^2 - \Gamma_{12}^2 \Gamma_{12}^2) \sigma_v \\ &= -(FK) \sigma_u + (EK) \sigma_v \quad (\text{by Proposition 5.54}) \\ &= K(E\sigma_v - F\sigma_u) \\ &= K(|\sigma_u|^2 \sigma_v - \langle \sigma_u, \sigma_v \rangle \sigma_u) \\ &= K|\sigma_u|^2 \left(\sigma_v - \frac{\langle \sigma_u, \sigma_v \rangle \sigma_u}{|\sigma_u|^2} \right) \\ &= K|\sigma_u|^2 (\sigma_v)^\perp = K|\sigma_u| |\sigma_v^\perp| \left(|\sigma_u| \frac{\sigma_v^\perp}{|\sigma_v^\perp|} \right) \\ &= K|\sigma_u| \times \sigma_v |R_{90}(\sigma_u)|, \end{aligned}$$

where $(\sigma_v)^\perp$ denotes the projection of σ_v orthogonal to σ_u .

This proves the result when $w = \sigma_u$. The result when $w = \sigma_v$ is proven analogously. Since an arbitrary tangent vector field on S has the form $w = a\sigma_u + b\sigma_v$ for some smooth functions $a, b : U \rightarrow \mathbb{R}$, the general result will follow from these two special cases, provided we can establish the following linearity property:

$$\left(\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv} \right) (a\sigma_u + b\sigma_v) = a \left(\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv} \right) \sigma_u + b \left(\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv} \right) \sigma_v.$$

This linearity property really says that the value at p of $\left(\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv} \right) w$ depends only on the value of w at p (not at nearby points). Its proof is straightforward calculation using the algebraic properties of covariant differentiation, and is left to the reader in Exercise 5.66. \square

Proposition 5.56 and its proof seem purely algebraic, but the next section will include a geometric interpretation of the expression $\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv}$ as “infinitesimal holonomy.”

EXERCISES

EXERCISE 5.56. Prove Proposition 4.12 on page 210.

EXERCISE 5.57. In normal coordinates, prove that all Christoffel symbols vanish at the origin. *HINT: See Corollary 5.18(1).*

EXERCISE 5.58 (Fermi Coordinates). Let $\gamma : [a, b] \rightarrow S$ be a unit-speed geodesic without self-intersections in a regular surface S . Let w be a unit-length parallel vector field along γ that is orthogonal to γ' . For sufficiently small $\epsilon > 0$, prove that the function $\sigma : (-\epsilon, \epsilon) \times (a, b) \rightarrow S$ defined as

$$\sigma(s, t) = \exp_{\gamma(t)}(s \cdot w(t))$$

is a surface patch. Prove that all Christoffel symbols vanish at points of the domain where $s = 0$. What can be said about the first fundamental form of σ ?

EXERCISE 5.59. Use Lemma 5.51 and Proposition 5.53 to give an alternative proof that covariant differentiation is intrinsic (Proposition 5.49 on page 287).

EXERCISE 5.60. For the natural surface patch of a graph (Example 3.80 on page 185), find formulas for all Christoffel symbols in terms of the functions f .

EXERCISE 5.61. If $F = 0$, prove that

$$K = -\frac{1}{2\sqrt{EG}} \left(\frac{\partial}{\partial u} \left(\frac{G_u}{\sqrt{EG}} \right) + \frac{\partial}{\partial v} \left(\frac{E_v}{\sqrt{EG}} \right) \right).$$

EXERCISE 5.62. If $F = 0$ and $E = G$, prove that

$$K = -\frac{(\ln G)_{uu} + (\ln G)_{vv}}{2G}.$$

EXERCISE 5.63. Expand each side of the equation $N_{uv} = N_{vu}$ as a linear combination of the basis $\{\sigma_u, \sigma_v, N\}$. Show that equating coefficients yields equations that are redundant with those from Proposition 5.54.

EXERCISE 5.64. Consider the parametrized surfaces

$$\sigma_1(u, v) = (u \cos v, u \sin v, v), \quad \sigma_2(u, v) = (u \cos v, u \sin v, \ln u).$$

- (1) Prove that they have the same Gaussian curvature: $K_1(u, v) = K_2(u, v)$.
- (2) Prove that $\sigma_1(u, v) \mapsto \sigma_2(u, v)$ is not an isometry.

COMMENT: Like Exercise 5.40 on page 275, this demonstrates that the converse to the Theorema Egregium is false in the sense that a Gaussian-curvature-preserving diffeomorphism between surfaces need not be an isometry.

EXERCISE 5.65. If a regular surface is covered by a surface patch all of whose u - and v - coordinate curves are geodesics, prove that it has Gaussian curvature zero.

EXERCISE 5.66. Prove the linearity property described at the end of the proof of Proposition 5.56.

EXERCISE 5.67. Let $\gamma : [a, b] \rightarrow \mathbb{R}^3$ be a unit-speed space curve with nonzero curvature and with no self-intersections. Prove there exists a regular surface S such that γ is a geodesic in S . *HINT:* Define $\sigma(s, t) = \gamma(t) + s \cdot \mathbf{b}(t)$, where $\mathbf{b}(t)$ is the unit binormal.

EXERCISE 5.68. Suppose that $E = 1$ and $F = 0$.

- (1) Verify that

$$\Gamma_{11}^1 = \Gamma_{11}^2 = \Gamma_{12}^1 = 0, \quad \Gamma_{12}^2 = \frac{G_u}{2G}, \quad \Gamma_{22}^1 = \frac{-G_u}{2}, \quad \Gamma_{22}^2 = \frac{G_v}{2G}.$$

- (2) Verify that the geodesic equations become

$$u'' - (v')^2 \frac{G_u}{2} = 0 \quad \text{and} \quad v'' + (u'v') \frac{G_u}{G} + (v')^2 \frac{G_v}{2G} = 0.$$

- (3) If additionally $G_v = 0$, show that the first equation of (2) is equivalent to

$$(*) \quad (u'G)' = 0.$$

Verify that these hypotheses are satisfied by our usual coordinate chart for a surface of revolution (with $u \leftrightarrow t$ and $v \leftrightarrow \theta$). In this case, confirm that $(*)$ is equivalent to Eq. 5.4 in the proof of Clairaut's theorem (on page 253).

- (4) In normal polar coordinates (with the usual variable names $u \leftrightarrow r$ and $v \leftrightarrow \theta$), the first equation of (2) becomes

$$(**) \quad r'' = \frac{1}{2}(\theta')^2 G_r.$$

Use this to improve the claim of Exercise 5.28(1) (on page 267) by showing that $\sqrt{f} \circ \beta$ is concave up, provided β is not a portion of a geodesic through p .

- (5) Let σ be the surface patch for normal polar coordinates, and suppose that $\beta(t) = \sigma(r(t), \theta(t))$ is a geodesic expressed in these coordinates. Let $\Psi(t) \in [0, 2\pi)$ denote the angle from σ_r to $\beta'(t)$ (counterclockwise with respect to the orientation induced by σ). Prove that

$$\Psi' = -\theta' \left(\sqrt{G} \right)_r.$$

HINT: If β is of unit speed, then $\sin \Psi = |\sigma_r \times \beta'| = \theta' \sqrt{G}$ and $\cos \Psi = \langle \sigma_r, \beta' \rangle = r'$. Therefore

$$r'' = (\cos \Psi)' = -\Psi'(\sin \Psi) = -\Psi'(\theta' \sqrt{G}).$$

*Now use (**).*

- (6) (**Gauss–Bonnet for small geodesic triangles**) In Fig. 5.19, the three edges of the illustrated “triangle” T are geodesics within a normal coordinate chart at the vertex p . Prove that the sum of the interior angles is

$$a + b + c = \pi + \iint_T K \, dA.$$

HINT: With the notation of (5), reparametrize β with $\theta(t) = t$, so that

$$\iint_T K \, dA = \int_{t_1}^{t_2} \int_{r(t_1)}^{r(t_2)} K \|\sigma\| \, dr \, dt = \int_{t_1}^{t_2} \int_{r(t_1)}^{r(t_2)} - \left(\sqrt{G} \right)_{rr} \, dr \, dt,$$

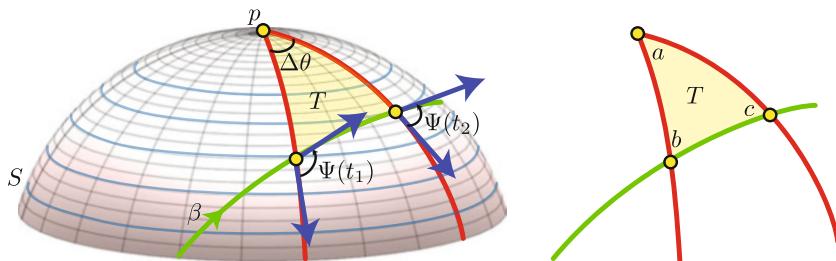


FIGURE 5.19. Gauss–Bonnet for small geodesic triangles

and apply the fundamental theorem of calculus. Even though β no longer has constant speed, the conclusion of (5) can still help simplify the resulting expression to $\iint_T K \, dA = \Delta\Psi + \Delta\theta$, where $\Delta\Psi = \Psi(t_2) - \Psi(t_1)$ and $\Delta\theta = t_2 - t_1$.

□

7. Gaussian Curvature Measures Infinitesimal Holonomy

In this section, we will generalize our previous definitions of *parallel transport*, *angle function*, and *holonomy* to curves that are only *piecewise* regular. This will allow us to interpret the expression $\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv}$ from Proposition 5.56 (on page 293) as “infinitesimal holonomy.”

Assume throughout this section that S is an oriented regular surface and that $\gamma : [a, b] \rightarrow S$ is a *piecewise*-regular curve in S . According to Definition 2.6 on page 68, this means that γ is continuous and there exists a partition, $a = t_0 < t_1 < \dots < t_n = b$, such that the restriction of γ to each subinterval $[t_i, t_{i+1}]$ (which we denote by γ_i) is a regular curve in S . As before, γ is called **closed** if $\gamma(a) = \gamma(b)$, **simple** if γ is one-to-one on the domain $[a, b]$, and of **unit speed** if each γ_i is of unit speed.

Several key definitions generalize naturally to this setting:

DEFINITION 5.57.

Let v be a **vector field along** γ , which means a continuous function $v : [a, b] \rightarrow \mathbb{R}^3$ whose restriction to each subinterval $[t_i, t_{i+1}]$ is a vector field along γ_i .

- (1) v is called **parallel** along γ if its restriction to each subinterval $[t_i, t_{i+1}]$ is parallel along γ_i .
- (2) **Parallel transport** along γ is the function $P_\gamma : T_{\gamma(a)}S \rightarrow T_{\gamma(b)}S$ defined by parallel-transporting one smooth segment at a time:

$$P_\gamma = P_{\gamma_{n-1}} \circ P_{\gamma_{n-2}} \circ \cdots \circ P_{\gamma_1} \circ P_{\gamma_0}.$$

- (3) If γ is closed, then P_γ is also called the **holonomy** around γ .

For every $w \in T_{\gamma(a)}S$, notice that $P_\gamma(w) \in T_{\gamma(b)}S$ is the value at $t = b$ of the unique parallel vector field along γ whose value at $t = a$ equals w , just as in the smooth case.

Figure 5.20 illustrates a parallel vector field along a piecewise-regular curve in the plane (where parallel vector fields are constant). The visual idea would be the same in an arbitrary surface: at each corner, the parallel field changes continuously, while the curve’s direction changes abruptly. We next discuss a meaningful way to measure this abrupt change in the curve’s direction.

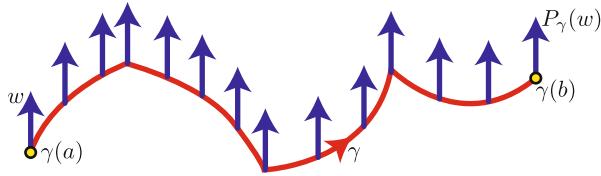
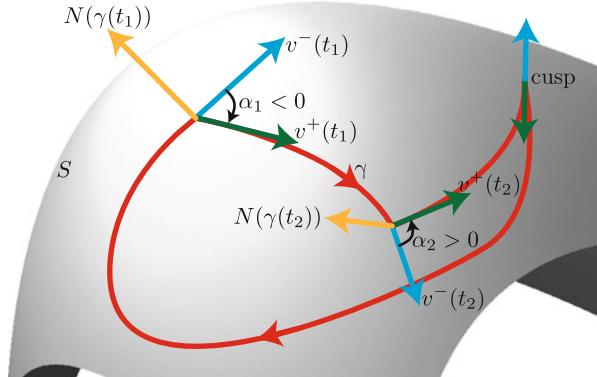


FIGURE 5.20. Parallel transport along a piecewise-regular curve

Each nonsmooth point $\gamma(t_i)$ is called a **vertex**² (or a **corner**) of γ . At each vertex $\gamma(t_i)$, the **signed angle**, $\alpha_i \in [-\pi, \pi]$, is defined exactly as in Sect. 1 of Chap. 2. More precisely, the absolute value of α_i equals the smallest determination of the angle between $v^-(t_i)$ and $v^+(t_i)$ (defined exactly as before), while the sign of α_i is positive if $v^+(t_i)$ is a counterclockwise rotation of $v^-(t_i)$ through this angle (and negative if clockwise). The meaning of “clockwise/counterclockwise” comes from the orientation (as in Eqs. 3.11 and 3.12 on page 150). Equivalently, α_i is positive if and only if $\{v^-(t_i), v^+(t_i), N(\gamma(t_i))\}$ is a positively oriented ordered basis of \mathbb{R}^3 , where N denotes the orientation of S ; see Fig. 5.21.

If γ is not closed, then $\gamma(a)$ and $\gamma(b)$ do *not* count as vertices. If γ is closed and $\gamma'(a) \neq \gamma'(b)$, then $\gamma(a) = \gamma(b)$ counts as a single vertex, and the corresponding signed angle is defined as above, but with $v^-(a)$ replaced by $v^-(b)$.

FIGURE 5.21. The signed angles at the vertices of a piecewise-regular simple closed curve in S

²For the remainder of the book, forget about the other unrelated meaning for “vertex” from the four vertex theorem in Sect. 2 of Chap. 2.

As before, the vertex $\gamma(t_i)$ is called a **cusp** if $v^+(t_i)$ is a negative scalar multiple of $v^-(t_i)$. The decision whether the signed angle at a cusp is π or $-\pi$ is made using the same process as before, but this process requires a notion of “orientation” and “interior,” which we will not define until Sect. 1 of Chap. 6. So for now, we will assume for simplicity there are no cusps. The notions of *angle function* and *angle displacement* generalize naturally to this setting:

PROPOSITION AND DEFINITION 5.58.

Assume that γ is of unit speed. Let $\{\alpha_i\}$ denote the signed angles at its vertices.

- (1) A function $\theta : [a, b] \rightarrow \mathbb{R}$ is called an **angle function** of γ if for each $i \in \{1, \dots, n\}$, θ has a jump discontinuity by α_i at t_i , and elsewhere satisfies $\theta'(t) = \kappa_g(t)$. An angle function exists and is unique up to an additive constant.
- (2) The **angle displacement** along γ means the net change in an angle function: $\Delta\theta = \theta(b) - \theta(a)$. Equivalently,

$$\Delta\theta = \int_a^b \kappa_g(t) dt + \sum_i \alpha_i,$$

where “ $\int_a^b \kappa_g(t) dt$ ” is shorthand for the sum of the integrals of the geodesic curvature over the smooth segments of γ .

- (3) If γ is closed, then P_γ equals a clockwise rotation of $T_{\gamma(a)} S$ by $\Delta\theta$.

PROOF. Exercise 5.69. □

Notice that (3) generalizes Proposition 5.46 on page 284. It is straightforward to show that the angle function, θ , relates γ' to an arbitrary parallel vector field along γ exactly as in Proposition 5.42. Furthermore, as in the proof of Theorem 2.7 on page 69, each vertex can be smoothed, and the signed angle at the vertex represents the net change in an angle function of the smoothed version (in the limit as the smoothing occurs in a smaller and smaller neighborhood of the vertex).

We are now able to provide an “infinitesimal holonomy” interpretation for the expression $\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv}$ from Proposition 5.56 on page 293:

PROPOSITION 5.59 (Gaussian curvature measures infinitesimal holonomy).

Let S be an oriented regular surface, $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ a surface patch compatible with the orientation, $p \in V$, and $w \in T_p S$. For small $s > 0$, let $\gamma_s : [0, 4\sqrt{s}] \rightarrow S$ denote the composition with σ of the unit-speed counterclockwise square in U with bottom-left corner $q = \sigma^{-1}(p)$ and side length \sqrt{s} (as in Fig. 5.22). Let $w(s) \in T_p S$ be the parallel transport of w around γ_s . Then

$$w'(0) = \left(\left(\frac{D}{dv} \frac{D}{du} - \frac{D}{du} \frac{D}{dv} \right) W \right) (p) \underbrace{= K(q) \cdot |\sigma_u(q) \times \sigma_v(q)| \cdot R_{90}(w)}_{\text{by Proposition 5.56}},$$

where W is any extension of w to a tangent vector field in a neighborhood of p .

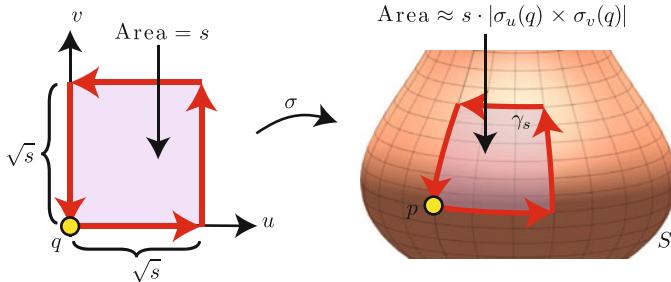


FIGURE 5.22. γ_s is a small “parameter square” at p

PROOF. We know from the proof of Proposition 5.56 that the left side is independent of the choice of extension W , so we begin by choosing a convenient one. For this, we assume for simplicity that $q = (0, 0)$, and for (u, v) close to $(0, 0)$, we let $W(u, v) \in T_{\sigma(u,v)} S$ denote the parallel transport of w along the u -parameter curve from p to $\sigma(u, 0)$ followed by the v -parameter curve from $\sigma(u, 0)$ to $\sigma(u, v)$. With this choice, $\frac{D}{dv} W = \mathbf{0}$ near p , so it remains to prove that $w'(0) = (\frac{D}{dv} \frac{D}{du} W)(p)$.

Define $A(u, v) \in T_{\sigma(0,v)} S$ to be the parallel transport of $W(u, v)$ along the u -parameter curve from $\sigma(u, v)$ to $\sigma(0, v)$. Notice that for fixed v_0 , $u \mapsto A(u, v_0)$ is a curve in $T_{\sigma(0,v_0)} S$ whose initial derivative can be computed by Exercise 5.53 (on page 288) as

$$\frac{d}{du} \Big|_{u=0} A(u, v_0) = \left(\frac{D}{du} W \right) (0, v_0).$$

Next define $B(u, v) \in T_p S$ to equal the parallel transport of $A(u, v)$ along the v -parameter curve from $\sigma(0, v)$ to p . Using Exercise 5.53 once more gives

$$\left(\frac{\partial^2}{\partial v \partial u} B \right) (0, 0) = \left(\frac{D}{dv} \frac{D}{du} W \right) (p).$$

Since $B : (\text{neighborhood of } (0, 0) \text{ in } \mathbb{R}^2) \rightarrow T_p S$ sends the u -axis and the v -axis to the vector w , the second-order Taylor polynomial for B (Eq. 3.6 on page 118) simplifies to

$$B(u, v) - B(0, 0) \approx uv \left(\frac{\partial^2}{\partial v \partial u} B \right) (0, 0) = uv \left(\frac{D}{dv} \frac{D}{du} W \right) (p).$$

Since $w(s) = B(\sqrt{s}, \sqrt{s})$,

$$w'(0) = \lim_{s \rightarrow 0} \frac{B(\sqrt{s}, \sqrt{s}) - B(0, 0)}{s} = \left(\frac{D}{dv} \frac{D}{du} W \right) (p).$$

□

The conclusion of the proposition is

$$(5.12) \quad w'(0) = K(q) \cdot |\sigma_u(q) \times \sigma_v(q)| \cdot R_{90}(w).$$

The choice of $w \in T_p S$ is not important here, since all choices work the same. To understand this, let $(\Delta\theta)(s)$ denote the angle displacement around a unit-speed reparametrization of γ_s . For every fixed choice of w , $w(s)$ equals the clockwise rotation of w by $(\Delta\theta)(s)$; that is,

$$w(s) = \cos((-\Delta\theta)(s)) w + \sin(-(\Delta\theta)(s)) R_{90}(w).$$

Notice that $w'(0) = -(\Delta\theta)'(0) \cdot R_{90}(w)$. Thus, Eq. 5.12 really says

$$(5.13) \quad (\Delta\theta)'(0) = -K(q) \cdot |\sigma_u(q) \times \sigma_v(q)|.$$

Since $\lim_{s \rightarrow 0} (\Delta\theta)(s) = 2\pi$, the first-order approximation is

$$(\Delta\theta)(s) \approx 2\pi - K(q) \cdot s \cdot \underbrace{|\sigma_u(q) \times \sigma_v(q)|}_{\text{approx area in } \gamma_s}.$$

In summary, for small s , the holonomy around γ_s is a clockwise rotation by approximately 2π minus $K(q)$ times the area enclosed in γ_s . It is equivalent and somewhat simpler so say, *For small s , the holonomy around γ_s is a counterclockwise rotation by approximately $K(q)$ times the area enclosed in γ_s .* Either interpretation justifies the title of Proposition 5.59.

The main goal of the next chapter is to obtain a global version of this infinitesimal result. Under certain conditions, even around a large closed loop, the holonomy equals a counterclockwise rotation by an angle equal to the integral of K over the region enclosed!

EXERCISES

EXERCISE 5.69. Prove Proposition 5.58.

EXERCISE 5.70. Compute in terms of θ and ϕ_0 the angle displacement around the piecewise-regular loops in S^2 illustrated in Fig. 5.23.

EXERCISE 5.71. Prove that every angle can be realized as the angle displacement around a piecewise-regular loop in S^2 .

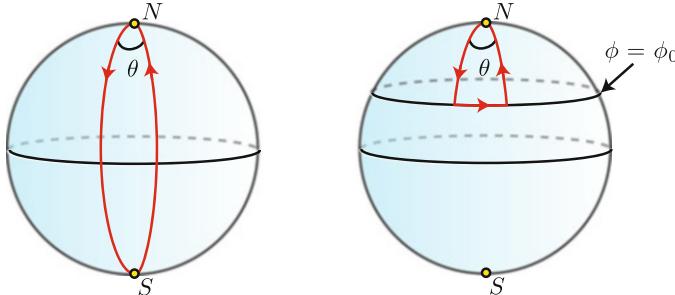


FIGURE 5.23. What is the angle displacement around these loops?

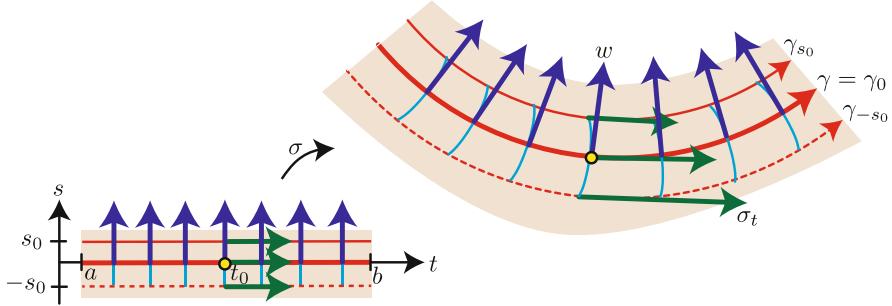
8. Arc-Length Variation: Tire Tracks on a Curved Surface (Optional)

In Sect. 6 of Chap. 1 (page 37), we studied the tire tracks of a chariot driven along a curve in the plane. At any instant, the speed difference between the right and left wheels is proportional to the signed curvature of the path traversed by the center. A south-pointing chariot (also discussed in Sect. 6 of Chap. 1) has an upright statue whose gearing causes it to spin clockwise (relative to the chariot) at a rate proportional to this speed difference between the wheels. The result justifies its name—it always points in the same direction, even as the chariot turns left or right.

One goal of this section is to study a regular chariot or a south-pointing chariot that is ridden along a curve in an *arbitrary* surface. This problem (and others) can be framed using the following very general vocabulary:

DEFINITION 5.60.

Let $\gamma : I \rightarrow S$ be a regular curve in a regular surface. A **variation** of γ is a smooth function $\sigma : (-\epsilon, \epsilon) \times I \rightarrow S$ (for some $\epsilon > 0$) such that $\gamma(t) = \sigma(0, t)$ for all $t \in I$. For each fixed $s \in (-\epsilon, \epsilon)$, the path $\gamma_s(t) = \sigma(s, t)$ is called a **curve of the variation**. The vector field $w(t) = \sigma_s(0, t)$ along γ is called the **variational vector field** of σ . The variation is called **orthogonal** if $w(t) \perp \gamma'(t)$ for all $t \in I$. It is called **proper** if $I = [a, b]$ and the curves of the variation have common endpoints; that is, $\sigma(s, a) = \gamma(a)$ and $\sigma(s, b) = \gamma(b)$ for all $s \in (-\epsilon, \epsilon)$.

FIGURE 5.24. An orthogonal variation of γ

A variation is a very general construction with many applications including the chariot problem. Working in this generality, we begin by computing the rate of change of the arc length of the curves of the variation with respect to s :

PROPOSITION 5.61 (First variation of arc length).

Let S be a regular surface and $\gamma : [a, b] \rightarrow S$ a unit-speed curve. If σ is an orthogonal variation of γ and w is the variational vector field of σ , then

$$\frac{d}{ds} \Big|_{s=0} \text{length}(\gamma_s) = - \int_a^b \langle \gamma''(t), w(t) \rangle dt.$$

PROOF. Fix $t_0 \in [a, b]$, and consider the vector field $s \mapsto \sigma_t(s, t_0)$ along the curve $s \mapsto \sigma(s, t_0)$. In Fig. 5.24, this is illustrated as a green vector field along a turquoise curve. Its squared length has the following initial derivative:

$$\begin{aligned} \frac{d}{ds} \Big|_{s=0} |\sigma_t(s, t_0)|^2 &= \frac{d}{ds} \Big|_{s=0} \langle \sigma_t(s, t_0), \sigma_t(s, t_0) \rangle = 2 \langle \sigma_{ts}(0, t_0), \sigma_t(0, t_0) \rangle \\ &= 2 \langle \sigma_{st}(0, t_0), \sigma_t(0, t_0) \rangle = 2 \langle w'(t_0), \gamma'(t_0) \rangle \\ &= 2 \left(\underbrace{\frac{d}{dt} \Big|_{t=t_0} \langle w(t), \gamma'(t) \rangle}_0 - \langle w(t_0), \gamma''(t_0) \rangle \right) = -2 \langle w(t_0), \gamma''(t_0) \rangle. \end{aligned}$$

The initial derivative of the square root of the above expression is

$$(5.14) \quad \frac{d}{ds} \Big|_{s=0} |\gamma'_s(t_0)| = \frac{d}{ds} \Big|_{s=0} |\sigma_t(s, t_0)| = -\langle w(t_0), \gamma''(t_0) \rangle.$$

Therefore,

$$\begin{aligned} \frac{d}{ds} \Big|_{s=0} \text{length}(\gamma_s) &= \frac{d}{ds} \Big|_{s=0} \int_a^b |\gamma'_s(t)| dt \\ &= \int_a^b \frac{d}{ds} \Big|_{s=0} |\gamma'_s(t)| dt = - \int_a^b \langle w(t), \gamma''(t) \rangle dt. \end{aligned}$$

□

EXAMPLE 5.62 (Chariot tire tracks). To model the chariot problem, we must assume that S is oriented, so that it makes sense to imagine the chariot riding on the “top” of S , with γ representing the curve traversed by the center of the chariot (the midpoint on the axle between its wheels). We also assume that w is of unit length. We want w to point toward the left tire, so we assume that the orientation is such that $R_{90}(\gamma'(t)) = w(t)$. If the axle length is $2s_0$, then the curves γ_{s_0} and γ_{-s_0} respectively approximate the paths of the left and right chariot tires; see Fig. 5.24. With these assumptions, the integrand in Proposition 5.61 becomes

$$\langle \gamma''(t), w(t) \rangle = \langle \gamma''(t), R_{90}(\gamma'(t)) \rangle = \kappa_g(t).$$

Therefore, Eq. 5.14 and Proposition 5.61 become

(5.15)

$$\frac{d}{ds}\Big|_{s=0} |\gamma'_s(t_0)| = -\kappa_g(t_0) = -\theta'(t_0), \quad \frac{d}{ds}\Big|_{s=0} \text{length}(\gamma_s) = - \int_a^b \kappa_g(t) dt = -\Delta\theta,$$

where $\Delta\theta$ denotes the net change in an angle function, θ , of γ .

Equations 5.15 generalize the conclusions of Example 1.41 from Sect. 6 of Chap. 1. They report the rate of change of tire speed (left formula) and of tire-track length (right formula) with respect to axle length. The left formula determines how a south-pointing chariot will behave:

EXAMPLE 5.63 (South-pointing chariot). Suppose the chariot from Example 5.62 is a south-pointing chariot. Assume that the chariot’s axle length, $2s_0$, is small compared to the features of the surface. More precisely, assume that Eq. 5.15 (left) gives rise to a good first-order approximation of tire speed. Then at every instant t_0 ,

$$(\text{right-wheel speed minus left-wheel speed}) \approx s_0 \cdot \kappa_g(t_0) = s_0 \cdot \theta'(t_0).$$

Since the gearing of a south-facing chariot causes the statue’s clockwise angular speed (relative to the chariot) to be proportional to this wheel-speed difference, the statue will approximately point in the direction of a parallel vector field. In other words, the statue’s clockwise rotation relative to the chariot (the rotation as seen by an observer riding on the chariot) will approximately be an angle function for γ .

These conclusions remain valid when γ is only piecewise regular, provided the chariot “pivots” at each vertex, so that the chariot’s direction changes but the direction in which the statue points does not.

Proposition 5.61 has several applications unrelated to the chariot problem. For example, it provides a useful alternative characterization of geodesics:

PROPOSITION 5.64 (Variational characterization of geodesics).

A unit-speed curve $\gamma : [a, b] \rightarrow S$ in a regular surface is a geodesic if and only if $\frac{d}{ds}\Big|_{s=0} \text{length}(\gamma_s) = 0$ for every proper variation of γ .

SKETCH OF PROOF. If γ is a geodesic, then $\langle \gamma''(t), w(t) \rangle = 0$ for every variation, so the first derivative of arc length equals zero by Proposition 5.61.

Conversely, suppose that γ is not a geodesic, so there exists $t_0 \in (a, b)$ such that $\gamma''(t_0)$ is *not* normal to S . By continuity, there exists $\delta > 0$ such that $\gamma''(t)$ is not normal to S for all $t \in I = (t_0 - \delta, t_0 + \delta)$. Let w be a vector field along γ that is nonzero for all $t \in I$ and is zero for all values of t outside of I . It is possible to construct a *proper* variation whose variational vector field is w (Exercise 5.72). The above choices can be made such that $\langle \gamma''(t), w(t) \rangle$ has a constant sign on I , and equals zero outside of I . So according to Proposition 5.61, the first derivative of arc length is nonzero. \square

In Proposition 5.64, imagine the trace of γ as a rubber band stretched taught along S , pinned at the endpoints. A taught rubber band wants to shrink itself. If γ is not a geodesic, then it is able to do so by deforming along a variation for which the first derivative of arc length is negative. If γ is a sufficiently short geodesic, then it cannot shrink itself, because it is minimizing (Proposition 5.19). If γ is a long geodesic, then it might or might not be able to shrink itself. Computing higher derivatives of arc length can sometimes help us decide. The second derivative is described in the following:

PROPOSITION 5.65 (Second variation of arc length).

If S is a regular surface, $\gamma : [a, b] \rightarrow S$ is a unit-speed geodesic, σ is a proper orthogonal variation of γ , and w is the variational vector field of σ , then

$$\frac{d^2}{ds^2} \Big|_{s=0} \text{length}(\gamma_s) = \int_a^b (|w^\theta(t)|^2 - K(t)|w(t)|^2) dt,$$

where $K(t) = K(\gamma(t))$ is the Gaussian curvature.

PROOF. Even though σ is not a surface patch, we will apply to σ some definitions and formulas that were derived in Sect. 6 for surface patches. The reader can check that these applications are valid in this context. In particular, the expressions $\frac{D}{dt}$ and $\frac{D}{ds}$ make sense (as in the discussion following Definition 5.50), and

$$(5.16) \quad \frac{D}{dt} \sigma_s = \frac{D}{ds} \sigma_t,$$

because both equal the projection onto the tangent plane of $\sigma_{st} = \sigma_{ts}$. Furthermore, Proposition 5.56 is valid (subbing $u \leftrightarrow s$ and $v \leftrightarrow t$), provided that “ R_{90} ” is interpreted with respect to the local orientation $N = \frac{\sigma_s \times \sigma_t}{|\sigma_s \times \sigma_t|}$. Notice that “ R_{90} ” is not defined at a point of the domain of σ where $|\sigma_s \times \sigma_t| = 0$, but this will not affect the calculations below.

The first derivative at an arbitrary point of the domain of the variation is

$$\frac{\partial}{\partial s} |\sigma_t|^2 = \frac{\partial}{\partial s} \langle \sigma_t, \sigma_t \rangle = 2 \langle \sigma_{ts}, \sigma_t \rangle = 2 \left\langle \frac{D}{ds} \sigma_t, \sigma_t \right\rangle.$$

The second derivative is

$$\begin{aligned}
 \frac{\partial^2}{\partial s^2} |\sigma_t|^2 &= 2 \frac{\partial}{\partial s} \left\langle \frac{D}{ds} \sigma_t, \sigma_t \right\rangle \\
 &= 2 \left\langle \frac{D}{ds} \frac{D}{ds} \sigma_t, \sigma_t \right\rangle + 2 \left| \frac{D}{ds} \sigma_t \right|^2 \quad (\text{Lemma 5.39}) \\
 &= 2 \left\langle \frac{D}{ds} \frac{D}{dt} \sigma_s, \sigma_t \right\rangle + 2 \left| \frac{D}{dt} \sigma_s \right|^2 \quad (\text{Equation 5.16}) \\
 &= 2 \left\langle \frac{D}{dt} \frac{D}{dt} \sigma_s - K \cdot |\sigma_s \times \sigma_t| \cdot R_{90}(\sigma_s), \sigma_t \right\rangle + 2 \left| \frac{D}{dt} \sigma_s \right|^2 \quad (\text{Proposition 5.56}) \\
 &= 2 \frac{\partial}{\partial t} \left\langle \frac{D}{ds} \sigma_s, \sigma_t \right\rangle - 2 \left\langle \frac{D}{ds} \sigma_s, \frac{D}{dt} \sigma_t \right\rangle - 2K \cdot |\sigma_s \times \sigma_t| \cdot \langle R_{90}(\sigma_s), \sigma_t \rangle + 2 \left| \frac{D}{dt} \sigma_s \right|^2.
 \end{aligned}$$

When $s = 0$, $\frac{D}{dt} \sigma_t = \mathbf{0}$, because γ is a geodesic, so the above expression simplifies to

$$\begin{aligned}
 \frac{\partial^2}{\partial s^2} \Big|_{s=0} |\sigma_t|^2 &= 2 \frac{d}{dt} \left\langle \frac{D}{ds} \sigma_s, \gamma'(t) \right\rangle - 2K \cdot |w(t) \times \gamma'(t)| \underbrace{\left\langle R_{90}(w(t)), \gamma'(t) \right\rangle}_{|w(t)| \gamma'(t)} + 2 |w^\vartheta(t)|^2 \\
 &= 2 \frac{d}{dt} \left\langle \frac{D}{ds} \sigma_s, \gamma'(t) \right\rangle - 2K \cdot |w(t)|^2 + 2 |w^\vartheta(t)|^2.
 \end{aligned}$$

The second derivative of the square root of the above expression is therefore

$$\frac{\partial^2}{\partial s^2} \Big|_{s=0} |\sigma_t| = \frac{d}{dt} \left\langle \frac{D}{ds} \sigma_s, \gamma'(t) \right\rangle - K|w(t)|^2 + |w^\vartheta(t)|^2.$$

Thus,

$$\begin{aligned}
 \frac{d^2}{ds^2} \Big|_{s=0} \text{length}(\gamma_s) &= \frac{d^2}{ds^2} \Big|_{s=0} \int_a^b |\gamma'_s(t)| dt = \int_a^b \frac{d^2}{ds^2} \Big|_{s=0} |\gamma'_s(t)| dt \\
 &= \int_a^b \left(\frac{d}{dt} \left\langle \frac{D}{ds} \sigma_s, \gamma'(t) \right\rangle - K|w(t)|^2 + |w^\vartheta(t)|^2 \right) dt \\
 &= \int_a^b (-K|w(t)|^2 + |w^\vartheta(t)|^2) dt.
 \end{aligned}$$

The last equality uses the fundamental theorem of calculus and the fact that w is proper and orthogonal. \square

Even in the simplest example we can think of, the second variation formula gives very interesting results:

EXAMPLE 5.66. Let $\gamma : [0, l] \rightarrow S^2$ be a unit-speed geodesic in S^2 . If $l \leq \pi$, then γ is minimizing, so there is no shorter path between its endpoints. Assume now that $\pi < l < 2\pi$. In this case, γ is not minimizing, but the obvious shorter path between its endpoints is far away from γ . It is perhaps not visually clear whether any length-decreasing proper variation of γ exists. If a rubber band is stretched more than halfway around a great circle of a sphere and pinned at its endpoints, will it slip? Try a physical experiment. Use a sphere that is frictionless (or at least slippery). Does the rubber band slip? Proposition 5.65 can help us decide whether any particular proper orthogonal variation is length-decreasing. The most geometrically natural variation is

pictured in Fig. 5.25 and described as follows. Let $u(t) = R_{90}(\gamma'(t))$, which is a unit-length parallel field along γ . Let $w(t) = \sin\left(\frac{t\pi}{l}\right) u(t)$, which is an orthogonal vector field along γ . Let σ be a proper variation whose variational vector field equals w , which exists by Exercise 5.72. Proposition 5.61 gives $\frac{d}{ds}|_{s=0} \text{length}(\gamma_s) = 0$, while Proposition 5.65 gives

$$\begin{aligned} \frac{d^2}{ds^2} \Big|_{s=0} \text{length}(\gamma_s) &= \int_0^l (|w^\theta(t)|^2 - K(s)|w(s)|^2) \, dt \\ &= \int_0^l \left(\left(\frac{\pi}{l} \cos\left(\frac{t\pi}{l}\right) \right)^2 - 1 \left(\sin\left(\frac{t\pi}{l}\right) \right)^2 \right) \, dt \\ &= \frac{\pi^2 - l^2}{2l} < 0, \end{aligned}$$

so this variation is length-decreasing.

The calculation in this example suggests that the rubber band will slip. This calculation also led Bonnet to prove a beautiful global theorem:

THEOREM 5.67 (Bonnet's Theorem).

If the Gaussian curvature of a complete connected regular surface S satisfies $K \geq \delta$ for some $\delta > 0$, then S is compact and

$$\text{diam}(S) \leq \frac{\pi}{\sqrt{\delta}}.$$

Recall from Exercise 5.42 on page 279 the definition of **diameter**,

$$\text{diam}(S) = \sup\{d(p, q) \mid p, q \in S\},$$

and the assertion that a complete surface with finite diameter must be compact. Thus, the first claim of Bonnet's theorem follows from the second.

PROOF. We'll first prove the case $\delta = 1$. In this case, suppose to the contrary that $\text{diam}(S) > \pi$. Then there must exist a pair of points $p, q \in S$

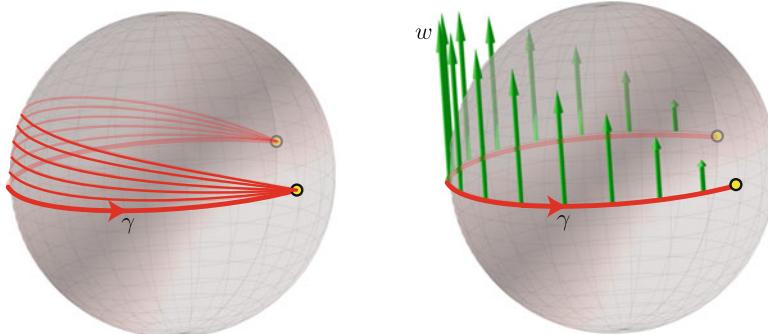


FIGURE 5.25. Will the rubber band slip?

with $l = d(p, q) > \pi$. By the Hopf–Rinow theorem (on page 277), there exists a minimizing geodesic γ from p to q . Consider the variation σ of γ defined exactly as in Example 5.66. The calculation from this example demonstrates that

$$\frac{d^2}{ds^2} \Big|_{s=0} \text{length}(\gamma_s) < \frac{\pi^2 - l^2}{2l} < 0,$$

contradicting the fact that γ is minimizing.

The case of arbitrary δ can be handled by a straightforward modification of the calculation in Example 5.66, or by rescaling to achieve $\delta = 1$. \square

EXERCISES

EXERCISE 5.72. Let $\gamma : [a, b] \rightarrow S$ be a unit-speed curve in a regular surface S . If w is a vector field along γ , prove that there exists a variation of γ whose variational vector field equals w . If $w(a) = w(b) = \mathbf{0}$, prove that the variation can be chosen to be proper.

EXERCISE 5.73. When $l = \pi$, show that the vector field w defined in Example 5.66 is the variational vector field of a variation whose curves are all geodesics.

EXERCISE 5.74. Generalize Proposition 5.65 to variations that are not necessarily proper.

EXERCISE 5.75. Show that the formula from Proposition 5.65 is equivalent to

$$\frac{d^2}{ds^2} \Big|_{s=0} \text{length}(\gamma_s) = - \int_a^b \langle w^{\theta\theta}(t) + K(\gamma(t))w(t), w(t) \rangle dt.$$

9. Jacobi Fields (Optional)

This section explores how nearby geodesics spread apart from a single geodesic along its entire length. The question is best framed in terms of *Jacobi fields*. Building on the definition of a *variation* (on page 303), we make the following definition:

DEFINITION 5.68.

Let S be a regular surface and $\gamma : I \rightarrow S$ a geodesic.

- (1) A variation of γ is called a **geodesic variation** if every curve of the variation is a geodesic.
- (2) A **Jacobi field** along γ means the variational vector field of a geodesic variation of γ .
- (3) A Jacobi field, J , is called **tangent** (respectively **orthogonal**) if $J(t)$ is parallel (respectively orthogonal) to $\gamma'(t)$ for all $t \in I$.

In Sect. 3 of this chapter, we already studied the following important type of Jacobi field:

EXAMPLE 5.69 (Jacobi fields from normal polar coordinates). Let $\sigma : (0, \epsilon) \times (0, 2\pi) \rightarrow \mathcal{O}_\epsilon(p) \subset S$ be the surface patch for normal polar coordinates at a point p of a regular surface S . For any fixed $\theta_0 \in (0, 2\pi)$, the curve $\gamma : (0, \epsilon) \rightarrow S$ defined as $\gamma(t) = \sigma(t, \theta_0)$ is a unit-speed geodesic. The function $(s, t) \mapsto \sigma(t, \theta_0 + s)$ is a geodesic variation of γ whose variational vector field is

$$J(t) = \sigma_\theta(t, \theta_0).$$

This Jacobi field is just the coordinate field σ_θ along γ , as illustrated in Fig. 5.26. According to Gauss's lemma (on page 260), it is an orthogonal Jacobi field, so we can express it as $J(t) = g(t) \cdot R_{90}(\gamma'(t))$, where $g(t) = |J(t)|$ and R_{90} is with respect to the orientation induced by σ . Theorem 5.26 (on page 269) says that for all $t \in (0, \epsilon)$,

$$(5.17) \quad [g''(t) = -K(\gamma(t)) \cdot g(t)].$$

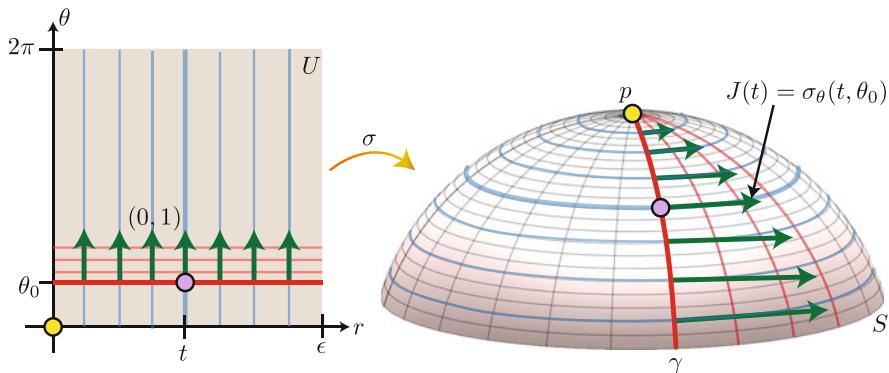


FIGURE 5.26. σ_θ is a Jacobi field along the radial geodesic γ

As we discussed in Sect. 3, Eq. 5.17 says that K controls how nearby geodesics spread apart from γ . Unfortunately, this formula applies only within a normal neighborhood (for $t \in (0, \epsilon)$), even though all of the geodesics of this variation continue on for values $t > \epsilon$. In fact, they continue on forever if S is complete, determining a Jacobi field that continues on forever. One goal of this section is to prove that Eq. 5.17 remains true for as long as the geodesics are defined. Furthermore, Eq. 5.17 will turn out to be the key to understanding *all* Jacobi fields, not just the type described in this example. To broaden our view, let's explore other types of Jacobi fields, beginning with the dullest type:

EXAMPLE 5.70 (Tangent Jacobi fields). For any geodesic $\gamma : [a, b] \rightarrow S$ in a regular surface S , there are two natural tangent Jacobi fields along γ , namely:

- (1) $J(t) = \gamma'(t)$, from the geodesic variation $\sigma(s, t) = \gamma(s + t)$.
- (2) $J(t) = (t - a)\gamma'(t)$, from the geodesic variation $\sigma(s, t) = \gamma(t + s(t - a))$.

The first satisfies $J^0(a) = \mathbf{0}$, while the second satisfies $J(a) = \mathbf{0}$. Unlike the orthogonal Jacobi field from the previous example, these tangent Jacobi fields tell us nothing about γ or S , since they work the same for all geodesics in all surfaces. We will soon learn that an arbitrary tangent Jacobi field along γ must be a linear combination of these two examples, and hence encodes no geometric information.

The previous two examples are not the only types of Jacobi fields:

LEMMA 5.71 (Jacobi fields exist with arbitrary initial conditions).

Let $\gamma : [a, b] \rightarrow S$ be a geodesic in a regular surface S . For all $v_1, v_2 \in T_{\gamma(a)}S$, there exists a unique Jacobi field, J , along γ with $J(a) = v_1$ and $J^0(a) = v_2$.

This lemma reveals how many Jacobi fields there are along γ : there is exactly one for each pair $v_1, v_2 \in T_{\gamma(a)}S$. For now, we will only prove existence. The proof of uniqueness will be deferred until the last sentence of the proof of the forthcoming Theorem 5.72.

PROOF. Choose any curve $\alpha : (-\epsilon, \epsilon) \rightarrow S$ with $\alpha(0) = \gamma(a)$ and $\alpha'(0) = v_1$. Next choose a vector field u along α with $u(0) = \gamma'(a)$ and $u^0(0) = \frac{D}{ds}|_{s=0} u(s) = v_2$. Finally, define

$$\sigma(s, t) = \exp_{\alpha(s)}(t \cdot u(s)),$$

as in Fig. 5.27. Since $\sigma(0, t) = \gamma(t)$ is defined for values of t in an open interval containing $[a, b]$, it can be shown that the same is true for nearby curves of the variation. In particular, for sufficiently small s , the curve $t \mapsto \sigma(s, t)$ is defined for all $t \in [a, b]$.

Notice that σ is a geodesic variation of γ . Let $J(t) = \sigma_s(0, t)$ be its variational vector field, which is a Jacobi field by definition. It has the specified initial conditions, because

$$\begin{aligned} J(a) &= \sigma_s(0, a) = \alpha'(0) = v_1, \quad \text{and} \\ J^0(a) &= \underbrace{\frac{D}{dt} \sigma_s(0, a)}_{\text{Equation 5.16, page 306}} = \frac{D}{ds} \sigma_t(0, a) = \frac{D}{ds} \Big|_{s=0} u(s) = v_2. \end{aligned}$$

□

Let's summarize what we expect. Examples 5.69 and 5.70 hint that, among all Jacobi fields along γ , the *orthogonal* Jacobi fields are the only ones that encode geometric information about γ and S . Lemma 5.71 says that Jacobi fields are arbitrary enough to mix things up—a single Jacobi field can have both tangent and orthogonal components. It seems necessary

to understand this mixed-up-ness of the family of *all* Jacobi fields along γ in order to prove that the orthogonal ones are the only ones that deserve to be studied. The following theorem pulls it all together:

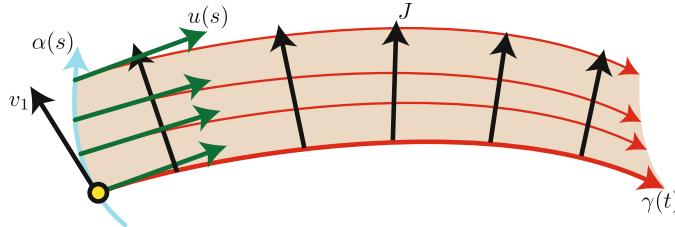


FIGURE 5.27. The construction of a Jacobi field with arbitrary initial conditions

THEOREM 5.72 (Analytic characterization of Jacobi fields).

Let S be a regular surface and $\gamma : [a, b] \rightarrow S$ a geodesic with speed λ . Let w be a unit-length parallel vector field along γ that is everywhere orthogonal to γ' . The following are equivalent properties for a vector field J along γ :

- (1) J is a Jacobi field.
- (2) J has the form

$$J(t) = \underbrace{(c_1 + c_2 t)\gamma'(t)}_{\text{tangent Jacobi field}} + \underbrace{g(t) \cdot w(t)}_{\text{orthogonal Jacobi field}}$$

for some $c_1, c_2 \in \mathbb{R}$ and some smooth function $g : [a, b] \rightarrow \mathbb{R}$ such that for all $t \in [a, b]$,

$$g''(t) = -\lambda^2 \cdot K(\gamma(t)) \cdot g(t).$$

- (3) J satisfies

$$\boxed{\text{the Jacobi equation: } J^{\partial\partial}(t) = -\lambda^2 \cdot K(\gamma(t)) \cdot J^\perp(t)}$$

for all $t \in [a, b]$, where $J^\perp(t)$ denotes the component of $J(t)$ orthogonal to $\gamma'(t)$.

If γ is of unit speed ($\lambda = 1$), then the boxed equation in (2) is exactly Eq. 5.17.

PROOF. $\boxed{(2) \implies (3)}$ Since both γ' and w are parallel along γ , we have

$$\begin{aligned} J^{\partial\partial}(t) &= 0 + g''(t)w(t) = -\lambda^2 \cdot K(\gamma(t)) \cdot g(t) \cdot w(t) \\ &= -\lambda^2 \cdot K(\gamma(t)) \cdot J^\perp(t). \end{aligned}$$

$\boxed{(3) \implies (2)}$ Express $J(t) = f(t) \cdot \gamma'(t) + g(t) \cdot w(t)$ for some smooth functions $f, g : [a, b] \rightarrow S$. Again using the fact that both γ' and w are

parallel along γ , we have

$$\begin{aligned} f''(t) \cdot \gamma'(t) + g''(t) \cdot w(t) &= J^{\theta\theta}(t) = -\lambda^2 \cdot K(\gamma(t)) \cdot J^\perp(t) \\ &= -\lambda^2 \cdot K(\gamma(t)) \cdot g(t) \cdot w(t). \end{aligned}$$

This implies that $f'' = 0$ (so f is a linear function) and $g''(t) = -\lambda^2 \cdot K(\gamma(t)) \cdot g(t)$.

(1) \implies (3) Suppose that J is a Jacobi field, which means it is the variational vector field of a geodesic variation σ of γ . As discussed in the previous section, even though σ is not necessarily a surface patch, Proposition 5.56 (on page 293, subbing $u \leftrightarrow t$ and $v \leftrightarrow s$) and Eq. 5.16 (on page 306) are valid for σ , provided that “ R_{90} ” is interpreted with respect to the local orientation $N = \frac{\sigma_t \times \sigma_s}{|\sigma_t \times \sigma_s|}$. Observe that $\frac{D}{dt}\sigma_t = \mathbf{0}$, because the curves of the variation are geodesics. Therefore,

$$\frac{D}{dt} \frac{D}{dt} \sigma_s = \frac{D}{dt} \frac{D}{ds} \sigma_t = \frac{D}{ds} \underbrace{\frac{D}{dt} \sigma_t}_0 - K \cdot |\sigma_s \times \sigma_t| \cdot R_{90}(\sigma_t) = -K \cdot |\sigma_s \times \sigma_t| \cdot R_{90}(\sigma_t),$$

where R_{90} is counterclockwise with respect to the orientation induced by σ , so that

$$\frac{J^\perp(t)}{|J^\perp(t)|} = R_{90} \left(\frac{\gamma'(t)}{\lambda} \right).$$

This orientation is not defined where $|\sigma_s \times \sigma_t| = 0$, but it doesn't matter, because the above expression vanishes at such points. At $s = 0$ and arbitrary t , this equation says that

$$\begin{aligned} J^{\theta\theta}(t) &= -K(\gamma(t)) \cdot |J(t) \times \gamma'(t)| \cdot R_{90}(\gamma'(t)) \\ &= -K(\gamma(t)) \cdot |\lambda \cdot J^\perp(t)| \cdot \lambda \cdot R_{90} \left(\frac{\gamma'(t)}{\lambda} \right) = -\lambda^2 \cdot K(\gamma(t)) \cdot J^\perp(t). \end{aligned}$$

(3) \implies (1) Suppose that J satisfies the Jacobi equation. By Lemma 5.71, there exists a Jacobi field \tilde{J} that has the same initial data as J ; that is, $\tilde{J}(a) = J(a)$ and $\tilde{J}^\theta(a) = J^\theta(a)$. According to the already completed portions of the current proof, both J and \tilde{J} satisfy property (2); that is, we can write

$$\begin{aligned} J(t) &= (c_1 + c_2 t)\gamma'(t) + g(t) \cdot w(t), \\ \tilde{J}(t) &= (\tilde{c}_1 + \tilde{c}_2 t)\gamma'(t) + \tilde{g}(t) \cdot w(t), \end{aligned}$$

for some $c_1, c_2, \tilde{c}_1, \tilde{c}_2 \in \mathbb{R}$ and some $g, \tilde{g} : [a, b] \rightarrow \mathbb{R}$ that both satisfy the boxed equation in part (2) of the theorem. Since the $J(a)$ and $\tilde{J}(a)$ have the same projections onto $\gamma'(a)$, we have $\tilde{c}_1 = c_1$. Since the $J^\theta(a)$ and $\tilde{J}^\theta(a)$ have the same projections onto $\gamma'(a)$, we have $\tilde{c}_2 = c_2$. By similarly considering the projections onto $w(t)$, we learn that $\tilde{g}(0) = g(0)$ and $\tilde{g}'(0) = g'(0)$. In summary, \tilde{g} and g have the same initial data and satisfy the same differential equation, so they must be equal. This observation also justifies the uniqueness claim of Lemma 5.71. \square

By characterizing Jacobi fields as solutions to an analytic equation, Theorem 5.72 implies certain algebraic and linearity properties that are not at all obvious from our original variational definition of Jacobi fields. For example:

COROLLARY 5.73 (Algebraic properties of Jacobi fields).

Let $\gamma : [a, b] \rightarrow S$ be a geodesic in a regular surface S . Let J , J_1 , and J_2 be Jacobi fields along γ .

- (1) $c_1 J_1 + c_2 J_2$ is a Jacobi field along γ for every $c_1, c_2 \in \mathbb{R}$.
- (2) $\langle J_1^\theta(t), J_2(t) \rangle - \langle J_1(t), J_2^\theta(t) \rangle$ is constant on $[a, b]$.
- (3) If there exists a single $t_0 \in [a, b]$ such that $J(t_0)$ and $J^\theta(t_0)$ are both orthogonal to $\gamma'(t_0)$, then J is an orthogonal Jacobi field.
- (4) If J vanishes at two distinct times, then J is an orthogonal Jacobi field.

PROOF. Exercise 5.82. □

EXAMPLE 5.74 (A Jacobi field on S^2). Let $p \in S^2$ and let $v \in T_p S^2$ be of unit length. As discussed in Example 5.6 on page 249, the geodesic $\gamma : \mathbb{R} \rightarrow S^2$ through p in the direction v is described as

$$\gamma(t) = (\cos t)p + (\sin t)v.$$

According to Theorem 5.72, the vector field $J(t) = g(t) \cdot R_{90}(\gamma'(t))$ along γ is a Jacobi field if and only if $g''(t) = -g(t)$ for all $t \in \mathbb{R}$. The solution $g(t) = \sin(t)$ is illustrated in Fig. 5.28.

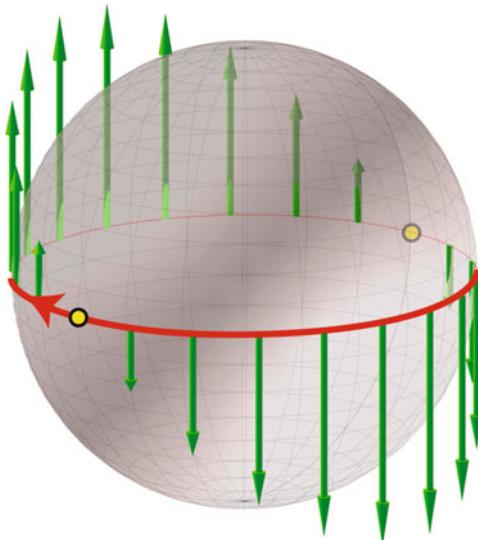


FIGURE 5.28. A Jacobi field on S^2

In Fig. 5.28, the pair of yellow points are called *conjugate* along the red geodesic:

DEFINITION 5.75.

Let S be a regular surface and $\gamma : [a, b] \rightarrow S$ a geodesic. For $t_0 \in (a, b]$, the point $q = \gamma(t_0)$ is called a **conjugate point** to $p = \gamma(a)$ along γ if there exists a Jacobi field J along γ (not everywhere zero) with $J(a) = J(t_0) = \mathbf{0}$.

In Sect. 3 of this chapter, we measured the initial rate of separation of two friends walking away from p along different geodesics. If one friend walks along γ and the other along a nearby geodesic at the same speed, then derivative information predicts that they will meet again at the first conjugate point to p along γ . Thus, conjugate points are where friends reunite. But there are no reunions when the Gaussian curvature is nonpositive:

PROPOSITION 5.76.

A geodesic in a regular surface S with $K \leq 0$ has no conjugate points.

PROOF. Assume to the contrary that there exist a geodesic $\gamma : [a, b] \rightarrow S$ and a Jacobi field J along γ (not everywhere zero) such that $J(a) = J(t_0) = \mathbf{0}$ for some $t_0 \in (a, b]$. By Corollary 5.73(4), J is orthogonal. Express J as

$$J(t) = g(t) \cdot w(t),$$

as in Theorem 5.72(2), with $g(a) = g(t_0) = 0$. We know that $g'(a) \neq 0$, for otherwise J would have the same initial conditions as (and hence be equal to) the everywhere-zero Jacobi field along γ . By redefining w to be its negative if necessary, we can ensure that $g'(a) > 0$.

In summary, we have $g(a) = g(t_0) = 0$, $g'(a) > 0$, and

$$g''(t) = \underbrace{-\lambda^2 \cdot K(\gamma(t)) \cdot g(t)}_{\geq 0}$$

for all $t \in [a, b]$. But no smooth function g could have this collection of properties. To understand why, we can assume without loss of generality that t_0 is the *first* root of g to the right of a , so on $[a, t_0]$ we know that $g(t) \geq 0$ and therefore that $g'' \geq 0$. In other words, g' is increasing and hence positive on $[a, t_0]$, contradicting the fact that $g(a) = g(t_0) = 0$. \square

One fundamental application of Jacobi fields and conjugate points is to increase our understanding of the derivative of the exponential map. For this, we will think of $T_p S$ as a surface and $\exp_p : T_p S \rightarrow S$ as a smooth function between surfaces. Since $T_p S$ is a very special type of surface (a plane), it equals each of its tangent planes. In other words, for every $v \in T_p S$, $T_v(T_p S) = T_p S$. This is the domain of $d(\exp_p)_v$ (the derivative of \exp_p at v).

PROPOSITION 5.77 (Jacobi fields describe the derivative of the exponential map).

Let S be a regular surface, $p \in S$, and $v \in T_p S$ such that $\exp_p(v)$ is defined. Let $w \in T_p S = T_v(T_p S)$. Then

$$d(\exp_p)_v(w) = J(1),$$

where J is the unique Jacobi field along $\gamma(t) = \exp_p(tv)$ ($t \in [0, 1]$) with initial conditions $J(0) = 0$ and $J'(0) = w$. Further, if $w \perp v$, then J is an orthogonal Jacobi field.

PROOF. Consider the following geodesic variation of γ :

$$\sigma(s, t) = \exp_p(t(v + sw)).$$

In Fig. 5.29, the curves of this variation are illustrated in orange. Since $\exp_p(v)$ is defined, it can be shown that σ is defined on the domain $(-\epsilon, \epsilon) \times [0, 1]$ for sufficiently small ϵ .

Let J be the variational vector field of σ , which is a Jacobi field. We claim that

$$(5.18) \quad J(t) = d(\exp_p)_{tv}(tw)$$

for all $t \in [0, 1]$. To justify this, notice that the s -parameter curve of σ corresponding to a fixed value $t_0 \in [0, 1]$ can be described as

$$\beta(s) = \sigma(s, t_0) = \exp_p(t_0(v + sw)) = \exp_p\left(\tilde{\beta}(s)\right),$$

where $\tilde{\beta}(s) = t_0(v + sw)$ is the straight line in $T_p S$ passing through $\tilde{\beta}(0) = t_0 v$ in the direction $\tilde{\beta}'(0) = t_0 w$. Thus,

$$J(t_0) = \beta'(0) = d(\exp_p)_{\tilde{\beta}(0)}\left(\tilde{\beta}'(0)\right) = d(\exp_p)_{t_0 v}(t_0 w),$$

which proves Eq. 5.18.

Clearly $J(0) = \mathbf{0}$. To compute the initial derivative, rewrite Eq. 5.18 as $J(t) = t(d(\exp_p)_{tv}(w))$ and apply Lemma 5.39(2) on page 280:

$$J'(0) = 0 + \underbrace{1 \cdot (d(\exp_p)_{0v}(w))}_{\text{because } d(\exp_p)_0 = \text{identity}} = w.$$

If $w \perp v$, then Lemma 5.73(3) implies that J is an orthogonal Jacobi field. \square

Gauss's lemma (on page 260) promises that within a small ball about $\mathbf{0}$ in $T_p S$, the derivative of $\exp_p : T_p S \rightarrow S$ preserves the norm of the radial σ_r -vector and also preserves the orthogonality between σ_r and the angular σ_θ -vector. We can now prove that this is true not only within a small ball, but everywhere that \exp_p is defined:

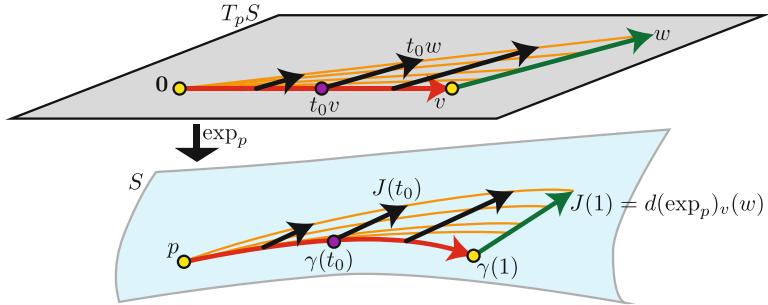


FIGURE 5.29. $d(\exp_p)_v(w) = J(1)$, where $J(0) = \mathbf{0}$ and $J^\theta(0) = w$

COROLLARY 5.78 (The generalized Gauss's Lemma).

With the assumptions of Proposition 5.77, the linear transformation $d(\exp_p)_v$ acts as follows:

- (1) $|d(\exp_p)_v(v)| = |v|$ (it is length-preserving in the v -direction), so its rank is at least 1.
- (2) If $w \perp v$, then $(d(\exp_p)_v(w)) \perp (d(\exp_p)_v(v))$; that is, it preserves orthogonality with the v -direction.
- (3) It is singular (has rank 1) if and only if $\exp_p(v)$ is conjugate to p along γ .

PROOF. For part (1), observe that $|v|$ is the speed of γ , and $d(\exp_p)_v(v) = \gamma'(1)$. Part (2) comes from the fact that $d(\exp_p)_v(w) = J(1)$, where J is an orthogonal Jacobi field.

For (3), first suppose that $\exp_p(v)$ is conjugate to p along γ , which means that there is a Jacobi field J along γ with $J(0) = J(1) = \mathbf{0}$. Setting $w = J^\theta(0)$, we have $\mathbf{0} = J(1) = d(\exp_p)_v(w)$, so the linear transformation is singular. The other direction is similar. \square

COROLLARY 5.79.

If S is a complete surface with $K \leq 0$, then for every $p \in S$, $\exp_p : T_p S \rightarrow S$ is a local diffeomorphism; that is, every $v \in T_p S$ has a neighborhood restricted to which \exp_p is a diffeomorphism onto its image.

PROOF. Let $p \in S$. By Proposition 5.76, geodesics through p have no conjugate points. According to Corollary 5.78(3), this implies that the derivative of \exp_p at every $v \in T_p S$ is nonsingular. The result now follows from the inverse function theorem for surfaces (on page 143). \square

This corollary does *not* say that \exp_p is a diffeomorphism. For example, the cylinder from Example 5.7 (on page 250) has $K = 0$, yet the exponential map is not one-to-one at any point.

EXERCISES

EXERCISE 5.76. Precisely formulate and prove the assertion that Jacobi fields are intrinsic.

EXERCISE 5.77. Prove that on a surface of revolution (Example 3.25 on page 131), every longitudinal curve is a geodesic with no conjugate points.

EXERCISE 5.78. Explicitly describe a geodesic variation for the Jacobi field in Example 5.74.

EXERCISE 5.79. Give an example of a periodic geodesic $\gamma : \mathbb{R} \rightarrow C$ in the cylinder C from Example 5.7 (on page 250) and a Jacobi field along it that is not periodic.

EXERCISE 5.80. Let $\gamma : [a, b] \rightarrow S$ be a geodesic in a regular surface S . Under what conditions does there exist an orthogonal Jacobi field along γ (not everywhere zero) that is parallel? Give examples.

EXERCISE 5.81. Prove that the following is an equivalent formulation of the Jacobi equation:

$$J^{\infty\infty}(t) = -K(\gamma(t)) \cdot (\gamma'(t) \times J(t)) \times \gamma'(t).$$

EXERCISE 5.82. Prove Corollary 5.73.

EXERCISE 5.83. Let $\gamma : [a, b] \rightarrow S$ be a geodesic in a regular surface S . Assume that $\gamma(b)$ is not conjugate to $\gamma(a)$ along γ . Prove that for every $v_a \in T_{\gamma(a)}S$ and $v_b \in T_{\gamma(b)}S$, there exists a unique Jacobi field J along γ with $J(a) = v_a$ and $J(b) = v_b$.

EXERCISE 5.84. Let S be a regular surface, $p \in S$, and $v \in T_p S$ with $\exp_p(v)$ defined. Define $\gamma : [0, 1] \rightarrow S$ as $\gamma(t) = \exp_p(tv)$.

- (1) Let $\beta : [0, 1] \rightarrow T_p S$ be any regular curve with $\beta(0) = \mathbf{0}$ and $\beta(1) = v$. If \exp_p is defined at all points of the trace of β , prove that

$$\text{length}(\exp_p \circ \beta) \geq \text{length}(\gamma).$$

- (2) (**Jacobi's Theorem**). Assume that $\gamma(t)$ is *not* a conjugate point to p along γ for all $t \in (0, 1]$. Prove that for every proper variation $\sigma : (-\epsilon, \epsilon) \times [0, 1] \rightarrow S$ of γ , there exists $\delta \in (0, \epsilon)$ such that each curve of the variation of the form $t \mapsto \sigma(s_0, t)$ with $|s_0| < \delta$ has length greater than or equal to the length of γ . In other words, γ is length-minimizing among the sufficiently nearby curves of every variation. *HINT: Prove that each sufficiently nearby curve of the variation can be expressed as $\exp_p \circ \beta$ as in part (1).*



CHAPTER

6



Carl Friedrich Gauss (1777–1855) was one of the founders of differential geometry. It is appropriate that his name, which already appeared prominently throughout the last two chapters, heads the title of this final chapter. His image is shown here on the German 10-mark note.

The Gauss–Bonnet Theorem

The Gauss–Bonnet theorem is the most famous result in the study of surfaces. It provides a satisfying final chapter of this textbook because it interrelates many fundamental concepts from the previous five chapters.

In Sect. 1, we will discuss the local version of the Gauss–Bonnet theorem, which is a generalization to surfaces of Hopf's Umlaufsatz (Theorem 2.7 on page 69). This generalization introduces a new term to the equation: the integral of the Gaussian curvature over the interior of the closed curve. The theorem will confirm our guess about how the holonomy around a closed curve is related to the Gaussian curvature enclosed inside it.

In Sect. 2, the global version of the Gauss–Bonnet theorem is stated and proven by stitching together sufficiently many applications of the local version. The result is mathematics at its best—simple, beautiful, powerful, with many remarkable applications! Let's begin.

1. The Local Gauss–Bonnet Theorem

This is a good time to reread Theorem 2.7 (on page 2.7), which we intend to generalize to surfaces. It is about a “positively oriented piecewise-regular simple closed curve in \mathbb{R}^2 . ” We know from Sect. 7 of Chap. 5 how most of these adjectives apply to curves in arbitrary oriented surfaces, but what about “positively oriented”?

To decide this, let S be an oriented surface. Let $\gamma : [a, b] \rightarrow S$ be a piecewise-regular simple closed curve in S with signed angles denoted by $\{\alpha_i\}$ (as defined in Sect. 7 of Chap. 5). In the case $S = \mathbb{R}^2$, we decided in Sect. 1 of Chap. 2 to call γ *positively oriented* if its interior is on one’s left as one traverses γ . But for arbitrary S , the term “interior” doesn’t necessarily make sense. The Jordan curve theorem is *not* valid in this setting. Who is to say which side of a curve deserves to be called the interior? For example, if γ traverses the equator of a sphere, then whether γ is positively or negatively oriented would depend on whether the southern or the northern hemisphere is regarded as its “interior” (and also would depend on the choice of orientation for the sphere). Thus, the adjective “positively oriented” makes sense only with respect to an orientation of S and a decision about which region is considered the interior. The story must begin with a region:

DEFINITION 6.1.

Let S be an oriented surface.

- (1) A subset $R \subset S$ is called a **region** of S if it equals the union of an open set in S with the boundary of that open set.
- (2) A **regular region** in S means a compact region $R \subset S$ whose boundary (denoted by ∂R) equals the union of the traces of finitely many nonintersecting piecewise-regular simple closed curves in S . Each individual trace is called a **boundary component** of R .
- (3) Let R be a regular region in S . A parametrization, $\gamma : [a, b] \rightarrow R$, of a boundary component of R is called **positively oriented** if R is on one’s left as one traverses γ ; more precisely, $R_{90}(\gamma'(t))$ points into R for all times $t \in [a, b]$ that don’t correspond to vertices.

A “**polygonal region**” (Definition 3.55 on page 161) is the same as a “regular region with one boundary component that is covered by a single surface patch.” In Fig. 6.1, the left image shows a polygonal region, while the right image shows a regular region with four boundary components. The boundary components of both regions are shown positively oriented.

The *global* Gauss–Bonnet theorem (in the next section) applies to complicated regular regions like what is shown in the right figure, but the *local* version applies only to polygonal regions like what is shown in the left figure:

THEOREM 6.2 (The Local Gauss–Bonnet Theorem).

Let S be an oriented regular surface and $R \subset S$ a polygonal region. Let $\gamma : [a, b] \rightarrow R$ be a unit-speed positively oriented parametrization of ∂R , with signed angles denoted by $\{\alpha_i\}$. Then

$$\underbrace{\int_a^b \kappa_g(t) dt + \sum_i \alpha_i}_{\text{angle displacement around } \gamma} = 2\pi - \iint_R K dA.$$

As before, “ $\int_a^b \kappa_g(t) dt$ ” is shorthand for $\sum_i \left(\int_{t_i}^{t_{i+1}} \kappa_g(t) dt \right)$, which means the sum of the integrals of the geodesic curvature over the smooth segments of γ . When γ is smooth (no vertices), this is called the **total geodesic curvature** of γ .

We previously confirmed some special cases of the theorem. First, if S is the xy -plane with the orientation $N = (0, 0, 1)$, then $\iint_R K dA = 0$ and $\kappa_g = \kappa_s$, so the conclusion reduces to that of Theorem 2.7 (on page 69).

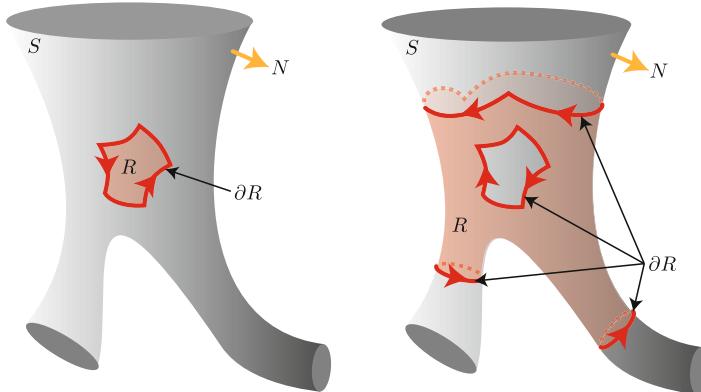


FIGURE 6.1. Regular regions with positively oriented boundary components

Second, when γ is a latitude of S^2 , the theorem was verified in Exercise 5.52 on page 288. Third, the theorem was confirmed in Exercise 5.68 on page 296 for a geodesic triangle that is small enough to lie within a normal ball of one of its vertices.

As we discussed in Sect. 7 of Chap. 5, $\Delta\theta = \int_a^b \kappa_g(t) dt + \sum_i \alpha_i$ represents the angle displacement around γ , which means the net change in an angle function for γ . The holonomy around γ equals a clockwise rotation by this amount. In Example 5.63 on page 305, we visualized this by imagining that a south-pointing chariot is driven along γ , so that $\Delta\theta$ represents the total clockwise rotation of the statue as seen by the charioteer. The local Gauss–Bonnet theorem says that this angle displacement differs from 2π by the

integral of the Gaussian curvature over the interior of γ . Since a clockwise rotation by $2\pi - \iint_R K dA$ is the same as a *counterclockwise* rotation by $\iint_R K dA$, this can be rephrased thus:

COROLLARY 6.3.

With the notation and assumptions of Theorem 6.2, the holonomy around γ equals a counterclockwise rotation of $T_{\gamma(a)}S$ by the angle $\iint_R K dA$.

PROOF OF THEOREM 6.2. Let $\sigma : U \subset \mathbb{R}^2 \rightarrow V \subset S$ be a surface patch with $R \subset V$. By the definition of a *polygonal region*, $\tilde{\gamma}(t) = \sigma^{-1}(\gamma(t))$, $t \in [a, b]$, is a piecewise-regular simple closed curve in U that bounds $\tilde{R} = \sigma^{-1}(R)$. Assume that σ is compatible with the orientation of S , so that $\tilde{\gamma}$ is positively oriented. Express $\tilde{\gamma}$ in coordinates as $\tilde{\gamma}(t) = (u(t), v(t))$. Let \tilde{C} denote the trace of $\tilde{\gamma}$.

For every $q \in U$, define

$$X(q) = \frac{\sigma_u(q)}{|\sigma_u(q)|}, \quad Y(q) = R_{90}(X(q)).$$

Notice that X and Y determine vector fields on V such that for each $q \in U$, $\{X(q), Y(q)\}$ is a positively oriented orthonormal ordered basis of $T_{\sigma(q)}S$. In particular, the function $N : U \rightarrow \mathbb{R}^3$ defined as $N(q) = X(q) \times Y(q)$ equals the given orientation expressed in local coordinates. Since $\{X, Y, N\}$ is everywhere orthonormal, Proposition 1.17 (on page 13) yields a collection of derivative relationships:

$$(6.1) \quad \begin{aligned} \langle X_u, Y \rangle &= -\langle X, Y_u \rangle, & \langle X_u, N \rangle &= -\langle X, N_u \rangle, & \langle Y_u, N \rangle &= -\langle Y, N_u \rangle, \\ \langle X_v, Y \rangle &= -\langle X, Y_v \rangle, & \langle X_v, N \rangle &= -\langle X, N_v \rangle, & \langle Y_v, N \rangle &= -\langle Y, N_v \rangle, \\ \langle X_u, X \rangle &= \langle X_v, X \rangle = \langle Y_u, Y \rangle = \langle Y_v, Y \rangle = \langle N_u, N \rangle = \langle N_v, N \rangle = 0. \end{aligned}$$

Let $I = [t_i, t_{i+1}]$ be one of the subintervals on which γ is regular. For $t \in I$, define $X(t) = X(\tilde{\gamma}(t))$ and $Y(t) = Y(\tilde{\gamma}(t))$. As in the proof of Proposition 2.3 (on page 63), there exists a smooth function $\theta : I \rightarrow \mathbb{R}$ such that

$$\gamma'(t) = (\cos \theta(t)) \cdot X(t) + (\sin \theta(t)) \cdot Y(t)$$

for all $t \in I$. Notice that θ measures the angle that γ' makes with the positive u -coordinate direction.

If X happens to be parallel along γ , then θ agrees with the angle function that was defined in Proposition 5.42 (on page 283), so $\kappa_g = \theta'$. If X is not parallel along γ , then one might expect κ_g to equal θ' plus a term that

reflects the failure of X to be parallel along γ . That is exactly what happens. Suppressing the input variable, we have

$$\gamma'' = \underbrace{-(\theta' \sin \theta) X + (\theta' \cos \theta) Y}_{\theta' \cdot R_{90}(\gamma')} + \underbrace{(\cos \theta) X' + (\sin \theta) Y'}_{R_{90}(\gamma')}$$

Therefore,

$$\begin{aligned}\kappa_g &= \langle \gamma'', R_{90}(\gamma') \rangle \\ &= \langle \theta' \cdot R_{90}(\gamma') + (\cos \theta) X' + (\sin \theta) Y', R_{90}(\gamma') \rangle \\ &= \langle \theta' \cdot R_{90}(\gamma'), R_{90}(\gamma') \rangle + \left\langle (\cos \theta) X' + (\sin \theta) Y', \underbrace{-(\sin \theta) X + (\cos \theta) Y}_{R_{90}(\gamma')} \right\rangle \\ &= \theta' + \langle X', Y \rangle \quad (\text{using Equation 6.1}).\end{aligned}$$

So to prove the theorem, it will suffice to confirm that

$$(1) \int_a^b \theta'(t) dt + \sum_i \alpha_i = 2\pi, \quad \text{and}$$

$$(2) \int_a^b \langle X'(t), Y(t) \rangle dt = - \iint_R K dA,$$

where both integrals are shorthand for the sum of the corresponding integrals over the smooth segments of γ .

Claim (1) is perhaps not surprising. In the special case that σ is conformal¹ (angle-preserving), claim (1) follows immediately by applying Theorem 2.7 to $\tilde{\gamma}$. To handle the nonconformal case, we must appeal to the proof (rather than just the statement) of Theorem 2.7. Specifically, we first smooth the corners of γ to reduce to the case that γ is smooth, in which case θ is a single smooth function on all of $[a, b]$, and

$$\int_a^b \theta'(t) dt = \theta(b) - \theta(a) = 2\pi \cdot (\text{degree of } t \mapsto (\cos \theta(t), \sin \theta(t))).$$

The degree of the map $t \mapsto \frac{\tilde{\gamma}'(t)}{|\tilde{\gamma}'(t)|}$ is the rotation index of $\tilde{\gamma}$, which equals 1 by Hopf's Umlaufsatz. So to prove claim (1), we must confirm that the maps $t \mapsto \frac{\tilde{\gamma}'(t)}{|\tilde{\gamma}'(t)|}$ and $t \mapsto (\cos \theta(t), \sin \theta(t))$ have the same degree. For this, Exercise 3.58 on page 151 (applied with $g = d\sigma_{\tilde{\gamma}(t)}$ and $v = \tilde{\gamma}'(t)$) implies that these two maps never have outputs in opposite directions, which by Lemma 2.4 (on page 63) suffices for us to conclude that they have the same degree.

For claim (2), we will apply Green's theorem (on page 91) along $\tilde{\gamma}$ to the vector field $\mathcal{F} : U \rightarrow \mathbb{R}^2$ defined as

$$\mathcal{F} = (\underbrace{\langle X_u, Y \rangle}_P, \underbrace{\langle X_v, Y \rangle}_Q).$$

¹This special case is not as restrictive as it might seem; it can be proven that every regular surface admits an atlas of conformal surface patches.

Since $X' = u'X_u + v'X_v$, this yields

$$\begin{aligned} \int_a^b \langle X'(t), Y(t) \rangle dt &= \oint_{\tilde{C}} \mathcal{F} \cdot d\tilde{\gamma} = \iint_{\tilde{R}} (Q_u - P_v) du dv \\ &= \iint_{\tilde{R}} (\langle X_v, Y \rangle_u - \langle X_u, Y \rangle_v) du dv \\ &= \iint_{\tilde{R}} (\langle X_v, Y_u \rangle - \langle X_u, Y_v \rangle) du dv \\ &= \iint_{\tilde{R}} \frac{\langle X_v, Y_u \rangle - \langle X_u, Y_v \rangle}{\|d\sigma\|} \|d\sigma\| du dv \\ &= \iint_{\tilde{R}} -K \|d\sigma\| du dv = - \iint_R K dA. \end{aligned}$$

To finish the proof, we must prove that the above two blue expressions are equal at an arbitrary point of U . For this, let $q \in U$ be arbitrary, and define $p = \sigma(q)$. The Gaussian curvature at p equals the determinant of the matrix representing the Weingarten map with respect to any basis of $T_p S$. We will use the basis $\{X, Y\}$, but it will be necessary to convert between this basis and the basis $\{\sigma_u, \sigma_v\}$. Define the conversion matrix $M = \begin{pmatrix} a & c \\ b & d \end{pmatrix}$ such that

$$X = a \cdot \sigma_u + b \cdot \sigma_v, \quad Y = c \cdot \sigma_u + d \cdot \sigma_v$$

at q . Notice that

$$\|d\sigma_q\| = \det(M^{-1}) = \frac{1}{\det M} = \frac{1}{ad - bc}.$$

The Gaussian curvature of S at p is

$$\begin{aligned} K(p) &= \det \begin{pmatrix} \langle dN_p(X), X \rangle & \langle dN_p(Y), X \rangle \\ \langle dN_p(X), Y \rangle & \langle dN_p(Y), Y \rangle \end{pmatrix} \\ &= \det \begin{pmatrix} \langle dN_p(a \cdot \sigma_u + b \cdot \sigma_v), X \rangle & \langle dN_p(c \cdot \sigma_u + d \cdot \sigma_v), X \rangle \\ \langle dN_p(a \cdot \sigma_u + b \cdot \sigma_v), Y \rangle & \langle dN_p(c \cdot \sigma_u + d \cdot \sigma_v), Y \rangle \end{pmatrix} \\ &= \det \left(\begin{pmatrix} \langle N_u, X \rangle & \langle N_v, X \rangle \\ \langle N_u, Y \rangle & \langle N_v, Y \rangle \end{pmatrix} \cdot \begin{pmatrix} a & c \\ b & d \end{pmatrix} \right) \\ &= \frac{1}{\|d\sigma_q\|} (\langle N_u, X \rangle \cdot \langle N_v, Y \rangle - \langle N_u, Y \rangle \cdot \langle N_v, X \rangle) \\ &= \frac{1}{\|d\sigma_q\|} (\langle N, X_u \rangle \cdot \langle N, Y_v \rangle - \langle N, Y_u \rangle \cdot \langle N, X_v \rangle) \quad (\text{by Equation 6.1}) \\ &= \frac{1}{\|d\sigma_q\|} (\langle (X_u)^N, (Y_v)^N \rangle - \langle (Y_u)^N, (X_v)^N \rangle) \\ &= \frac{1}{\|d\sigma_q\|} (\langle X_u, Y_v \rangle - \langle Y_u, X_v \rangle), \end{aligned}$$

where the N -superscripts denote projections in the direction of N . The final equality comes from Eq. 6.1, which implies that the X - and Y -projections cancel out. This completes the proof. \square

It is sometimes convenient to rephrase the local Gauss–Bonnet theorem in terms of *interior* angles. With the notation and assumptions of Theorem 6.2, define the **interior angle** at the i th vertex of γ as

$$(6.2) \quad \beta_i = \pi - \alpha_i.$$

Figure 2.10 (on page 71) illustrates the internal angles of a plane curve. The geometric interpretation is the same for curves in general surfaces. The conclusion of the theorem becomes

$$\boxed{\int_a^b \kappa_g(t) dt = \sum_i \beta_i - (m-2)\pi - \iint_R K dA,}$$

where m is the number of vertices. If the smooth segments of γ are geodesics, this becomes

$$(6.3) \quad \boxed{\sum_i \beta_i = (m-2)\pi + \iint_R K dA,}$$

which generalizes a well-known formula for the sum of the interior angles of a polygon in \mathbb{R}^2 . The case $m = 3$ of this formula was previously discussed in Exercise 5.68(6). In this case, γ is called a **geodesic triangle**, and the formula says that

$$\beta_1 + \beta_2 + \beta_3 = \pi + \iint_R K dA.$$

As illustrated in Fig. 6.2, a geodesic triangle is called “fat” if $\beta_1 + \beta_2 + \beta_3 > \pi$, which occurs in regions of positive curvature, and is called “thin” if $\beta_1 + \beta_2 + \beta_3 < \pi$, which occurs in regions of negative curvature.

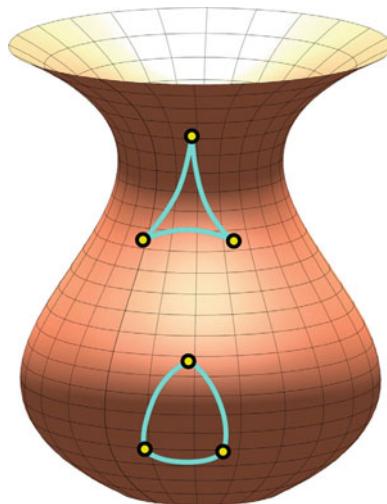


FIGURE 6.2. Geodesic triangles are fat in regions where $K > 0$, and thin in regions where $K < 0$

EXERCISES

EXERCISE 6.1. Use the local Gauss–Bonnet theorem to verify Eq. 5.13 on page 302 (which should be thought of as an infinitesimal version of the theorem).

EXERCISE 6.2. Explicitly verify the Gauss–Bonnet theorem for the regions of S^2 from Exercise 5.70 on page 302.

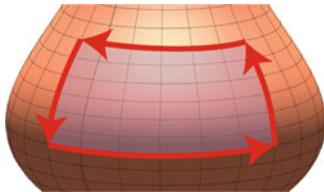


FIGURE 6.3. A parameter rectangle on a surface of revolution

EXERCISE 6.3. Let S be a surface of revolution (Example 3.25 on page 131). Let R be an arbitrary “parameter rectangle,” which means the region bounded by two fixed latitudes and two fixed longitudes (as in Fig. 6.3). Express $\int \kappa_g(t) dt$ and $\iint_R K dA$ in terms of the components, $x(t)$ and $z(t)$, of the generating curve, and explicitly verify the Gauss–Bonnet theorem for this example.

2. The Global Gauss–Bonnet Theorem

In this section, we generalize the local Gauss–Bonnet theorem to a result that applies to regular regions (rather than just polygonal regions).

More specifically, let S be an oriented surface and $R \subset S$ a regular region with positively oriented boundary components. If R is a polygonal region, recall that the local Gauss–Bonnet theorem says that

$$(6.4) \quad \int \kappa_g(t) dt + \sum \alpha_i = 2\pi - \iint_R K dA.$$

Even if R is not a polygonal region, the terms of this equation still make sense. We will interpret “ $\int \kappa_g(t) dt$ ” as the sum of the integrals of the geodesic curvature over all smooth segments of *all boundary components*. Similarly, we will interpret “ $\sum \alpha_i$ ” as the sum of the signed angles at all vertices of *all boundary components*. Finally, the term $\iint_R K dA$ makes sense via Definition 3.55 (on page 161), provided that R equals the union of finitely many polygonal regions intersecting only along boundaries. We will soon see that every regular region equals such a union.

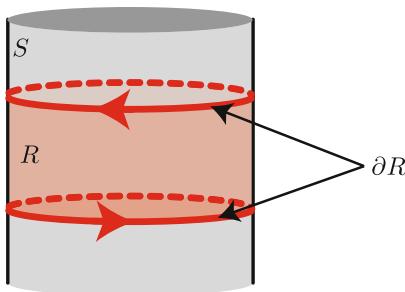


FIGURE 6.4. $\int \kappa_g(t) dt + \sum \alpha_i = 0$

With this interpretation, there are many examples of regular regions for which Eq. 6.4 is false. For example in Fig. 6.4, the two boundary components of R are geodesics with no vertices, so $\int \kappa_g(t) dt + \sum \alpha_i = 0$, but the right side of Eq. 6.4 equals 2π , because $K = 0$ at every point of the cylinder. As we will see, the issue is topological. Every polygonal region is homeomorphic to a disk, but the region R in Fig. 6.4 is not, and that's how it gets away with violating Eq. 6.4.

In order to fix Eq. 6.4, we must define a topological invariant of regular regions called the *Euler characteristic*:

DEFINITION 6.4.

Let R be a regular region of a regular surface S .

- (1) A **triangle** in S means a polygonal region in S with three vertices. The three smooth segments of the boundary of a triangle are called its **edges**.
- (2) A **triangulation** of R means a finite family $\{T_1, \dots, T_F\}$ of triangles such that:
 - (a) $\bigcup_i T_i = R$.
 - (b) If $i \neq j$, then $T_i \cap T_j$ either is empty or is a common edge or a common vertex of T_i and T_j .
- (3) The **Euler characteristic** of a triangulation $\{T_1, \dots, T_F\}$ of R is

$$\chi = V - E + F,$$

where F is the number of faces (triangles), E is the number of edges (with each edge counted only once, even if it is shared by two triangles), and V is the number of vertices (with each vertex counted only once, even if it is shared by multiple triangles).

The importance of this definitions stems from the following theorem:

THEOREM 6.5.

If R is a regular region of a regular surface S , then there exists a triangulation of R . Furthermore, every two triangulations of R have the same Euler characteristic.

We will not prove the first claim (that all regular regions can be triangulated), because the proof is somewhat technical. But the second claim (that every two triangulations have the same Euler characteristic) is an immediate consequence of the forthcoming Theorem 6.8.

Since the Euler characteristic of a triangulation of R depends only on R , we will henceforth denote it by $\chi(R)$ and call it the **Euler characteristic of R** .

If two regular regions, $R \subset S$ and $\tilde{R} \subset \tilde{S}$, are diffeomorphic, then they have the same Euler characteristic. This is because a diffeomorphism $f : R \rightarrow \tilde{R}$ sends a triangulation of R to a triangulation of \tilde{R} with the same values of V , E , and F .

The Euler characteristic can more generally be defined when S is a (not necessarily regular) surface and when the boundary components of R are traces of continuous (rather than piecewise-regular) curves. It is still true in this generality that a triangulation of R exists (with continuous rather than smooth edges), and that the Euler characteristic, $\chi = V - E + F$, depends only on R and not on the choice of triangulation. Working in this generality, two *homeomorphic* regions must have the same Euler characteristic.

An edge of a triangulation of $R \subset S$ is called an **interior edge** if it is shared by two triangles, or an **exterior edge** if it only belongs to one triangle. The exterior edges together comprise the boundary of R . If S is oriented, and if the boundaries of all triangles of the triangulation are positively oriented, then the two triangles sharing an interior edge will determine opposite orientations for this shared edge. Furthermore, the orientation of each exterior edge will agree with the positive orientation of the boundary of R . This is illustrated in Fig. 6.5.

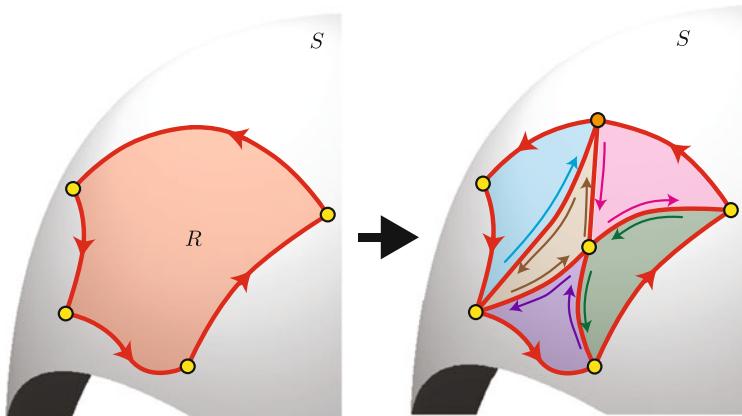


FIGURE 6.5. A triangulation of R . Each interior edge receives opposite orientations from the two triangles that share it. Here, $\chi(R) = V - E + F = 6 - 10 + 5 = 1$

EXAMPLE 6.6 (A Simple Region Has $\chi = 1$). A simple region means a regular region that is homeomorphic to the disk $\{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 \leq 1\}$. This is the same as a polygonal region, except that we are not assuming that it can be covered by a single surface chart. The region illustrated in Fig. 6.5 is simple. The pictured triangulation shows that its Euler characteristic (and hence the Euler characteristic of every simple region) equals 1.

EXAMPLE 6.7 (A Cylinder Has $\chi = 0$). A model for the bounded cylinder region, R , in Fig. 6.4 can be obtained by gluing a pair of opposite edges of a rectangle. Figure 6.6 shows a triangulation of the rectangle that induces a triangulation of R , from which we compute that $\chi(R) = 0$.

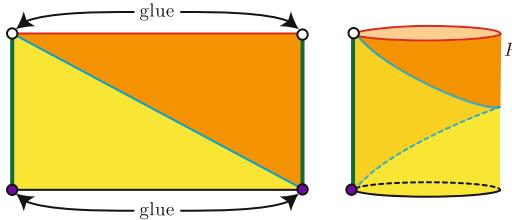


FIGURE 6.6. $\chi(R) = V - E + F = 2 - 4 + 2 = 0$

THEOREM 6.8 (The Global Gauss–Bonnet Theorem).

If S is an oriented regular surface and $R \subset S$ is a regular region with unit-speed positively oriented boundary components, then

$$\int \kappa_g(t) dt + \sum \alpha_i = 2\pi\chi(R) - \iint_R K dA,$$

where “ $\int \kappa_g(t) dt$ ” denotes the sum of the integrals of the geodesic curvature over all smooth segments of all boundary components of R , and “ $\sum \alpha_i$ ” denotes the sum of all vertices of all boundary components of R .

As in Definition 3.55 (on page 161), the term “ $\iint_R K dA$ ” means the sum of the integrals of K over the faces of a triangulation, which can be shown to be independent of the choice of triangulation.

PROOF. Let m denote the total number of vertices of all boundary components of R , so the expression “ $\sum \alpha_i$ ” from the theorem could more properly be written as “ $\sum_{i=1}^m \alpha_i$.” For each $i = 1, \dots, m$, let $\beta_i = \pi - \alpha_i$ denote the corresponding internal angle (as in Eq. 6.2 on page 325).

Choose a triangulation of R , denoted by $\{T_1, \dots, T_F\}$. Denote the number of vertices, edges, and faces of this triangulation by V , E , and F respectively. Let E_{ext} and E_{int} denote the numbers of exterior and interior edges respectively, so

$$E = E_{\text{ext}} + E_{\text{int}}.$$

Similarly, let V_{ext} and V_{int} respectively denote the numbers of exterior vertices (belonging to ∂R) and interior vertices (not belonging to ∂R). Notice that

$$V = V_{\text{ext}} + V_{\text{int}} = m + V_{\text{ext}}^+ + V_{\text{int}},$$

where V_{ext}^+ is the number of exterior vertices that were introduced by the triangulation (such as the top orange vertex in Fig. 6.5).

Choose a positively oriented unit-speed parametrization of each triangle of the triangulation. Applying the local Gauss–Bonnet theorem to each triangle and summing the results gives

$$(6.5) \quad \underbrace{\sum_{i=1}^F \left(\int_{\partial T_i} \kappa_g(t) dt \right)}_{\boxed{\int \kappa_g(t) dt}} + \underbrace{\sum_{i=1}^F \left(\begin{array}{l} \text{sum of the three} \\ \text{signed angles of } T_i \end{array} \right)}_{\boxed{A}} = 2\pi F - \underbrace{\sum_{i=1}^F \iint_{T_i} K dA}_{\boxed{\iint_R K dA}}$$

The left sum simplifies to $\int \kappa_g(t) dt$, because each interior edge receives opposite orientations from the two triangles that it bounds. This causes the integrals over interior edges to cancel, leaving only the integrals over exterior edges, which constitute the boundary of R .

It remains to simplify the middle sum, which is the sum of all signed angles of all triangles:

$$\begin{aligned} \boxed{A} &= 3\pi F - \sum_{i=1}^F \left(\begin{array}{l} \text{sum of the three} \\ \text{interior angles of } T_i \end{array} \right) \\ &= 3\pi F - \left(\begin{array}{l} \text{sum of all interior angles} \\ \text{at all interior vertices} \end{array} \right) - \left(\begin{array}{l} \text{sum of all interior angles} \\ \text{at all exterior vertices} \end{array} \right) \\ &= 3\pi F - (2\pi V_{\text{int}}) - \left(\sum_{i=1}^m \beta_i + \pi V_{\text{ext}}^+ \right) \\ &= 3\pi F - (2\pi V_{\text{int}}) - \left(\left(\pi m - \sum_{i=1}^m \alpha_i \right) + \pi V_{\text{ext}}^+ \right) \\ &= \sum_{i=1}^m \alpha_i + \pi (2E_{\text{int}} + E_{\text{ext}}) - 2\pi V_{\text{int}} - \pi V_{\text{ext}}^+ - \pi m \quad (3F = 2E_{\text{int}} + E_{\text{ext}}) \\ \\ &= \sum_{i=1}^m \alpha_i + \pi (2E_{\text{int}} + 2E_{\text{ext}}) - 2\pi V_{\text{int}} \underbrace{- \pi V_{\text{ext}}^+ - \pi m - \pi V_{\text{ext}}}_{-\pi V_{\text{ext}}} \quad (V_{\text{ext}} = E_{\text{ext}}) \\ &= \sum_{i=1}^m \alpha_i + 2\pi E - 2\pi V. \end{aligned}$$

Substituting this expression into Eq. 6.5 proves the theorem. \square

There are many powerful consequences of the global Gauss–Bonnet theorem. For example, the conclusion of the local Gauss–Bonnet theorem remains true when R is any simple region (rather than just a polygonal region), because the Euler characteristic of every simple region equals 1, according to Example 6.6. This observation will allow us to prove the following corollary:

COROLLARY 6.9.

If S is a regular surface with $K \leq 0$, then two geodesics from a point $p \in S$ cannot meet again at a point $q \in S$ in such a way that their traces form the boundary of a simple region of S .

PROOF. If they did, Eq. 6.3 would give $\beta_1 + \beta_2 = \iint_R K dA \leq 0$, where β_1, β_2 are the interior angles formed at p and q . Thus, $\beta_1 = \beta_2 = 0$, contradicting the uniqueness of geodesics (Theorem 5.3 on page 248). \square

Corollary 6.9 is a strengthening of Proposition 5.76 (on page 315), which claims *infinitesimally* that geodesics never rejoin in a surface with $K \leq 0$. On the other hand, Fig. 6.7 shows how geodesics in surfaces with $K \leq 0$ can rejoin without violating this corollary. This violation is possible because the cylinder is not *simply connected*:

DEFINITION 6.10.

*A regular surface S is called **simply connected** if every piecewise-regular simple closed curve in S bounds a simple region.*

A consequence of Corollary 6.9 is the following:

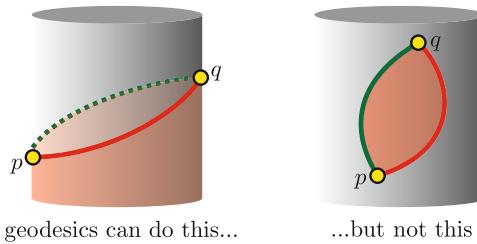


FIGURE 6.7. How geodesics can rejoin on the cylinder

COROLLARY 6.11.

If S is a connected, simply connected, complete regular surface with $K \leq 0$, then for every $p \in S$, $\exp_p : T_p S \rightarrow S$ is a diffeomorphism.

PROOF. The exponential map is surjective because S is complete, is a *local* diffeomorphism by Corollary 5.79 (on page 317), and is one-to-one by Corollary 6.9, so it is a diffeomorphism. \square

Readers familiar with covering space theory could instead prove Corollary 6.11 directly from Corollary 5.79 without requiring the Gauss–Bonnet theorem.

Our next corollary generalizes Exercise 5.20 on page 257:

COROLLARY 6.12.

If S is a regular surface that is diffeomorphic to a cylinder and has Gaussian curvature $K < 0$, then S has at most one simple closed geodesic (up to reparametrization).

PROOF. Suppose S had two distinct simple closed geodesics, such as the green and red ones in Fig. 6.8. Denote their traces by C_1, C_2 . By considering their images under a diffeomorphism from S to the punctured plane, $\mathbb{R}^2 - \{(0, 0)\}$, one can conclude that:

- (1) If C_1 and C_2 intersected in one point, this would contradict the uniqueness of geodesics.
- (2) If C_1 and C_2 intersected in multiple points, then their segments between two consecutive intersection points would bound a simple region, contradicting Corollary 6.9.
- (3) If C_1 and C_2 do not intersect, then they together bound a region, R , that is diffeomorphic to (and hence has the same Euler characteristic as) the bounded cylinder from Example 6.7. The Gauss–Bonnet theorem gives $\int_R K dA = 2\pi\chi(R) = 0$, contradicting the assumption that $K < 0$.

□

EXERCISES

EXERCISE 6.4. State and prove a generalization of Green’s theorem (on page 91) that applies to an arbitrary regular region of \mathbb{R}^2 .

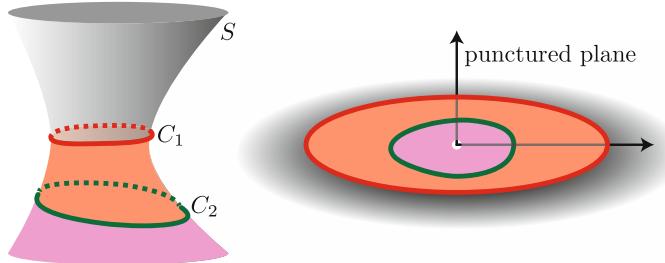


FIGURE 6.8. The proof of Corollary 6.12

EXERCISE 6.5 (Jacobi’s Theorem). Let $\gamma : [a, b] \rightarrow \mathbb{R}^3$ be a simple closed unit-speed space curve with nowhere-vanishing curvature. For all $t \in [a, b]$, let $\{\mathbf{t}(t), \mathbf{n}(t), \mathbf{b}(t)\}$ denote the Frenet frame at $\gamma(t)$. Notice that $t \mapsto \mathbf{n}(t)$ is a regular parametrized curve on the sphere S^2 . If this curve on S^2 is

reparametrized by arc length, prove that its total geodesic curvature equals zero. If it is simple, conclude that it divides S^2 into two regions of equal area.

EXERCISE 6.6 (Discrete Holonomy). Let S be a complete oriented regular surface and $\gamma : [a, b] \rightarrow S$ a regular curve in S . Let $a = t_0 < t_1 < \dots < t_n = b$ be a regular partition of $[a, b]$, which means that $t_i = a + i\Delta t$, where $\Delta t = \frac{b-a}{n}$. Explain why for sufficiently large n , there exists a unique geodesic from $\gamma(t_i)$ to $\gamma(t_{i+1})$ for each i , which together form a piecewise-geodesic curve β from $\gamma(a)$ to $\gamma(b)$, as in Fig. 6.9. As $n \rightarrow \infty$, prove that the sum of the signed angles of β converges to the angle displacement along γ .

EXERCISE 6.7. A **generalized triangulation** of a regular region R means a finite family $\{T_1, \dots, T_F\}$ of polygonal regions whose union equals R such that if $i \neq j$, then $T_i \cap T_j$ is either empty or is a common edge or a common vertex of T_i and T_j . This is the same as a triangulation, except that the faces are allowed to have any number of edges. Prove that for every generalized triangulation of R , the value $V - E + F$ equals the Euler characteristic of R .

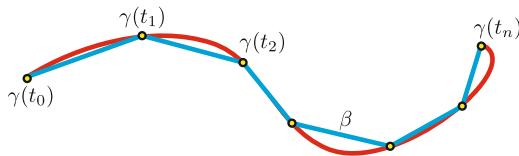


FIGURE 6.9. β is a piecewise-geodesic approximation of γ

EXERCISE 6.8. For the torus of revolution (Exercise 3.23 on page 138), Let R be the region bounded by two arbitrary latitudinal curves or by two arbitrary longitudinal curves, as in Fig. 6.10. Compute all terms of the global Gauss–Bonnet theorem, and verify the theorem for such regions.

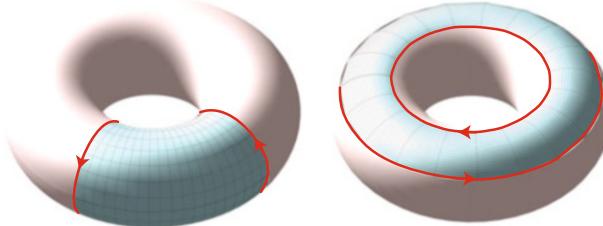


FIGURE 6.10. Regions of the torus for Exercise 6.8

3. Compact Surfaces

A compact regular surface can be considered a regular region with no boundary components. Since the boundary is empty, all of the edges of a triangulation will be interior edges. In this situation, the global Gauss–Bonnet theorem says the following:

COROLLARY 6.13.

If S is a compact connected regular surface, then

$$\iint_S K \, dA = 2\pi\chi(S).$$

Some sources refer to this corollary as “the Gauss–Bonnet theorem,” because it is important enough to deserve this title, even though it’s really a special case of the global Gauss–Bonnet theorem. In this section, we discuss some of its powerful implications, including Poincaré’s theorem.

We begin by computing the Euler characteristics of some compact surfaces, beginning with the sphere:

EXAMPLE 6.14 ($\chi(S^2) = 2$). Figure 6.11 shows two different triangulations of the sphere, S^2 . The first is formed by intersecting S^2 with the three coordinate planes. The second is modeled after a tetrahedron. Using either triangulation, we compute that $\chi(S^2) = 2$. Therefore, Corollary 6.13 says that $\iint_S K \, dA = 4\pi$ for every regular surface, S , that is diffeomorphic to S^2 , which is quite remarkable!

The sphere S^2 has genus zero, according to the following definition:

DEFINITION 6.15.

The **genus** of a compact connected regular surface S is

$$g = 1 - \frac{1}{2}\chi(S).$$

For several choices of g , Fig. 6.12 illustrates a surface that looks like the exterior crust of g bagels that merged together as they baked in the

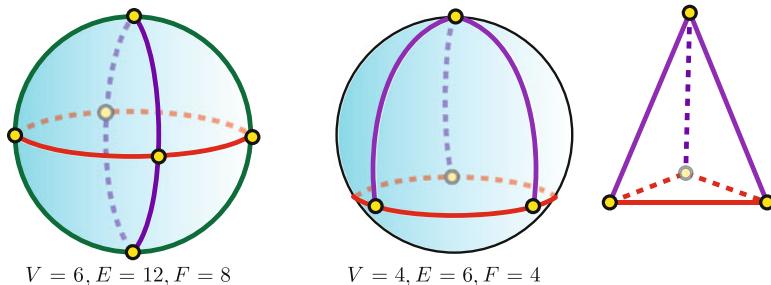


FIGURE 6.11. $\chi(S^2) = 2$

oven. The illustrated *generalized triangulations* of these surfaces (as defined in Exercise 6.7 on page 333) verify that the bagel count g equals the genus. Although only the choices $g = 1, 2, 3$ are shown, the bottom row indicates how the pattern continues. Thus for every integer $g \geq 1$, the exterior crust of g merged bagels has Euler characteristic $2 - 2g$ and hence has genus g .

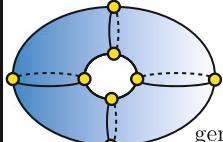
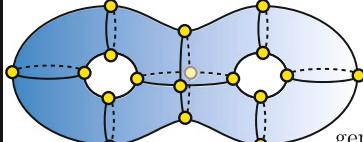
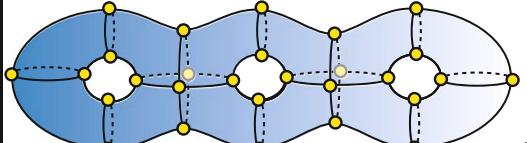
| | V | E | F | χ |
|---|-----------|-----------|------|----------|
|  genus 1 | 8 | 16 | 8 | 0 |
|  genus 2 | 18 | 36 | 16 | -2 |
|  genus 3 | 28 | 56 | 24 | -4 |
| surface of genus g | $10g - 2$ | $20g - 4$ | $8g$ | $2 - 2g$ |

FIGURE 6.12. The exterior crust of g merged bagels has genus g

In summary, S^2 has genus 0, while for every integer $g \geq 1$, the exterior crust of g merged bagels has genus g . We state without proof the following theorem, which says that other than these examples, there are no other topologically distinct compact regular surfaces:

THEOREM 6.16.

The genus of every compact connected regular surface is an integer ≥ 0 . Furthermore, every pair of compact connected regular surfaces with the same genus must be diffeomorphic.

Equivalently, the Euler characteristic of every compact connected regular surface is an element of $\{\dots, -6, -4, -2, 0, 2\}$, and every two compact connected regular surfaces with the same Euler characteristic must be diffeomorphic. This implies the following:

COROLLARY 6.17.

A compact connected regular surface S with $K > 0$ must be diffeomorphic to S^2 .

PROOF. If S were not diffeomorphic to S^2 , then $\chi(S) \neq 2$, so $\chi(S) \leq 0$. Thus,

$$\iint_S K \, dA = 2\pi\chi(S) \leq 0,$$

contradicting the positive curvature hypothesis. \square

We conclude this section with Poincaré’s theorem, which is a powerful application of the Gauss–Bonnet theorem. Our setup begins with some terminology:

DEFINITION 6.18.

*Let V be a tangent field on a regular surface S (Definition 3.43 on page 152). We call $p \in S$ a **singular point** of V if $V(p) = \mathbf{0}$. In this case, p is called **isolated** if there exists a neighborhood of p in S on which p is the only singular point of V .*

If S is compact and all of the singular points of V are isolated, then there are only finitely many singular points (Exercise 6.16).

Assuming that S is oriented, we will next define the *index* of V at p as an integer that answers the following question: if I walk counterclockwise once around a small loop centered at p while turning my body so that I am always facing in the direction of the vectors I encounter along the way, how many total counterclockwise rotations does this activity force me to perform? As a warmup, we’ll first make this idea precise for a vector field on \mathbb{R}^2 , using the notion of *degree* from Definition 2.3 on page 63:

DEFINITION 6.19.

*The **index** at an isolated singular point p of a vector field V on \mathbb{R}^2 equals the degree of the function $f_\epsilon : [a, b] \rightarrow S^1$ defined as*

$$f_\epsilon(t) = \frac{V(\gamma_\epsilon(t))}{|V(\gamma_\epsilon(t))|},$$

where $\gamma_\epsilon : [a, b] \rightarrow \mathbb{R}^2$ parametrizes the counterclockwise circle of radius ϵ centered at p , with $\epsilon > 0$ chosen small enough that V has no singular points other than p on or inside this circle.

Some examples are illustrated in Fig. 6.13. This definition is independent of the choice of ϵ because the degree of f_ϵ is an integer-valued quantity, and Lemma 2.4 (on page 63) can be used to show that it changes continuously with ϵ , and must therefore be constant.

To define “index” when $S \neq \mathbb{R}^2$, we essentially apply Definition 6.19 in local coordinates. But to make the measurement more versatile, we’ll replace the circles with arbitrary polygonal regions:

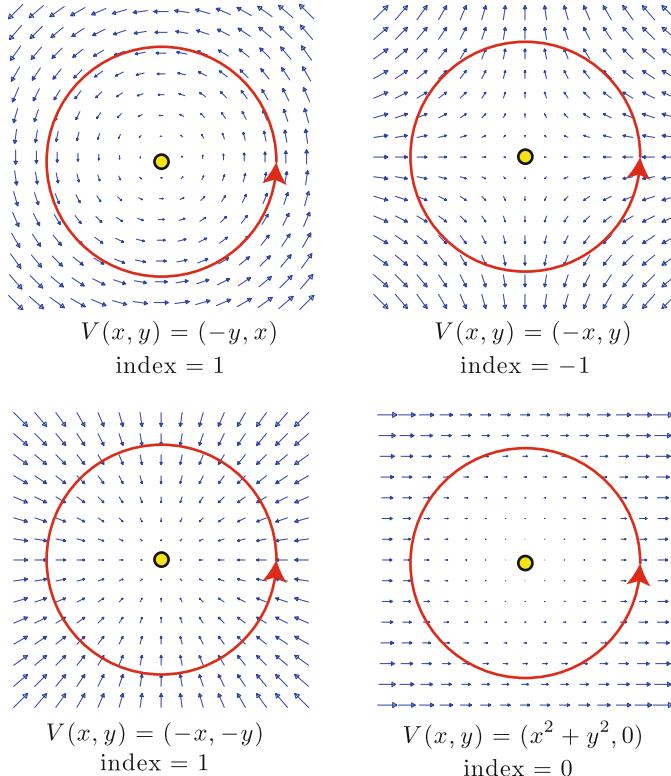


FIGURE 6.13. The index of some vector fields on \mathbb{R}^2 that all have isolated singular points at $(0, 0)$

DEFINITION 6.20.

Let p be an isolated singular point of a tangent field V on an oriented regular surface S . Let $R \subset S$ be a polygonal region whose interior contains p , with no other singular points in its interior or boundary. Let $\gamma : [a, b] \rightarrow S$ be a positively oriented parametrization of ∂R . Let $\sigma : U \subset \mathbb{R}^2 \rightarrow S$ be a surface patch covering R that is compatible with the orientation. For $t \in [a, b]$, let $\sigma_u(t)$ and $V(t)$ denote the values of σ_u and V at $\gamma(t)$. Let $\varphi : [a, b] \rightarrow \mathbb{R}$ be a continuous function such that for all $t \in [a, b]$, $\varphi(t)$ equals the angle from $\sigma_u(t)$ to $V(t)$; more precisely, $V(t)$ is a positive scalar multiple of the counterclockwise rotation of $\sigma_u(t)$ by the angle $\varphi(t)$. The **index** of V at p is defined as

$$(6.6) \quad I(p) = \frac{1}{2\pi} \Delta \varphi = \frac{1}{2\pi} (\varphi(b) - \varphi(a)).$$

For example, each of the tangent fields on S^2 illustrated in Fig. 6.14 has a pair of singular points of index 1 (at the north and south poles).

Definition 6.20 is independent of the choice of polygonal region. Although we will omit the details, proving this assertion involves showing that every polygonal region can be continuously deformed into any other; since the index is integer-valued and changes continuously through the deformation, it must remain constant.

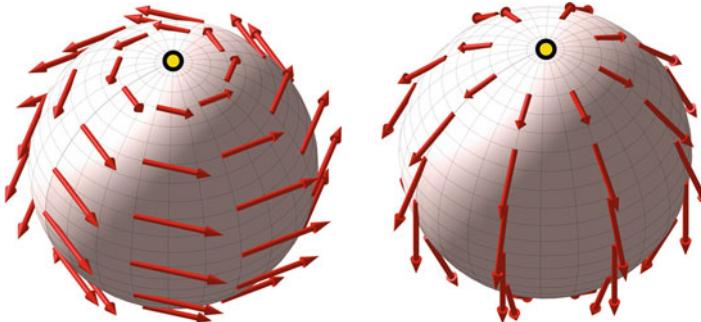


FIGURE 6.14. Each tangent field on S^2 has a pair of singular points of index 1

Definition 6.20 is also independent of the choice of (local) orientation of S (Exercise 6.15), so it makes sense to measure the index at an isolated singular point of a vector field on a nonoriented surface.

We will now prove that Definition 6.20 is also independent of the choice of surface patch σ . For this, with all assumptions and notation as in the definition, let $w : [a, b] \rightarrow \mathbb{R}^3$ be a parallel vector field along γ . Let $\phi : [a, b] \rightarrow \mathbb{R}$ be a continuous function such that for all $t \in [a, b]$, $\phi(t)$ equals the angle from $\sigma_u(t)$ to $w(t)$. According to Corollary 6.3 (on page 322),

$$(6.7) \quad \Delta\phi = \phi(b) - \phi(a) = \int_R K \, dA.$$

Subtracting Eq. 6.6 from Eq. 6.7 yields

$$(6.8) \quad \Delta(\phi - \varphi) = (\phi - \varphi)(b) - (\phi - \varphi)(a) = \iint_R K \, dA - 2\pi \cdot I(p).$$

Since ϕ is the angle from σ_u to w , while φ is the angle from σ_u to V , their difference $(\phi - \varphi)$ equals the angle from V to w . Therefore, the value $\Delta(\phi - \varphi)$ equals the net rotation of V relative to a parallel field along γ , which has nothing to do with σ . So solving Eq. 6.8 for $I(p)$ yields an expression involving only measurements that are independent of σ .

Equation 6.8 is also the key to proving the following theorem:

THEOREM 6.21 (Poincaré's Theorem).

If S is a compact connected regular surface and V is a tangent field on S all of whose singular points are isolated, then the sum of the indices of V at all the singular points equals $\chi(S)$.

PROOF. It is possible to find a triangulation of S such that none of the singular points of V lie on edges, and no face contains more than one singular point.

If T is a face whose interior contains a singular point p , Eq. 6.8 gives

$$\iint_T K \, dA - 2\pi \cdot I(p) = \text{the net rotation of } V \text{ relative to a parallel field around } \partial T.$$

If T is a face whose interior contains no singular points, the logic of Eq. 6.8 reduces to

$$\iint_T K \, dA - 0 = \text{the net rotation of } V \text{ relative to a parallel field around } \partial T.$$

Notice that each “net rotation” term is independent of the choice of parallel field along the boundary triangle. In fact, each such term is the sum over the three boundary edges of the net rotation of V relative to a parallel field along that edge. This measurement makes sense for a single edge (there is no need for the choice of parallel field to be consistent even between the three edges). The measurement changes by a sign if the orientation of the edge is reversed.

Summing over all of the faces yields

$$\iint_S K \, dA - 2\pi \cdot (\text{sum of indices of all singular points}) = 0.$$

This is because all edges are interior edges, so the net rotation terms cancel. Since $\iint_S K \, dA = 2\pi\chi(S)$, this completes the proof. \square

COROLLARY 6.22.

A compact connected regular surface S on which there exists a nowhere-vanishing tangent field must be diffeomorphic to a torus.

PROOF. A nowhere-vanishing tangent field has no singular points, so Poincaré’s theorem says that $0 = \chi(S)$. In other words, the genus of S equals 1. \square

Figure 6.15 demonstrates that a torus *does* support nowhere-vanishing vector fields. The illustrated vector fields are the coordinate fields σ_θ and σ_t associated to the natural parametrization σ defined in Exercise 3.23 on page 138.

The corollary implies in particular that S^2 *does not* support a nowhere-vanishing vector field; this fact is sometimes called the **hairy ball theorem** and paraphrased as, “you can’t comb a hairy ball flat.” Each tangent field on S^2 in Fig. 6.14 has a pair of singular points with index 1. The Wikipedia article on the “Hairy Ball Theorem” includes an animated illustration of a tangent field on S^2 with only one singular point, of index 2.

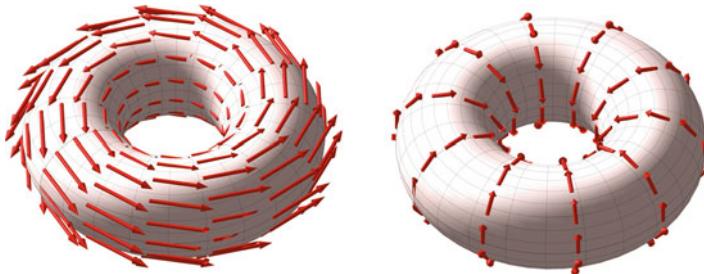


FIGURE 6.15. Nowhere-vanishing tangent fields on a torus of revolution

EXERCISES

EXERCISE 6.9. Let S be a compact connected regular surface with Gaussian curvature $K > 0$. Prove that every two simple closed geodesics in S must intersect.

EXERCISE 6.10. The proof of the global Gauss–Bonnet theorem simplifies in the special case that R is a compact connected regular surface (the special case of Corollary 6.13). Rewrite the proof for this special case.

EXERCISE 6.11. If S is a connected compact regular surface with non-negative Gaussian curvature ($K \geq 0$), prove that S is diffeomorphic to S^2 .
HINT: Use Exercise 4.16 on page 211.

EXERCISE 6.12. A soccer ball is constructed from 12 pentagonal faces and 20 hexagonal faces; see Fig. 6.16 (left). Consider this a generalized triangulation of S^2 . Determine V , E , and F , and use this to again verify that $\chi(S^2) = 2$.

EXERCISE 6.13. Figure 6.16 (right) shows a generalized triangulation of a torus with rectangular faces. Explain why $F = V = \frac{E}{2}$, and use this to again verify that the Euler characteristic of the torus equals 0.

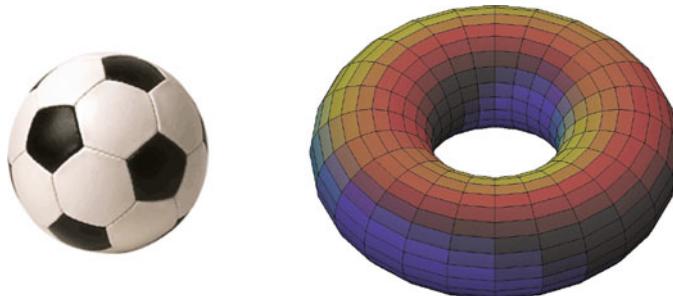


FIGURE 6.16. $\chi(\text{soccer ball}) = 2$, and $\chi(\text{torus}) = 0$

EXERCISE 6.14. On a regular surface that is diffeomorphic to S^2 , is it possible for the region on which $K \leq 0$ to have greater area than the region on which $K \geq 0$?

EXERCISE 6.15. Let p be an isolated singular point of a vector field V on an oriented surface S . Prove that the index of p (Definition 6.20) would be the same if measured with respect to the other orientation of S .

EXERCISE 6.16. If all of the singular points of a tangent field on a compact regular surface are isolated, prove that there are only finitely many of them.

EXERCISE 6.17. In each vector field in Fig. 6.13, verify computationally that the index at the origin is as claimed.

EXERCISE 6.18. Let k be a positive integer. Identify \mathbb{R}^2 with the complex plane by associating $(a, b) \in \mathbb{R}^2$ with $z = a + bi$.

- (1) Show that the vector field on \mathbb{R}^2 defined as $V(z) = z^k$ has an isolated singular point at the origin with index k .
- (2) Show that the vector field defined as $V(z) = \bar{z}^k$ has an isolated singular point at the origin with index $-k$, where the overline denotes complex conjugation.
- (3) Use a computer graphing application to graph these vector fields for several choices of k .

EXERCISE 6.19. Let $f : S \rightarrow \tilde{S}$ be a diffeomorphism between smooth surfaces. Let V be a vector field on S . Let \tilde{V} be the induced vector field on \tilde{S} , whose value at $f(p)$ equals $df_p(V(p))$ for each $p \in S$. Let p be an isolated singular point of V . Prove that the index of V at p equals the index of \tilde{V} at $f(p)$.

EXERCISE 6.20. Let $f : S^2 \rightarrow S^2$ be a smooth function. Prove that there exists $p \in S^2$ such that $f(p) = \pm p$.

EXERCISE 6.21. Construct a physical model of a surface with genus 2, and draw on it a vector field with exactly one singular point.

EXERCISE 6.22. Let S be a torus of revolution (Exercise 3.23 on page 138). Let $f : S \rightarrow \mathbb{R}$ be the “height function” defined such that for all $p = (x, y, z) \in S$, $f(p) = x$. Let ∇f be its gradient (Exercise 3.72 on page 160), which is illustrated in Fig. 6.17. With the terminology of Exercises 3.44 (on page 146) and 4.8 (on page 205), prove that f has exactly four critical points, all of which are nondegenerate. If p is any one of these four critical points, let $H_p : T_p S \rightarrow T_p S$ denote the self-adjoint linear transformation associated to the Hessian of f at p , and verify that

$$\begin{aligned} & \left(\begin{array}{l} \text{the eigenvalues of } H_p \\ \text{have the same sign} \end{array} \right) \iff \left(\begin{array}{l} p \text{ is a local} \\ \text{min or max of } f \end{array} \right) \iff \left(\begin{array}{l} \text{the index of } \nabla f \\ \text{at } p \text{ equals } 1 \end{array} \right) \\ & \left(\begin{array}{l} \text{the eigenvalues of } H_p \\ \text{have different signs} \end{array} \right) \iff \left(\begin{array}{l} p \text{ is a saddle point} \\ \text{of } f \end{array} \right) \iff \left(\begin{array}{l} \text{the index of } \nabla f \\ \text{at } p \text{ equals } -1 \end{array} \right) \end{aligned}$$

Comment: Although we will not prove it here, these conclusions apply much more generally to every nondegenerate critical point of every smooth

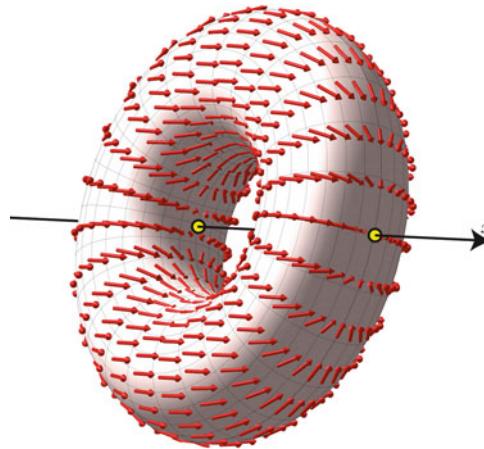


FIGURE 6.17. The gradient of the x -height function on the torus of revolution

function on every regular surface. Combining this fact with Poincaré’s theorem yields the following: if S is a compact connected regular surface and $f : S \rightarrow \mathbb{R}$ is a smooth function all of whose critical points are nondegenerate, then

$$\boxed{(\text{number of local extrema of } f) - (\text{number of saddle points of } f) = \chi(S).}$$

□

4. A Sampling of Other Global Theorems

A fundamental theme of differential geometry is the relationship between curvature and global shape. The global shape of a simple closed curve was related to its curvature by Hopf’s Umlaufsatz, the four vertex theorem, Fenchel’s theorem, and the Fary–Milnor theorem, to name a few. For compact surfaces, the Gauss–Bonnet theorem provided a beautiful relationship between curvature and global shape.

In this section, we overview (without proofs) a few additional fundamentally important theorems that relate the curvature of a surface to its global shape.

THEOREM 6.23 (Complete Surfaces with Constant Curvature).

Let $\lambda \in \mathbb{R}$, and let S be a connected complete regular surface with constant Gaussian curvature $K = \lambda$. Then:

- (1) (**Hilbert’s Theorem**) $\lambda \geq 0$; in other words, there does not exist a complete regular surface with constant negative Gaussian curvature.
- (2) If $\lambda = 0$, then S is a plane or a generalized cylinder.
- (3) If $\lambda > 0$, then S is a sphere of radius $\frac{1}{\sqrt{\lambda}}$.

A “generalized cylinder” is defined as a regular surface S such that through each point $p \in S$ there passes a unique line $L_p \subset S$, and for each pair $p, q \in S$, L_p is parallel to L_q (this is slightly more general than the definition in Exercise 3.18, because it doesn’t require the generating curve to be closed).

Each of these three statements is a significant result requiring a nontrivial proof. We will discuss only how these results fit together with other theorems and examples in this book.

Recall from Exercise 4.16 (on page 211) that every compact regular surface has a point of positive Gaussian curvature. Thus, in the case $\lambda < 0$ and $\lambda = 0$, we knew that S could not be compact.

On the other hand, in the case $\lambda > 0$, we knew from Bonnet’s theorem (on page 308) that S must be compact, and then from Corollary 6.17 (on page 336) that S must be diffeomorphic to a sphere. We also knew from Proposition 5.32 (on page 273) that S must be *locally* isometric to a sphere. It probably seems quite believable that something locally isometric to and diffeomorphic to a sphere must in fact be a sphere.

We next comment about the word “is” in the above theorem. Part (3) concluded that S *is* a sphere. A priori, this is potentially stronger than the conclusion that S *is isometric to* a sphere. Recall that Example 3.65 (on page 168) exhibited an isometric pair of compact surfaces that were not related by a rigid motion of \mathbb{R}^3 . However, part (3) implies that this doesn’t happen for spheres. If you’re isometric to a sphere, then you are a sphere. Similarly, part (2) classifies how complete $K = 0$ surfaces can be positioned in \mathbb{R}^3 , which is different from classifying their possible isometry types. For example, every generalized cylinder with a nonclosed generating curves is isometric to a plane.

Next observe that all three assertions of Theorem 6.23 would be false without the hypothesis that S is complete. Counterexamples include not only open subsets of the surfaces in the theorem, but also the following substantially different possibilities:

- $\boxed{\lambda < 0}$: The pseudosphere (page 225) and Dini’s surface (page 226).
- $\boxed{\lambda = 0}$: Generalized cones (pages 137 and 170), tangent developables (pages 138 and 188), and the flat generalized helicoid (page 226).
- $\boxed{\lambda > 0}$: Fake spheres (page 222).

In addition to these, there are many more examples in the literature of non-complete surfaces with constant Gaussian curvature. These examples even exhibit a wide variety of *local* behavior, for although every pair of surfaces with the same constant Gaussian curvature are *locally* isometric by Proposition 5.32 (on page 273), they need not be locally congruent; that is, it is not necessarily possible to find neighborhoods of points on these surfaces that differ by a rigid motion of \mathbb{R}^3 . This local flexibility indicates that truly global arguments are required to prove Theorem 6.23.

Once the constant-curvature surfaces are understood, the next most natural family of surfaces are those whose curvature is bounded (below or above) by zero. The global shape of surfaces with nonnegative Gaussian curvature is restricted as follows:

THEOREM 6.24 (Complete Surfaces with Nonnegative Curvature).

Let S be a connected complete regular surface with nonnegative Gaussian curvature ($K \geq 0$). Then:

- (1) *If S is compact, then S is diffeomorphic to S^2 .*
- (2) *If S is noncompact and has at least one point with $K > 0$, then S is diffeomorphic to \mathbb{R}^2 .*

Part (1) is proven in Exercise 6.11. Part (2) is more difficult, and will not be proven in this text. There is no part (3), because the apparently missing case is handled by Theorem 6.23: if S is noncompact and has no points of positive curvature, then $K = 0$, so S is a plane or a generalized cylinder.

In part (1), if S is compact with $K > 0$, it can be proven that S is **strictly convex**, some consequences of which were explored in Exercise 4.27 on page 217.

What about the other side of zero? If S is a complete surface with nonpositive curvature ($K \leq 0$), what can be said about S ? We know that S must be noncompact (Exercise 4.16 on page 211). We also know from Corollary 5.79 on page 317 that the exponential map at every point is a local diffeomorphism, and that if S is simply connected, then it is a global diffeomorphism by Corollary 6.11 (on page 331). But these observations are not as restrictive as one might have guessed. In fact, even the much smaller class of complete *minimal* surfaces (which satisfy $K \leq 0$ according to Exercise 4.47 on page 234) includes a wide variety of diffeomorphism types; see Fig. 6.18 or perform a web image search for “complete minimal surface.” Infinitely many copies of the right surface in Fig. 6.18 can be assembled to fill \mathbb{R}^3 , yielding a topologically complicated complete minimal surface.

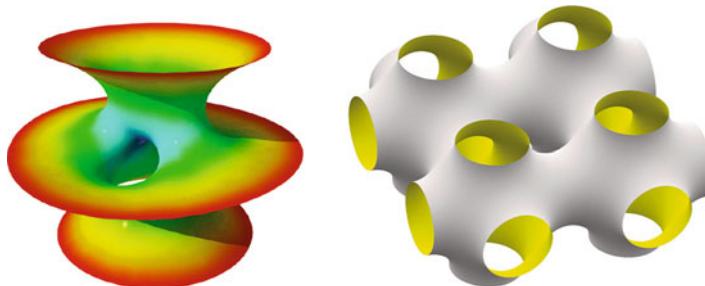


FIGURE 6.18. Minimal surfaces discovered by Costa (*left*) and Schwarz (*right*)

The Topology of Subsets of \mathbb{R}^n

In this appendix, we briefly review some notions from topology that are used throughout the book. The exposition is intended as a quick review for readers with some previous exposure to these topics.

1. Open and Closed Sets and Limit Points

The natural **distance function** on \mathbb{R}^n is defined such that for all $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$,

$$\text{dist}(\mathbf{a}, \mathbf{b}) = |\mathbf{a} - \mathbf{b}| = \sqrt{\langle \mathbf{a} - \mathbf{b}, \mathbf{a} - \mathbf{b} \rangle}.$$

Its most important property is the *triangle inequality*:

PROPOSITION A.1 (The Triangle Inequality).

For all $\mathbf{a}, \mathbf{b}, \mathbf{c} \in \mathbb{R}^n$,

$$\text{dist}(\mathbf{a}, \mathbf{c}) \leq \text{dist}(\mathbf{a}, \mathbf{b}) + \text{dist}(\mathbf{b}, \mathbf{c}).$$

PROOF. The Schwarz inequality (Lemma 1.12) says that $|\langle \mathbf{v}, \mathbf{w} \rangle| \leq |\mathbf{v}||\mathbf{w}|$ for all $\mathbf{v}, \mathbf{w} \in \mathbb{R}^n$. Thus,

$$\begin{aligned} |\mathbf{v} + \mathbf{w}|^2 &= |\mathbf{v}|^2 + 2\langle \mathbf{v}, \mathbf{w} \rangle + |\mathbf{w}|^2 \\ &\leq |\mathbf{v}|^2 + 2|\mathbf{v}| \cdot |\mathbf{w}| + |\mathbf{w}|^2 = (|\mathbf{v}| + |\mathbf{w}|)^2. \end{aligned}$$

So $|\mathbf{v} + \mathbf{w}| \leq |\mathbf{v}| + |\mathbf{w}|$. Applying this inequality to the vectors pictured in Fig. A.1 proves the triangle inequality. \square

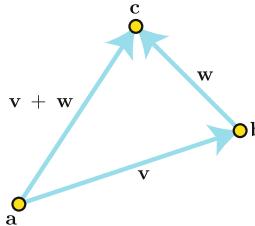


FIGURE A.1. Proof of the triangle inequality

Topology begins with precise language for discussing whether a subset of Euclidean space contains its boundary points. First, for $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$, we denote the **ball about \mathbf{p} of radius r** by

$$B(\mathbf{p}, r) = \{\mathbf{q} \in \mathbb{R}^n \mid \text{dist}(\mathbf{p}, \mathbf{q}) < r\}.$$

In other words, $B(\mathbf{p}, r)$ contains all points closer than a distance r from \mathbf{p} .

DEFINITION A.2.

A point $\mathbf{p} \in \mathbb{R}^n$ is called a **boundary point** of a subset $S \subset \mathbb{R}^n$ if for all $r > 0$, the ball $B(\mathbf{p}, r)$ contains at least one point in S and at least one point not in S . The collection of all boundary points of S is called the **boundary** of S .

Sometimes, but not always, boundary points of S are contained in S . For example, consider the “open upper half-plane”

$$H = \{(x, y) \in \mathbb{R}^2 \mid y > 0\},$$

and the “closed upper half-plane”

$$\overline{H} = \{(x, y) \in \mathbb{R}^2 \mid y \geq 0\}.$$

The x -axis, $\{(x, 0) \mid x \in \mathbb{R}^2\}$, is the boundary of H and also of \overline{H} . So H contains none of its boundary points, while \overline{H} contains all of its boundary points. This distinction is so central that we introduce vocabulary for it:

DEFINITION A.3.

Let $S \subset \mathbb{R}^n$ be a subset.

- (1) S is called **open** if it contains none of its boundary points.
- (2) S is called **closed** if it contains all of its boundary points.

In the previous example, H is open, while \overline{H} is closed. If part of the x -axis is adjoined to H (say the positive part), the result is neither closed nor open, since it contains some of its boundary points but not all of them.

A set $S \subset \mathbb{R}^n$ and its complement $S^c = \{\mathbf{p} \in \mathbb{R}^n \mid \mathbf{p} \notin S\}$ clearly have the same boundary. If S contains none of these common boundary points,

then S^c must contain all of them, and vice versa. So we learn a relationship between a subset and its complement:

LEMMA A.4.

A set $S \subset \mathbb{R}^n$ is closed if and only if its complement, S^c , is open.

The following provides a useful alternative definition of “open”:

LEMMA A.5.

A set $S \subset \mathbb{R}^n$ is open if and only if for all $\mathbf{p} \in S$, there exists $r > 0$ such that $B(\mathbf{p}, r) \subset S$.

PROOF. If S is not open, then it contains at least one of its boundary points, and no ball about such a boundary point is contained in S . Conversely, suppose that there is a point $\mathbf{p} \in S$ such that no ball about \mathbf{p} is contained in S . Then \mathbf{p} is a boundary point of S , so S is not open. \square

The proposition says that if you live in an open set, then so do all of your sufficiently close neighbors. How close is sufficient depends on how close you live from the boundary. For example, the set

$$S = (0, \infty) \subset \mathbb{R}$$

is open because for every $x \in S$, the ball $B(x, x/2) = (x/2, 3x/2)$ lies inside of S ; see Fig. A.2. When x is close to 0, the radius of this ball is small.

Similarly, for every $\mathbf{p} \in \mathbb{R}^n$ and $r > 0$, the ball $B = B(\mathbf{p}, r)$ is itself open, because about every $\mathbf{q} \in B$, the ball of radius $\frac{r - \text{dist}(\mathbf{p}, \mathbf{q})}{2}$ lies in B by the triangle inequality; see Fig. A.3.

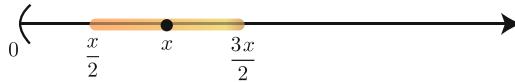


FIGURE A.2. The set $(0, \infty) \subset \mathbb{R}$ is open because it contains a ball about each of its points

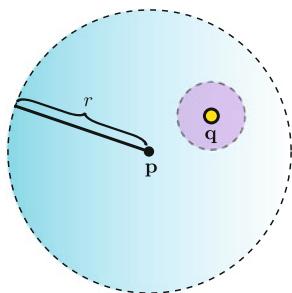


FIGURE A.3. The set $B(\mathbf{p}, r) \subset \mathbb{R}^n$ is open because it contains a ball about each of its points

LEMMA A.6.

The union of a collection of open sets is open. The intersection of a finite collection of open sets is open. The intersection of a collection of closed sets is closed. The union of a finite collection of closed sets is closed.

The collection of all open subsets of \mathbb{R}^n is called the **topology** of \mathbb{R}^n . It is surprising how many important concepts are topological, that is, definable purely in terms of the topology of \mathbb{R}^n . For example, the notion of whether a subset is closed is topological. The distance between points of \mathbb{R}^n is not topological. The notion of **convergence** is topological by the second definition below, although it may not initially seem so at first:

DEFINITION A.7.

An infinite sequence $\{\mathbf{p}_1, \mathbf{p}_2, \dots\}$ of points of \mathbb{R}^n is said to **converge** to $\mathbf{p} \in \mathbb{R}^n$ if either of the following equivalent conditions holds:

- (1) $\lim_{n \rightarrow \infty} \text{dist}(\mathbf{p}, \mathbf{p}_n) = 0$.
- (2) For every open set, U , containing \mathbf{p} , there exists an integer N such that $\mathbf{p}_n \in U$ for all $n > N$ (Fig. A.4).

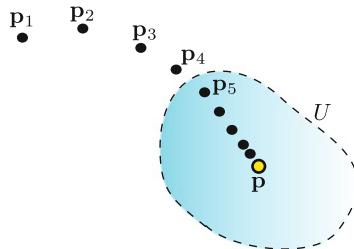


FIGURE A.4. A convergent sequence is eventually inside of every open set containing its limit

DEFINITION A.8.

A point $\mathbf{p} \in \mathbb{R}^n$ is called a **limit point** of a subset $S \subset \mathbb{R}^n$ if there exists an infinite sequence of points of S that converges to \mathbf{p} .

Every point $\mathbf{p} \in S$ is a limit point of S , as evidenced by the redundant infinite sequence $\{\mathbf{p}, \mathbf{p}, \mathbf{p}, \dots\}$. Every point of the boundary of S is a limit point of S as well. In fact, the collection of limit points of S equals the union of S and the boundary of S . Therefore, a set $S \subset \mathbb{R}^n$ is closed if and only if it contains all of its limit points, since this is the same as requiring it to contain all of its boundary points.

It is possible to show that a sequence converges without knowing its limit, just by showing that the terms get closer and closer to each other:

DEFINITION A.9.

An infinite sequence of points $\{\mathbf{p}_1, \mathbf{p}_2, \dots\}$ in \mathbb{R}^n is called a **Cauchy sequence** if for every $\epsilon > 0$, there exists an integer N such that $\text{dist}(\mathbf{p}_i, \mathbf{p}_j) < \epsilon$ for all $i, j > N$.

It is straightforward to prove that every convergent sequence is Cauchy. A fundamental property of Euclidean space is the converse:

PROPOSITION A.10.

Every Cauchy sequence in \mathbb{R}^n converges to some point of \mathbb{R}^n .

We end this section with an important *relative* notion of open and closed:

DEFINITION A.11.

Let $V \subset S \subset \mathbb{R}^n$ be subsets.

- (1) V is called **open in S** if there exists an open subset of \mathbb{R}^n whose intersection with S equals V .
- (2) V is called **closed in S** if there exists a closed subset of \mathbb{R}^n whose intersection with S equals V .

For example, setting $S^2 = \{(x, y, z) \in \mathbb{R}^3 \mid x^2 + y^2 + z^2 = 1\}$, notice that the hemisphere $\{(x, y, z) \in S^2 \mid z > 0\}$ is open in S^2 , because it is the intersection with S^2 of the following open set: $\{(x, y, z) \in \mathbb{R}^3 \mid z > 0\}$.

It is straightforward to show that V is open in S if and only if $\{\mathbf{p} \in S \mid \mathbf{p} \notin V\}$ is closed in S . The following is a useful equivalent characterization of “open in” and “closed in”:

PROPOSITION A.12.

Let $V \subset S \subset \mathbb{R}^n$ be subsets.

- (1) V is open in S if and only if for all $\mathbf{p} \in V$, there exists $r > 0$ such that $\{\mathbf{q} \in S \mid \text{dist}(\mathbf{p}, \mathbf{q}) < r\} \subset V$.
- (2) V is closed in S if and only if every $\mathbf{p} \in S$ that is a limit point of V is contained in V .

Part (1) says that if you live in a set that’s open in S , then so do all of your sufficiently close neighbors in S . Part (2) says that if you live in a set that’s closed in S , then you contain all of your limit points that belong to S . For example, the interval $[0, 1]$ is neither open nor closed in \mathbb{R} , but is open in $[0, 2]$ and is closed in $(-1, 1)$.

Let $\mathbf{p} \in S \subset \mathbb{R}^n$. A **neighborhood of \mathbf{p} in S** means a subset of S that is open in S and contains \mathbf{p} . For example, $(1 - \epsilon, 1 + \epsilon)$ is a neighborhood of 1 in $(0, 2)$ for every $\epsilon \in (0, 1]$. Also, $[0, \epsilon)$ is a neighborhood of 0 in $[0, 1]$ for every $\epsilon \in (0, 1]$.

The collection of all subsets of S that are open in S is called the **topology** of S . It has many natural properties. For example, the relative version of Lemma A.6 is true: the union of a collection of subsets of S that are open in S is itself open in S , and similarly for the other statements.

In the remainder of this appendix, pay attention to which properties of a set S are topological, that is, definable in terms of only the topology of S . For example, the notion of a sequence of points of S converging to $\mathbf{p} \in S$ is topological. Why? Because convergence means that the sequence is eventually inside of any neighborhood of \mathbf{p} in \mathbb{R}^n ; this is the same as being eventually inside of any neighborhood of \mathbf{p} in S , which has only to do with the topology of S . The idea is to forget about the ambient \mathbb{R}^n , and regard S as an independent object, with a topology and hence a notion of convergence.

2. Continuity

Let $S_1 \subset \mathbb{R}^{n_1}$ and $S_2 \subset \mathbb{R}^{n_2}$. A function $\mathbf{f} : S_1 \rightarrow S_2$ is called *continuous* if it maps nearby points to nearby points; more precisely:

DEFINITION A.13.

A function $\mathbf{f} : S_1 \rightarrow S_2$ is called **continuous** if for every infinite sequence $\{\mathbf{p}_1, \mathbf{p}_2, \dots\}$ of points in S_1 that converges to a point $\mathbf{p} \in S_1$, the sequence $\{\mathbf{f}(\mathbf{p}_1), \mathbf{f}(\mathbf{p}_2), \dots\}$ converges to $\mathbf{f}(\mathbf{p})$.

For example, the “step function” $f : \mathbb{R} \rightarrow \mathbb{R}$ defined as

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

is not continuous. Why? Because the sequence

$$\{1/2, 1/3, 1/4, \dots\}$$

in the domain of f converges to 0, but the sequence of images

$$\{f(1/2) = 1, f(1/3) = 1, f(1/4) = 1, \dots\}$$

converges to 1 rather than to $f(0) = 0$.

Notice that \mathbf{f} is continuous if and only if it is continuous when regarded as a function from S_1 to \mathbb{R}^{n_2} . It is nevertheless useful to forget about the ambient Euclidean spaces, and regard S_1 and S_2 as independent objects. This vantage point leads to the following beautiful, although less intuitive, way to define continuity:

PROPOSITION A.14.

For a function $\mathbf{f} : S_1 \rightarrow S_2$, the following are equivalent:

- (1) f is continuous.
- (2) For every set U that's open in S_2 , $\mathbf{f}^{-1}(U)$ is open in S_1 .
- (3) For every set U that's closed in S_2 , $\mathbf{f}^{-1}(U)$ is closed in S_1 .

Here, $\mathbf{f}^{-1}(U)$ denotes the set $\{\mathbf{p} \in S_1 \mid \mathbf{f}(\mathbf{p}) \in U\}$. The above step function fails this continuity test, because

$$\mathbf{f}^{-1}\left(\left(-\frac{1}{2}, \frac{1}{2}\right)\right) = (-\infty, 0],$$

which is not open in \mathbb{R} .

It is now clear that continuity is a topological concept, since this alternative definition involved only the topologies of S_1 and S_2 .

Familiar functions from \mathbb{R} to \mathbb{R} , such as polynomial, rational, trigonometric, exponential, and logarithmic functions, are all continuous on their domains. Furthermore, the composition of two continuous functions is continuous.

We next wish to describe what it means for S_1 and S_2 to be “topologically the same.” There should be a bijection between them that pairs open sets with open sets. More precisely:

DEFINITION A.15.

A function $\mathbf{f} : S_1 \rightarrow S_2$ is called a **homeomorphism** if \mathbf{f} is bijective and continuous and \mathbf{f}^{-1} is continuous. If such a function exists, then S_1 and S_2 are said to be **homeomorphic**.

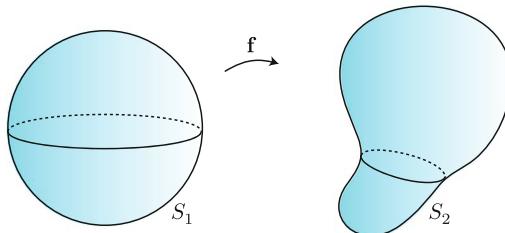


FIGURE A.5. Homeomorphic sets

Homeomorphic sets have the same “essential shape,” such as the two subsets of \mathbb{R}^3 pictured in Fig. A.5. The hypothesis that \mathbf{f}^{-1} is continuous is necessary. To see this, consider the function $\mathbf{f} : [0, 2\pi) \rightarrow S^1 \subset \mathbb{R}^2$ defined as $\mathbf{f}(t) = (\cos t, \sin t)$. It is straightforward to check that \mathbf{f} is continuous and bijective, but \mathbf{f}^{-1} is not continuous. (Why not?) We will see in Sect. 4 that $[0, 2\pi)$ is not homeomorphic to S^1 , since only the latter is compact.

3. Connected and Path-Connected Sets

DEFINITION A.16.

A subset $S \subset \mathbb{R}^n$ is called **path-connected** if for every pair $\mathbf{p}, \mathbf{q} \in S$, there exists a continuous function $\mathbf{f} : [0, 1] \rightarrow S$ with $\mathbf{f}(0) = \mathbf{p}$ and $\mathbf{f}(1) = \mathbf{q}$.

The terminology comes from visualizing the image of such a function \mathbf{f} as a path in S beginning at \mathbf{p} and ending at \mathbf{q} .

For example, the disk $A = \{(x, y) \in \mathbb{R}^2 \mid x^2 + y^2 < 1\}$ is path-connected, since every pair $\mathbf{p}, \mathbf{q} \in A$ can be connected by the straight line segment between them, explicitly parametrized as

$$\mathbf{f}(t) = \mathbf{p} + t(\mathbf{q} - \mathbf{p}).$$

But the disjoint union of two disks,

$$B = \{\mathbf{p} \in \mathbb{R}^2 \mid \text{dist}(\mathbf{p}, (-2, 0)) < 1 \text{ or } \text{dist}(\mathbf{p}, (2, 0)) < 1\},$$

is not path-connected, because no continuous path exists between points in different disks; see Fig. A.6.

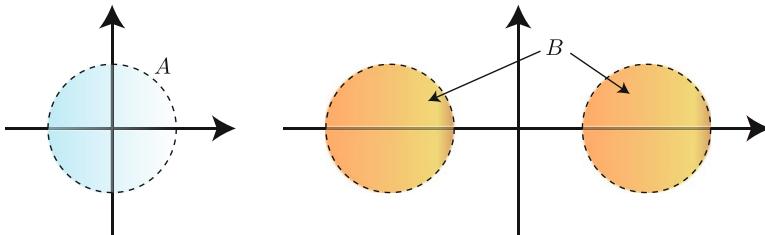


FIGURE A.6. A is path-connected, while B is not

In the non-path-connected example, the right disk is **clopen** (both open and closed) in B , and therefore so is the left disk. In other words, B decomposes into the disjoint union of two subsets that are both clopen in B .

DEFINITION A.17.

A set $S \subset \mathbb{R}^n$ is called **connected** if there is no subset of S (other than all of S and the empty set) that is **clopen in S** (both open in S and closed in S).

Such a separation of a path-connected set is impossible:

PROPOSITION A.18.

Every path-connected set $S \subset \mathbb{R}^n$ is connected.

PROOF. We first prove that the interval $[0, 1]$ has no clopen subsets other than itself and the empty set. Suppose $A \subset [0, 1]$ is another such set. Let t denote the infimum of A . Since A is closed, $t \in A$. Since A is open, there exists $r > 0$ such that all points of $[0, 1]$ with distance $< r$ from t lie in A . This contradicts the fact that t is the infimum of A unless $t = 0$. Therefore, $0 \in A$. Since the complement, A^c , of A is also clopen, the same argument proves that $0 \in A^c$, which is impossible.

Next, let $S \subset \mathbb{R}^n$ be a path-connected set. Suppose that $A \subset S$ is a clopen subset. Suppose there exist points $\mathbf{p}, \mathbf{q} \in S$ such that $\mathbf{p} \in A$ and $\mathbf{q} \notin A$. Since S is path-connected, there exists a continuous function $\mathbf{f} : [0, 1] \rightarrow S$ with $\mathbf{f}(0) = \mathbf{p}$ and $\mathbf{f}(1) = \mathbf{q}$. Then $\mathbf{f}^{-1}(A)$ is a clopen subset of $[0, 1]$ that contains 0 but not 1, contradicting the previous paragraph. \square

In practice, to prove that a property is true at all points of a connected set, it is often convenient to prove that the set of points where the property holds is nonempty, open, and closed. For example:

PROPOSITION A.19.

If $S \subset \mathbb{R}^n$ is connected, and $f : S \rightarrow \mathbb{R}$ is a continuous function that attains only integer values, then f is constant.

PROOF. Let $y_0 \in \mathbb{Z}$ denote an integer value attained by f . The nonempty set $f^{-1}(y_0) = \{x \in S \mid f(x) = y_0\}$ is closed in S because the singleton set $\{y_0\}$ is closed in \mathbb{R} . It is also open in S because it equals $f^{-1}(U)$, where $U = (y_0 - \frac{1}{2}, y_0 + \frac{1}{2})$ is a neighborhood of y_0 in \mathbb{R} that is small enough not to include any other integers. Since S is connected, $f^{-1}(y_0)$ must equal all of S , so f does not attain any other values. \square

Connectedness and path-connectedness are topological notions. In particular, If $S_1 \subset \mathbb{R}^{n_1}$ and $S_2 \subset \mathbb{R}^{n_2}$ are homeomorphic, then either both are path-connected or neither is path-connected. Similarly, either both are connected or neither is connected.

4. Compact Sets

The notion of compactness is fundamental to topology. We begin with the most intuitive definition.

DEFINITION A.20.

*A subset $S \subset \mathbb{R}^n$ is called **bounded** if $S \subset B(\mathbf{p}, r)$ for some $\mathbf{p} \in \mathbb{R}^n$ and some $r > 0$. Further, S is called **compact** if it is closed and bounded.*

Compact sets are those that contain their limit points and lie in a finite chunk of Euclidean space. Unfortunately, this definition is not topological, since “bounded” cannot be defined without referring to the distance function

on \mathbb{R}^n . In particular, boundedness is not preserved by homeomorphisms, since the bounded set $(0, 1)$ is homeomorphic to the unbounded set \mathbb{R} . Nevertheless, compactness is a topological notion, as is shown by the following alternative definition:

DEFINITION A.21.

Let $S \subset \mathbb{R}^n$.

- (1) An **open cover** of S is a collection, \mathcal{O} , of sets that are open in S and whose union equals S .
- (2) S is called **compact** if every open cover, \mathcal{O} , of S has a finite subcover, meaning a finite subcollection $\{U_1, \dots, U_n\} \subset \mathcal{O}$ whose union equals S .

The equivalence of our two definitions of compactness is called the **Heine–Borel theorem**. The easy half of its proof goes like this: Suppose that S is not bounded. Then the collection

$$\{\mathbf{p} \in S \mid \text{dist}(\mathbf{0}, \mathbf{p}) < k\},$$

for $k = 1, 2, 3, \dots$, is an open cover of S with no finite subcover. Next suppose that S is not closed, which means it is missing a limit point $\mathbf{q} \in \mathbb{R}^n$. Then the collection $\{\mathbf{p} \in S \mid \text{dist}(\mathbf{p}, \mathbf{q}) > 1/k\}$, for $k = 1, 2, 3, \dots$, is an open cover of S with no finite subcover.

The other half of the proof is substantially more difficult. We content ourselves with a few examples.

The open interval $(0, 1) \subset \mathbb{R}$ is not compact because it is not closed, or because

$$\mathcal{O} = \{(0, 1/2), (0, 2/3), (0, 3/4), (0, 4/5)\dots\}$$

is an open cover of $(0, 1)$ that has no finite subcover.

The closed interval $[0, 1]$ is compact because it is closed and bounded. It is somewhat difficult to prove directly that every open cover of $[0, 1]$ has a finite subcover; attempting to do so will increase your appreciation of the Heine–Borel theorem.

Our second definition of compactness is topological. Therefore, if $S_1 \subset \mathbb{R}^{n_1}$ and $S_2 \subset \mathbb{R}^{n_2}$ are homeomorphic, then either both are compact or neither is compact.

There is a third useful equivalent characterization of compactness, which depends on the notion of *subconvergence*.

PROPOSITION AND DEFINITION A.22.

- (1) An infinite sequence of points $\{\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3, \dots\}$ in \mathbb{R}^n is said to **subconverge** to $\mathbf{p} \in \mathbb{R}^n$ if there is an infinite subsequence $\{\mathbf{p}_{i_1}, \mathbf{p}_{i_2}, \mathbf{p}_{i_3}, \dots\}$ (with $i_1 < i_2 < i_3 < \dots$) that converges to \mathbf{p} .
- (2) A subset $S \subset \mathbb{R}^n$ is compact if and only if every infinite sequence of points in S subconverges to some $\mathbf{p} \in S$.

For example, the sequence $\{1/2, 2/3, 3/4, \dots\}$ in $S = (0, 1)$ subconverges only to $1 \notin S$, which gives another proof that $(0, 1)$ is not compact. One consequence of this “subconvergence” characterizations of compactness is the following:

PROPOSITION A.23.

If $S \subset \mathbb{R}^n$ is compact and \mathbb{O} is an open cover of S , then there exists a number $\delta > 0$ (called the **Lebesgue number** of \mathbb{O}) such that for any pair of points in S separated by a distance $< \delta$, there exists an open sets from the collection \mathbb{O} that contains both points.

PROOF. Suppose to the contrary that there is no such number δ . Then for each positive integer n , there exist points $p_n, q_n \in S$ with $\text{dist}(p_n, q_n) < \frac{1}{n}$ that do not both belong to any member of \mathbb{O} . The sequences $\{p_n\}$ and $\{q_n\}$ must subconverge to points $p, q \in S$ respectively, and it is straightforward to show that $p = q$. Since \mathbb{O} is an open cover, there exists $U \in \mathbb{O}$ such that $p \in U$. Since U is open, both subsequences are eventually inside U , contradicting the fact that for every n , p_n and q_n do not both belong to any single member of \mathbb{O} . \square

The next proposition says that the continuous image of a compact set is compact.

PROPOSITION A.24.

Let $S \subset \mathbb{R}^n$. Let $\mathbf{f} : S \rightarrow \mathbb{R}^m$ be continuous. If S is compact, then the image, $\mathbf{f}(S)$, is compact.

PROOF. The function \mathbf{f} is also continuous when regarded as a function from S to $\mathbf{f}(S)$. Let \mathbb{O} be an open cover of $\mathbf{f}(S)$. Then $\mathbf{f}^{-1}(U)$ is open in S for every $U \in \mathbb{O}$, so $\mathbf{f}^{-1}(\mathbb{O}) := \{\mathbf{f}^{-1}(U) \mid U \in \mathbb{O}\}$ is an open cover of S . Since S is compact, there exists a finite subcover $\{\mathbf{f}^{-1}(U_1), \dots, \mathbf{f}^{-1}(U_k)\}$ of $\mathbf{f}^{-1}(\mathbb{O})$. It is straightforward to check that $\{U_1, U_2, \dots, U_k\}$ is a finite subcover of \mathbb{O} . \square

COROLLARY A.25.

If $S \subset \mathbb{R}^n$ is compact and $f : S \rightarrow \mathbb{R}$ is continuous, then f attains its supremum and infimum.

The conclusion that f attains its supremum means two things. First, the supremum of $f(S)$ is finite (because $f(S)$ is bounded). Second, there is a point $\mathbf{p} \in S$ for which $f(\mathbf{p})$ equals this supremum (because $f(S)$ is closed).

Recommended Excursions

1. C. Adams, *The Knot Book*, American Mathematical Society, 2004.
2. M. Beeson, *Notes on Minimal Surfaces*, preprint, 2007,
<http://michaelbeeson.com/research/papers/IntroMinimal.pdf>
3. V. Blåsjö, *The Isoperimetric Problem*, American Mathematical Monthly. **112**, No. 6 (2005), 526–566.
4. D. DeTurck, H. Gluck, D. Pomerleano, and D. Shea Vick, *The Four Vertex Theorem and Its Converse*, Notices of the AMS. **54**, No. 2 (2007), 192–206.
5. R. Osserman, *A Survey of Minimal Surfaces*, Dover (1986).
6. S. Sawin, *South Point Chariot: An Invitation to Differential Geometry*, preprint, 2015.
7. D. Sobel, *Longitude: The True Story of a Lone Genius Who Solved the Greatest Scientific Problem of His Time*, Walker Books, 2007.

Image Credits

- Chapter 1 title photo, page 1: *Like pearls on a string*, Susanne Nilsson, Creative Commons License 2.0,
<https://www.flickr.com/photos/infomastern/22687968229/>
- Figure 1.5 (right), page 7: *A public address horn speaker in a train station in Smíchov, Prague, the Czech Republic*, Wikimedia user: ŠJů,
https://en.wikipedia.org/wiki/Horn_loudspeaker
- Figure 1.25, page 39: *South-pointing chariot photo*, Property of Maker Media, Inc., used with permission.
[http://makezine.com/2010/01/19/
toys-tricks-teasers-the-chinese-south-pointing-chariot/](http://makezine.com/2010/01/19/toys-tricks-teasers-the-chinese-south-pointing-chariot/)
- Page 41: *Illustration of the right-hand rule for the cross product*, Wikimedia user Acdx, Creative Commons License 3.0,
[https://commons.wikimedia.org/wiki/File:Right_hand_rule_cross_product
.svg](https://commons.wikimedia.org/wiki/File:Right_hand_rule_cross_product.svg)
- Figure 1.31, page 53: derivative of *Image of a chiral object and its reflection in a mirror*, Johannes Rössel, public domain,
<https://commons.wikimedia.org/wiki/File:Chirality.svg>
- Chapter 2 title photo, page 62: *Ein REISS-Kompensations-Polarplanimeter 3003 von 1958 (VEB Mess- und Zeichnungsgerätebau Bad Liebenwerda)*, Wikimedia user Moorochs, public domain,
[https://commons.wikimedia.org/wiki/File:Polarplanimeter-30-08-2009-055
.png](https://commons.wikimedia.org/wiki/File:Polarplanimeter-30-08-2009-055.png)
- Image 2.39 (right), page 109: *Reflexion of a plane wave by a cylindrical mirror with circular base. Image realized with POVRay*, Wikimedia user Jean-Michel Courty commonswiki, Creative Commons License 3.0,
<https://commons.wikimedia.org/wiki/File:Miroir-cercle.jpg>

- Chapter 3 title photo, page 113: *a scan of a work published originally in 1613 (François d'Aiguillon “Opticorum libri sex philosophis juxta ac mathematicis utiles” (Six Books of Optics, useful for philosophers and mathematicians alike), Anvers, 1613) by Rubens*, public domain, <https://commons.wikimedia.org/wiki/File:RubensAguilonStereographic.jpg>
- Figure 3.24 (left), page 156: *A photograph of a green paper Möbius strip*, David Benbennick, Creative Commons License 3.0, https://commons.wikimedia.org/wiki/File:M%C3%B6bius_strip.jpg
- Figure 3.24 (right), page 156: *Möbius strip*, Wikimedia user JoshDif, Creative Commons License 3.0, <https://commons.wikimedia.org/wiki/File:MobiusJoshDif.jpg>
- Figures 3.31 and 3.32, page 173: derivatives of *Archimedes sphere and cylinder*, Wikimedia user Pbroks13, Creative Commons License 2.5, https://commons.wikimedia.org/wiki/File:Archimedes_sphere_and_cylinder.svg
- Figure 3.33 (top), page 174: *cylindrical equal-area projection*, Wikimedia user KoenB, public domain, <https://commons.wikimedia.org/wiki/File:Cilinderprojectie-constructie.jpg>
- Figure 3.33 (middle), page 174: *Map of the world in a Lambert cylindrical equal-area projection with Tissot’s Indicatrix of deformation*, Eric Gaba—Wikimedia Commons user: Sting, Creative Commons License 2.0, https://commons.wikimedia.org/wiki/File:Tissot_indicatrix_world_map_Lambert_cyl_equal-area_proj.svg
- Figure 3.33, page 174 (bottom): *Map of the world in a Gall-Peters cylindrical equal-area projection also known as Gall’s orthographic or Peters, with Tissot’s Indicatrices of deformation*, Eric Gaba—Wikimedia Commons user: Sting, Creative Commons License 2.0, https://commons.wikimedia.org/wiki/File:Tissot_indicatrix_world_map_Gall-Peters_equal-area_proj.svg
- Figure 3.36 (left), page 177: see above.
- Figure 3.36 (right), page 177: *Polar stereographic projection*, Lars H. Rohwedder—Wikimedia user:RokerHRO, Creative Commons License 3.0, https://commons.wikimedia.org/wiki/File:Stereographic_Projection_Polar.jpg
- Figure 3.37, page 178: *a Mercator projection map with Tissot’s indicatrices*, Stefan Kühn, Creative Commons License 3.0, https://commons.wikimedia.org/wiki/File:Tissot_mercator.png
- Chapter 4 title photo, page 194: used with permission of the Children’s Museum of Virginia, Portsmouth, photographer: Jeffrey Ringer with Arengue Branding, <https://plus.google.com/u/0/+ChildrensMuseumofVirginiaPortsmouth/photos>

- Figure 4.14 (right), page 226: *Dog looking at and listening to a phonograph*, Francis Barraud, 1895, public domain,
https://en.wikipedia.org/wiki/Horn_loudspeaker#/media/File:OriginalNipper.jpg
- Figure 4.19 (left), page 231: *Photo personnelle destinée à illustrer l'article hélicoïde en même temps qu'une propriété physique des bulles de savon*, Wikimedia user: Blinking Spirit, Creative Commons License 1.2,
https://commons.wikimedia.org/wiki/File:Bulle_de_savon_h%C3%A9lico%C3%AFde.PNG
- Figure 4.19 (right), page 231: *Photo d'un film de savon matérialisant une caténoïde*, Wikimedia user: Blinking Spirit, Creative Commons License 1.0,
https://commons.wikimedia.org/wiki/File:Bulle_cat%C3%A9noide.PNG
- Chapter 5 title photo, page 248: *Péndulo de Foucault situado en la Ciudad de las Artes y de las Ciencias de Valencia*, Wikimedia user: Daniel Sancho, Creative Commons License 2.0,
https://commons.wikimedia.org/wiki/File:Foucault_pendulum_closeup.jpg
- Figure 5.17 (left), page 286: see above.
- Figure 5.17, page 286: *Foucault pendulum at North Pole*, Wikimedia user: en:User:Krallja, Creative Commons License 3.0,
https://commons.wikimedia.org/wiki/File:Foucault_pendulum_at_north_pole_accurate.PNG
- Chapter 6 title photo: *German banknote of the fourth series (since 1989/90)*, public domain,
https://commons.wikimedia.org/wiki/File:10_DM_Serie4_Vorderseite.jpg
- Figure 6.18 (left), page 344: *Costa's minimal surface, a three ended, complete embedded minimal surface. Rendered from a polygon mesh from the Scientific Graphics Project at MSRI*, Anders Sandberg, Creative Commons License 3.0,
https://commons.wikimedia.org/wiki/File:Costa%27s_Minimal_Surface.png
- Figure 6.18 (right), page 344: *Approximation of the Schwarz P minimal surface generated using Ken Brakke's Surface Evolver program*, Anders Sandberg, Creative Commons License 3.0,
https://commons.wikimedia.org/wiki/File:Schwarz_P_Surface.png

Index

- acceleration function, 16
angle displacement, 283
angle function, 35, 63, 283
antipodal map, 53
arc length, 4
Archimedes's theorem, 171
area, 161
area distortion, 147
 infinitesimal, 164
astroid, 8, 110
asymptotic, 212
atlas, 125

ball, 346
Bonnet's theorem, 308
boundary component, 320
boundary point, 346
bounded, 353

Catalan's surface, 235
catenary, 8
catenoid, 232
Cauchy sequence, 349
 d -, 275
chain rule, 119
 for surfaces, 146
Christoffel symbols, 289
circulation, 83
Clairaut's Theorem, 251
closed in, 349
closed set, 346
compact, 353
complete surface, 276
component, 10
component functions, 2, 82, 113

cone, 137
conformal, 173
conjugate point, 315
connected, 352
conservation of energy, 88
conservative vector field, 86
continuous, 350
convergence, 348
convex
 curve, 72, 78
 neighborhood, 265
 surface, 217, 344
coordinate chart, 125
corner, 68
Costa surface, 344
covariant derivative, 280
critical point, 123, 146, 205
cross product, 41
curvature
 Gaussian, 196
 geodesic, 209
 mean, 196
 minimal, 168
 normal, 206
 of a curve, 26
 signed, 33
curve, 2
 parametrized, 2
 piecewise-regular, 68
 piecewise-regular in a surface, 298
 regular, 5
 in a surface, 141
 simple closed, 21
cusp, 69, 300

- cycloid, 101, 103
 - generalized, 109
- cylinder, 128
 - generalized, 136, 343
- deformation, 228
- deltoid curve, 8
- derivative
 - of a curve, 3
 - of a function between Euclidean spaces, 116
 - of a function between surfaces, 143
- diameter, 279, 308
- diffeomorphic, 122
- diffeomorphism, 122
 - local, 169
- Dini's surface, 227
- directional derivative, 86, 115
- distance circle, 271
- distance function
 - intrinsic, 254
 - on \mathbb{R}^n , 345
- divergence, 94
- edge, 327
 - interior/exterior, 328
- ellipse, 72, 77, 101, 110
- Enneper's surface, 139, 182, 225, 235
- envelope, 108
- epicycloid, 110
- epitrochoid, 8, 40
- equiareal, 170
- escape velocity, 96
- Euclidean space, \mathbb{R}^n , 1
- Euler characteristic, 327
- evolute, 30, 106
- exponential map, 258
- fake sphere, 222
- Fary–Milnor theorem, 238
- Fenchel's theorem, 79, 237
- Fermi coordinates, 295
- first fundamental form, 166
 - in local coordinates, 183
- flux, 94
- Foucault's pendulum, 284
- four vertex theorem, 72
- Frenet equations, 45
- Frenet frame, 42
- Gall–Peters map, 173
- Gauss map, 196
- Gauss's lemma, 260
 - generalized, 317
- Gauss's Theorema Egregium, 269
- Gauss–Bonnet theorem
 - for compact surfaces, 334
 - for small geodesic triangles, 297
 - global, 329
 - local, 321
- genus, 334
- geodesic, 248
 - equations, 290
 - variational characterization, 305
- geodesic curvature, 248
- geodesic triangle, 325
- gnomonic projection, 181
- gradient, 86, 160
- Gram–Schmidt process, 16
- gravitational force, 89
- great circle, 79, 249
- Green's theorem, 91
 - flux version, 95
- hairy ball theorem, 339
- height function, 76, 240
- Heine–Borel theorem, 354
- helicoid, 186, 225, 233
 - generalized, 226
- helix, 2, 186, 250
- Hessian, 124, 146, 205, 224
- Hilbert's theorem, 342
- holonomy, 283, 298
- homeomorphism, 351
- homogeneity, 273
- Hopf's Umlaufsatz, 62
 - generalized, 69
- Hopf–Rinow theorem, 277
- hyperboloid, 134
- hypocycloid, 110
- index, 336, 337
- infinitesimal area distortion, 161
- infinitesimal circulation, 90
- inner product, 9
- interior angle, 71, 325
- interval, 2
- intrinsic, 167, 183, 184
 - covariant differentiation, 287
 - distance function, 254
 - Gaussian curvature, 269
 - geodesic curvature, 266
 - geodesics, 266
 - Jacobi fields, 318
 - minimal curvature is not, 168
 - volume is not, 169
- inverse function theorem
 - for Euclidean space, 120
 - for surfaces, 143

- involute, 105
- isometric, 167
- isometry, 167
 - local, 169
- isoperimetric inequality, 98
- Jacobi equation, 312
- Jacobi field, 309
- Jacobi's theorem, 318, 332
- Jacobian matrix, 117
- Jordan curve theorem, 62
- kinetic energy, 85
- knotted, 238
- Lambert projection, 173
- latitude, 132
- Lebesgue number, 355
- lemniscate of Bernoulli, 8
- level set, 134
- limit point, 348
- line integral, 83
- line of curvature, 211, 224
- Lissajous curve, 40
- local coordinates, 127
- locally isometric, 273
- logarithmic spiral, 6, 15
- longitude, 132
- longitude problem, 101
- loxodrome, 182
- Möbius strip, 156, 188, 225
- mean curvature, 196
 - field, 230
 - flow, 232
- Mercator projection, 177
- minimal surface, 231
 - conjugate pair, 235
- minimizing, 255
 - property of geodesics, 263
- monkey saddle, 217, 226, 275
- norm, 4
- normal coordinates, 260
- normal field, 152
- normal neighborhood, 159, 259
- normal polar coordinates, 260
- normal section, 208
- normal vector, 151
- open cover, 354
- open in, 349
- open set, 346
- orientation
 - of a basis of \mathbb{R}^n , 53
 - of a curve, 20, 24
- of a plane in \mathbb{R}^3 , 149
- of a simple closed curve, 62
- of a surface, 153
- orientation-preserving
 - diffeomorphism between surfaces, 155
 - linear transformation, 150
 - reparametrization, 20
- orthogonal
 - matrix, 49
 - vectors, 9
- orthonormal, 11
- osculating circle, 30
- osculating plane, 29
- parallel, 9
- parallel transport, 283, 298
- partial derivative, 114
- path-connected, 352
- path-independent, 88
- permutation matrix, 58
- planimeter, 61, 97, 101
- Plateau's problem, 231
- Poincaré's theorem, 338
- polygonal region, 161, 320
- potential function, 86
- preimage, 133
- principal curvatures, 206
- principal directions, 206
- projection, 10, 12
- pseudosphere, 225
 - twisted, 227
- quadratic form, 203
- ray, 279
- rectifying plane, 48
- region, 320
- regular region, 320
- regular value, 133
- reparametrization
 - of a closed curve, 22
 - of a curve, 19
- Reuleaux triangle, 41
- rhumb line, 182
- rigid motion, 49
 - proper/improper, 52
- rotation index, 37, 40
- ruled parametrized surface, 140
- Scherk's surface, 235
- Schwarz inequality, 9
- Schwarz surface, 344
- second fundamental form, 206
 - in local coordinates, 218
- self-adjoint, 202

- signed angle, 68, 299
- simple region, 329
- singular point, 336
- skew-symmetric matrix, 59
- smooth
 - multivariable, 114
 - on an arbitrary set, 121
 - single variable, 2
- south-pointing chariot, 38, 305
- spanning disk, 238
- speed, 4
- sphere, S^n , 26
- spherical cap, 165
- spherical coordinates, 129
- stereographic projection, 175
- subconvergence, 355
- surface
 - of revolution, 131, 182, 186, 188, 221, 226, 251
 - generalized, 226
 - oriented/orientable, 153
 - parametrized, 135
 - regular, 125
 - ruled, 139
- surface patch, 125
- tangent developable, 138, 188, 225
- tangent field, 152
- tangent plane, 141
- tautochrone clock, 102
- tautochrone curve, 102
- tire tracks, 37, 303
- topology, 348
- toroidal spiral, 8
- torsion, 43
- torus, 138
- torus knot, 244
- total curvature, 78
- total geodesic curvature, 321
- trace, 18
- tractrix, 6, 225
- transition map, 162
- translation, 52
- trefoil knot, 8
- triangle, 327
- triangle inequality, 254, 345
- triangulation, 327
 - generalized, 333
- tubular neighborhood, 66, 124, 238
- twisted cubic, 8
- twisted sphere, 227
- umbilical point, 206
- unit binormal vector, 42
- unit normal vector, 27, 42
- unit tangent vector, 27, 42
- unknotted, 238
- variation, 303
 - geodesic, 309
 - of arc length, 304, 306
- vector field
 - along a curve, 280
 - along a piecewise-regular curve, 298
 - on \mathbb{R}^n , 82
 - on a surface, 152
 - parallel, 280
- velocity function, 16
- vertex, 72, 299
- Weingarten map, 196
- work, 83