Geometric Modeling with Splines

An Introduction

Elaine Cohen
Richard F. Riesenfeld
Gershon Elber

CRC Press
Taylor & Francis Group
AN A K PETERS BOOK

Geometric Modeling with Splines:
An Introduction

Geometric Modeling with Splines: An Introduction

Elaine Cohen
University of Utah
School of Computing

Richard F. Riesenfeld
University of Utah
School of Computing

Gershon Elber
Technion
Israel Institute of Technology

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
AN A K PETERS BOOK

First published 2001 by A K Peters, Ltd.

Published 2018 by CRC Press
Taylor & Francis Group
6000 Broken Sound Parkway NW, Suite 300
Boca Raton, FL 33487-2742

CRC Press is an imprint of Taylor & Francis Group, an Informa business
No claim to original U.S. Government works

ISBN-13: 978-1-56881-137-6 (hbk)

Visit the Taylor & Francis Web site at
http://www.taylorandfrancis.com

and the CRC Press Web site at
http://www.crcpress.com

Library of Congress Cataloging-in-Publication Data

Cohen, Elaine, 1946–
Geometric modeling with splines : an introduction / Elaine Cohen, Richard F. Riesenfeld, Gershon Elber.
p. cm.
Includes bibliographical references and index.
ISBN 1-56881-137-3
1. Curves on surfaces–Mathematical models. 2. Surfaces–Mathematical models. 3. Spline theory. 4. Computer-aided design. I. Riesenfeld, Richard F. II. Elber, Gershon, 1960– III. Title.

QA565 .C656 2001
516.3′52–dc21 2001032921

Dedicated

to Pierre Bézier, Steven A. Coons, I. J. Schoenberg, and Ivan E. Sutherland, the giants on whose shoulders we stand. Their diverse, fundamental works and powerful visions provided the critical elements that have been forged into this increasingly important, multi-disciplinary field.

Foreword

Geometric modelling with splines has been an important and exciting field for many years covering numerous applications. The subject draws on various topics from mathematical approximation theory, numerical analysis, classical and discrete geometry, engineering, and computer science.

The field grew out of pioneering work in the 1960s on modelling of complex objects like ship hulls and car bodies. At that time, the topic of spline functions developed into an active area of research in approximation theory due to the fundamental work of Schoenberg. His work on B-splines opened new perspectives, and it gradually became clear that these functions were well suited for geometric modelling of physical objects. The seminal work of one of the authors played a crucial role in this development.

Geometric modelling with splines has been a significant area of research for almost 40 years with applications ranging from animated films to simulated surgery. Although many of the fundamental mathematical results have been established, the field is still burgeoning due to a continuous need for new techniques. This book is a welcome text and is written by well-known experts in the field. The authors, and their many accomplished students, have significantly influenced the advancement of this subject. Combining mathematical rigor and the science of modelling in a fruitful way, the book is well-suited as a textbook for a course on this topic, per se, or one that draws on aspects of this material.

The text contains a comprehensive treatment of curves and surfaces with emphasis on B-spline techniques. In addition the book contains a wealth of material ranging from classical techniques to a broad coverage of more specialized topics including new techniques like subdivision of surfaces, all of which add to its value as a reference for professionals and researchers working in the field.

The authors are to be congratulated for writing such a comprehensive text.

Tom Lyche
University of Oslo

Preface

This book has evolved out of lecture notes created to introduce students to various aspects of geometric modeling with splines. The shape-mimicking properties of the NURBS control polygon allows the user to create a smooth curve by manipulating a simple polyline.

While B-spline and NURBS mathematics can seem unnecessarily abstract to students, the mathematical formalisms effectively camouflage some rather formidable machinery that allows geometric shapes to be expressed and manipulated through what appears as a rather straightforward and intuitive geometric design scheme. The constructive nature of many computational B-spline algorithms can enhance a student's intuitive sense for the shape properties of the curve under design. Many algorithms for representation, computation, and querying of B-spline models can be implemented as intuitive, efficient algorithms executing at interactive rates. In light of these characteristics, we have stressed the mathematical soundness of spline methods. We present algorithms, and in some cases, pseudo-code. Since many of the surface methods, i.e., representations and algorithms, rely on curve algorithms and properties, the earlier chapters of the book stress rigorously establishing the properties that will be used throughout. Generally, later chapters address material based on schemes described in the earlier chapters.

The goal of the book is to act both as a text and a reference book. We believe the breadth and depth of the included material are sufficient to give the reader a background suitable for implementing splines and for designing with splines. In addition, the reader who attains a solid understanding of the underlying mathematical approaches, concepts, and logic, as well as a practical understanding, will have a sufficient background to conduct geometric modeling research using splines.

The book is structured as follows:

- A background review of mathematics used in this book is presented in Chapter 1. Chapter 2 presents an overview of the most common types of representations and their characteristics.

- It is in Chapter 3 that we begin the discussion of curve forms for geometric modeling with the oldest representation: conic sections. We show the equivalences of various traditional definitions, the ones most likely to have been seen by readers through the middle of the undergraduate college years. Then, a constructive geometric approach is presented. As the sections progress, we develop this approach into the typical *blending* formulation, and then derive a parametric *blending* representation, which enables writing a curve as a convex combination of geometric points. The last formulation allows us to develop curve properties which will recur in the Bézier formulation (Chapter 5) and the B-spline representation (Chapter 6). Subdivision algorithms are first presented with respect to conic sections in Chapter 3.

- Elements of differential geometry for curves, a branch of mathematics concerned with characterizing the behavior (and shape) of a parametric curve by its differential properties, are presented in Chapter 4. Concepts that recur throughout geometric modeling, including *regularity*, curvature, torsion, and Frenet equations are covered. In *bottom up* design, it is sometimes necessary to piece together pre-designed curves. We discuss methods for determining when a compound curve exhibits various types of parametric smoothness.

- The constructive approach first introduced with conics is generalized to the constructive approximation curves in Chapter 5. Then it is shown that these curves are Bézier curves, and we treat many characteristics for Bézier curve. We also introduce the idea of subdivision of Bézier curves by developing a special case algorithm for subdivision at a curve's midpoint. The Bernstein blending functions are the blending functions used in the Bézier method. Having developed formulations for the blending functions, we discuss using them for approximation and interpolation, and relate approximating a pre-existing parametric function by a Bézier curve to the Bernstein approximation method from classical approximation theory. Finally we apply results from Chapter 4 to discuss smoothly piecing together Bézier curves.

- In Chapter 6 the constructive approximation curves, or Bézier curves, are further generalized to piecewise smooth constructive approximation curves, which are then shown to be equivalent to B-spline curves. Basic properties of B-spline curves are shown. Chapters 7 and 8 reveal more characteristics and properties of spline spaces. Proving the representational power of splines and showing the linear independence of B-splines, Chapter 7 develops the more abstract properties. This is accomplished using inductive proofs like those used in Chapter 3. The idea of refinement is introduced in its simplest form, that of knot insertion of a single knot. This result is further developed for quadratic and cubic subdivision curves, which in Chapter 20 are shown to be the basis for Doo-Sabin and Catmull-Clark surfaces, respectively. In Chapter 8, frequently occurring knot vector configurations for spline curves are presented and their effects on curve shape are discussed. Rational spline curves are finally taken on, as are methods for their computation.

- Various forms of interpolation and approximation with splines are presented in Chapter 9. Two widely used interpolation methods, nodal and complete spline interpolation, are covered. In addition to basic discrete and continuous least squares approximation, the Schoenberg variation diminishing spline, the abstract quasi-interpolation method, and a more widely employed multiresolution constrained decomposition method are treated.

- Chapter 10 is devoted to various types of interpolation using classical polynomial bases.

- In Chapter 11 we present other derivations of B-splines to give a flavor of the origins of B-spline methods. One derivation manifests splines as shadows of higher-dimensional simplices, while another uses generalizations of divided differences to higher dimensions. These can be shown to be equivalent. A third view develops B-splines in terms of signal processing and filtering. The last approach, the original one presented by Schoenberg many decades ago, is once again becoming topical.

- The section on surfaces starts in Chapter 12 with differential geometry for surfaces. Using the results of Chapter 4 for curves, this chapter develops formulations for first and second fundamental forms, normal and geodesic curvation, principal curvatures, and principal directions. Surface shape analysis for design depends on such variables.

- Chapter 13 defines a tensor product surface and presents methods for evaluating position and derivatives using properties of curve forms. Emphasis is placed on B-spline and Bézier tensor product surfaces. Matrix methods for transforming between various bases are described. Based on their tensor product structure, surface forms are generalized from quadratic and cubic subdivision curves schemes.

- In Chapter 14, the ideas presented in Chapter 9 for fitting curves to data are generalized to surfaces. The classis Coons surface is detailed. Finally, an *operator* approach for transforming methods for fitting curves into methods for fitting surfaces is described.

- In Chapter 15, practical aspects of actually creating representations for specific surfaces are confronted. Methods to create representations for ruled surfaces and various surfaces of revolution are given.

- Starting with Chapter 16, more advanced techniques are taken on. These approaches and methods are necessary to create, manipulate, render, query, and fabricate B-spline representations for the complex shapes needed in applications ranging from animation to solid modeling.

- In Chapter 16 general algorithms for subdivision and refinement for B-splines are developed as well as specialized algorithms for subdivision of Bézier curves and surfaces. Pseudo-code clarifies the algorithms. Chapter 17 presents algorithms and pseudo-code of methods, based on the refinement approach, for rendering, computing intersections, and adding degrees of freedom to support hierarchical top-down design.

- It is rare that a single surface can be used to model a complicated object. In earlier chapters we presented methods for piecing together curves and surfaces. Chapter 18 includes issues that arise when attempting to combine arbitrary pieces of tensor product surfaces to define complex models. The important topic of finding curve and surface intersections is introduced. We give methods for finding such intersections, as well as criteria for determining convergence. The idea of a well-formed 3-D model is developed. Detailed algorithms for constructing a model by applying Boolean operations on existing models are described for planar polygonal models. We then show how algorithms for polyhedral models are hierarchical and rely on the results from the planar cases.

- Out of the discussion of creating actual models comes an important realization that we need data structures to traverse models. In Chapter 19 we present the winged-edge data structure, a well-known and widely used topological representation, and then discuss a data structure suitable for models bounded by trimmed surfaces, the kind that result from Boolean operations on models bounded by sculptured surfaces.

- In Chapters 20 and 21 we describe generalizations of B-splines that are just starting to be more widely taught and adopted. Chapter 20 is focused on subdivision surfaces. We show that the *template* algorithms for generating the refined meshes are generalizations of spline refinement algorithms for Catmull-Clark subdivision surfaces (a generalization of bi-cubic uniform floating spline surfaces), Doo-Sabin subdivision surfaces (a generalization of bi-quadratic uniform floating spline surfaces), and Loop subdivision surfaces (a generalization of box spline surfaces and refinement algorithms). Chapter 21 is focused on algorithms and uses for trivariate volumetric splines.

Versions of this manuscript have been used in teaching quarter, semester, and year long course sequences. Advanced undergraduate and beginning graduate students with good backgrounds in mathematics, i.e., advanced calculus and matrix algebra, and programming experience have been typical participants in these classes. The material can be taught with emphasis on the behaviors, algorithms, and implementations of splines in geometric modeling schemes, or with an emphasis on the proofs and proper mathematical development of the subject matter, depending on the goals of the course at hand.

If Chapters 1 and 2 are considered as background material, Chapters 3, 4, 5, 6, 7 (without Section 7.4), 8, 12, and 13 (without Sections 13.5 and 13.6) could form the basis for a one semester introductory class. A two semester class would include Chapters 9, 14, 15, 16, 17, as well as a selected subset of Chapters 10, 18, 19, 20, and 21, depending on the interests of the class.

We have tried to make this book broad enough to be appealing for many readers and many class situations. While some topics rely on earlier material, other chapters can be read directly. Many variations are possible in choosing a rewarding path though this book. Although considerable effort has been devoted to "debug" the text, errors will inevitably turn up. We will try to provide corrections on a web site as we become aware of any inaccuracies, so check our personal web pages for the most current information. (`www.cs.utah.edu/~cohen` or `www.cs.utah.edu/~rfr`)

This field has brought us into contact with many challenging and important problems, and provided a rich area of research. We are hopeful that the perspective and understanding that we have gained over many years contributes to making this a good book for others to learn the fundamentals more quickly, or to refer to for specific information while engaged in the subject.

Acknowledgments

This book has grown out of lecture notes from an advanced undergraduate and graduate course taught repeatedly both at the University of Utah and the Technion. Over the years, numerous students, staff, and colleagues at Utah and the Technion have stimulated and challenged our understanding of fundamental issues, particularly the members of Utah's Alpha_1 Research Group. Interacting with them has provided a daily education leading to many advancements of understanding reflected in this volume. Thanks are due to all those students at Utah and the Technion who provided critical responses to various earlier stages of the book. Their comments have led to significant improvements in organization and detail. We are particularly grateful to Bill Martin for his enormous effort in providing detailed proofreading of the material, and to Matt Kaplan for making extensive contributions to the figures.

This volume would not be going to press on schedule without extraordinary efforts of our wonderful publisher Alice Peters, who worked long days, weekends, and holidays in order to meet critical publication deadlines. We have felt fortunate to be working with her.

Mentors and teachers, Steve Coons, Bill Gordon, Robin Forrest, Pierre Bzier, Charles Lang, and Ivan Sutherland, have inspired us directly and indirectly. Finally, we wish to thank our colleague Tom Lyche at the University of Oslo, with whom two of us have collaborated periodically over the last 20 years, for all we have learned from him about the art of splines. He has regularly encouraged this book's completion, and helped with reading of an early draft.

Cohen and Riesenfeld express appreciation to the National Science Foundation for The Science and Technology Center for Computer Graphics and Visualization, which has provided a rich intellectual environment.

Elaine Cohen
Richard F. Riesenfeld
University of Utah

Gershon Elber
Technion
Israel Institute of Technology

June 2001

Contents

I

Introduction

1

Review of Basic Concepts

In the sections that follow we provide a short presentation of some of the basic material which will be needed in various other chapters of the book. If the reader is familiar with the contents of some sections, he may prefer to skip those sections.

1.1 Vector Analysis

We shall provide a brief review of some necessary concepts and manipulation techniques. Many of these concepts are general and not dependent on any particular vector space. The concept of cross product, however is defined only in $\mathbb{R}^3$.

Definition 1.1. *A* vector space $\boldsymbol{V}$ *is defined over a set of elements, the vectors, that have two operations, addition* $+ : \boldsymbol{V} \times \boldsymbol{V} \to \boldsymbol{V}$*, and scalar multiplication* $\cdot : \mathbb{R}^1 \times \boldsymbol{V} \to \boldsymbol{V}$ *which satisfy the following rules:*

If $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{w} \in \boldsymbol{V}$ *and if* r, $s \in \mathbb{R}^1$, *then*

1. $r\boldsymbol{u} + s\boldsymbol{v} \in \boldsymbol{V}$;
2. *There is an element* $\mathbf{0} \in \boldsymbol{V}$ *such that for all* $\boldsymbol{v} \in \boldsymbol{V}$, $\mathbf{0}+\boldsymbol{v} = \boldsymbol{v}+\mathbf{0} = \boldsymbol{v}$;
3. $r(\boldsymbol{u} + \boldsymbol{v}) = r\boldsymbol{u} + r\boldsymbol{v}$;
4. $(r + s)\boldsymbol{u} = r\boldsymbol{u} + s\boldsymbol{u}$.

Definition 1.2. *A finite subset of vectors $\boldsymbol{C}$ in $\boldsymbol{V}$ is called* independent *if for every choice of n less than or equal to the number of elements of $\boldsymbol{C}$, and for all arbitrary choices of $r_1, \ldots, r_n \in \boldsymbol{R}^1$ and $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \boldsymbol{C}$ then*

$$r_1\boldsymbol{v_1} + \cdots + r_n\boldsymbol{v_n} = \boldsymbol{0}$$

implies that

$$r_1 = \cdots = r_n = 0.$$

This states that no element of the set $\boldsymbol{I}$ can be written as a finite linear combination of other elements of the set, i.e., it cannot depend on a finite number of the other elements.

Definition 1.3. *Let $\boldsymbol{S}$ be a subset of vectors of $\boldsymbol{V}$. The* span *of $\boldsymbol{S}$, written span $\boldsymbol{S}$, is the set of all finite linear combinations of elements of $\boldsymbol{S}$. That is, for an arbitrary integer $n > 0$, select n arbitrary vectors $\boldsymbol{v}_1, \ldots, \boldsymbol{v}_n \in \boldsymbol{S}$ and n arbitrary values $r_1, \ldots, r_n \in \boldsymbol{R}^1$ then $r_1\boldsymbol{v_1} + \cdots + r_n\boldsymbol{v_n} \in$ span $\boldsymbol{S}$.*

Example 1.4. For vector space $\boldsymbol{R}^2$, and $\boldsymbol{I_1} = \{(1,0),(1,1)\}$, $\boldsymbol{I_1}$ is independent, but $\boldsymbol{I_2} = \{(1,0),(3,0)\}$ is not since $(3,0) = 3(1,0)$. The span $\boldsymbol{I_1} = \boldsymbol{R}^2$, but span $\boldsymbol{I_2} = \{\, r(1,0) : r \in \boldsymbol{R}^1 \,\}$. □

Definition 1.5. *A* basis $\boldsymbol{B}$ *for a vector space $\boldsymbol{V}$ is a set of vectors that is independent and such that span $\boldsymbol{B} = \boldsymbol{V}$.*

It can be shown that all bases of the same vector space have the same number of elements. For infinite dimensional vector spaces, one must use techniques which show equivalence of the *size* of the infinity. It is left as an exercise for the reader to show that this is true for finite dimensional vector spaces.

Definition 1.6. *If $\boldsymbol{B}$ is a basis for $\boldsymbol{V}$ and has a finite number of elements, we say that $\boldsymbol{V}$ is a* finite dimensional vector space *with dimension equal to the number of elements of $\boldsymbol{B}$.*

Example 1.7. Examples of Vector Spaces:

1. $\boldsymbol{R}^1$, $\boldsymbol{R}^2$, $\boldsymbol{R}^3$.

2. Function space: polynomials of degree 1, 2, ..., n.

3. $C^{(0)}[a,b]$ the set of continuous functions on the interval $[a,b]$. Addition is standard function addition, scalar multiplication is standard multiplication of the function value at the value of x. It can easily be shown that the rest of the properties follow.

4. 2×2 matrices. □

Definition 1.8. *A (real)* inner product space *is a vector space with a second vector operation, $<,>: \boldsymbol{V} \times \boldsymbol{V} \to \boldsymbol{R}^1$, defined such that the following holds true for $\boldsymbol{u}, \boldsymbol{v} \in \boldsymbol{V}$ and $r, s \in \boldsymbol{R}^1$:*

1. $\langle r\boldsymbol{u}, s\boldsymbol{v}\rangle = rs\langle \boldsymbol{u}, \boldsymbol{v}\rangle$;
2. $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = \langle \boldsymbol{v}, \boldsymbol{u}\rangle$;
3. $\langle \boldsymbol{u}+\boldsymbol{v}, \boldsymbol{w}\rangle = \langle \boldsymbol{u}, \boldsymbol{w}\rangle + \langle \boldsymbol{v}, \boldsymbol{w}\rangle$;
4. $\langle \boldsymbol{v}, \boldsymbol{v}\rangle \geq 0$, *and* $= 0$ *if and only if* $\boldsymbol{v} = 0$.

The idea of length or magnitude of a vector can be introduced into an inner product space and a *norm* can be defined.

Definition 1.9. *The* length *or* magnitude *of a vector $\boldsymbol{v} \in \boldsymbol{V}$ is defined as $\|\boldsymbol{v}\| = \sqrt{\langle \boldsymbol{v}, \boldsymbol{v}\rangle}$. The* distance *between two vectors $\boldsymbol{u}$, $\boldsymbol{v}$ is defined as the magnitude of the difference vector, that is, $\|\boldsymbol{u} - \boldsymbol{v}\|$.*

Definition 1.10. *Some further definitions which are based on the inner product and have a geometric interpretation when applied to $\boldsymbol{R}^2$ and $\boldsymbol{R}^3$ are:*

1. *Vectors $\boldsymbol{u}$ and $\boldsymbol{v}$ are said to be* orthogonal *if for $\boldsymbol{u} \neq \boldsymbol{v}$, $\langle \boldsymbol{u}, \boldsymbol{v}\rangle = 0$;*
2. *A vector $\boldsymbol{u}$ is* a unit vector *if $\langle \boldsymbol{u}, \boldsymbol{u}\rangle = 1$;*
3. *A collection of unit vectors $\mathcal{V}$ is* orthonormal *if $\langle \boldsymbol{v}, \boldsymbol{w}\rangle = 0$ for all $\boldsymbol{v}, \boldsymbol{w} \in \mathcal{V}$.*

Example 1.11. Let the vector space under consideration be $C^{(0)}[a,b]$, the space of continuous functions on the interval $[a,b]$. Define

$$\begin{aligned} \langle f, g\rangle &= \int_a^b f(t)g(t)dt \\ &= \text{net area under curve } fg. \end{aligned}$$

To check that this is an inner product requires verifying that C is a vector space and exhibits the other properties of inner product. However,

1. For $f \not\equiv 0$, $\langle f, f \rangle > 0$;
2. Additivity, scalar multiplication properties, commutativity, and the distributive property all follow from properties of the integral;
3. The uniqueness of the zero element follows from properties of the integral. □

Lemma 1.12. *If $\boldsymbol{W}$ is a finite collection of orthonormal vectors in a space V, then the vectors in $\boldsymbol{W}$ are* linearly independent.

Proof: Suppose $\boldsymbol{W} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ and

$$0 = r_1\boldsymbol{e}_1 + r_2\boldsymbol{e}_2 + \cdots + r_n\boldsymbol{e}_n$$

then

$$0 = \langle 0, \boldsymbol{e}_j \rangle = r_1\langle \boldsymbol{e}_1, \boldsymbol{e}_j \rangle + r_2\langle \boldsymbol{e}_2, \boldsymbol{e}_j \rangle + \cdots + r_i\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle + \cdots + r_n\langle \boldsymbol{e}_n, \boldsymbol{e}_j \rangle.$$

Since $\langle \boldsymbol{e}_i, \boldsymbol{e}_j \rangle = \delta_{i,j}$, $0 = r_j$. Letting $j = 1, \ldots, n$, gives the result that if $\boldsymbol{0}$ is a linear combination of the elements of $\boldsymbol{W}$, then all coefficients must be zero. ■

Lemma 1.13. *If $\boldsymbol{W} = \{\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n\}$ is a collection of orthonormal vectors with*

$$\boldsymbol{v} = r_{v,1}\boldsymbol{e}_1 + \cdots + r_{v,n}\boldsymbol{e}_n$$

and

$$\boldsymbol{w} = r_{w,1}\boldsymbol{e}_1 + \cdots + r_{w,n}\boldsymbol{e}_n,$$

then

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{i=1}^{n} r_{v,i} r_{w,i}.$$

Proof: The proof uses straightforward properties of inner product and is left as Exercise 7 for the reader. ■

Corollary 1.14. *If $\boldsymbol{W}$ and $\boldsymbol{w}$ are defined as above in Lemma 1.13 then*

$$\|\boldsymbol{w}\| = \sqrt{\sum_{i=1}^{n} r_{w,i}^2}.$$

1.1.1 $\mathbb{R}^2$ and $\mathbb{R}^3$ as Vector Spaces

In the special cases where the vector space is $\mathbb{R}^2$ or $\mathbb{R}^3$, we note an especially simple representation. In physics and mathematics one says that vectors are uniquely defined by direction and magnitude. Two vectors with the same direction and magnitude are the same—no matter where they are located. If a vector is located with its tail at the origin and its head (arrow) at the position (x, y, z), then the head position uniquely defines the direction and the magnitude of that vector. Thus, by convention, every point in the plane or in 3-space has a one-to-one and onto correspondence with the vector space $\mathbb{R}^2$ or $\mathbb{R}^3$, respectively. This correspondence matches the point (x, y, z) with the vector whose tail is at the origin and head at the point (x, y, z). Sometimes the two notations are used interchangeably which can lead to great confusion on the part of the novice. We shall use the term *free* vectors to mean those whose position is not bound, and *fixed* vectors to mean those with a bound position.

Let $\boldsymbol{e}_1 = (1,0,0)$, $\boldsymbol{e}_2 = (0,1,0)$, and $\boldsymbol{e}_3 = (0,0,1)$. Then the set $\boldsymbol{E} = \{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\} \in \mathbb{R}^3$ is orthonormal. For all real numbers x, y, and z, $x\boldsymbol{e}_1 + y\boldsymbol{e}_2 + z\boldsymbol{e}_3$ represents a unique vector in the span of $\boldsymbol{E}$ which is contained in $\mathbb{R}^3$, since the set $\boldsymbol{E}$ is linearly independent. Further, to every point in $\mathbb{R}^3 = \mathbb{R}^1 \times \mathbb{R}^1 \times \mathbb{R}^1$ there corresponds a vector in span $\boldsymbol{E}$. Hence, $\boldsymbol{E}$ is a basis for $\mathbb{R}^3$. Thus, every set of three orthonormal vectors forms a basis for $\mathbb{R}^3$.

Corollary 1.15. *In general, $\boldsymbol{E}$ is a set of three orthonormal vectors in $\mathbb{R}^3$, so for every pair of vectors $\boldsymbol{v}$, $\boldsymbol{w} \in \mathbb{R}^3$ with*

$$\boldsymbol{v} = r_{v,1}\boldsymbol{e}_1 + r_{v,2}\boldsymbol{e}_2 + r_{v,3}\boldsymbol{e}_3 \qquad \text{and} \qquad \boldsymbol{w} = r_{w,1}\boldsymbol{e}_1 + r_{w,2}\boldsymbol{e}_2 + r_{w,3}\boldsymbol{e}_3,$$

we have

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{i=1}^{3} r_{v,i} r_{w,i} \qquad \text{and} \qquad \|\boldsymbol{v}\| = \sqrt{\sum_{i=1}^{3} r_{v,i}^2}.$$

Whenever the particular choice of $\boldsymbol{E}$ is understood, one can write (r_1, r_2, r_3) to mean $r_1\boldsymbol{e}_1 + r_2\boldsymbol{e}_2 + r_3\boldsymbol{e}_3$. Frequently the $\boldsymbol{e}_1$ direction is denoted as the "x" direction, the $\boldsymbol{e}_2$ direction is denoted the "y" direction, and $\boldsymbol{e}_3$ direction is called the "z" direction. The triple (x, y, z) is then used to mean the vector $x\boldsymbol{e}_1 + y\boldsymbol{e}_2 + z\boldsymbol{e}_3$.

Define scalar multiplication as $c\boldsymbol{v} = (cx, cy, cz)$. In Exercise 3 the reader must show that this definition makes the vector c times longer, but does not change the direction.

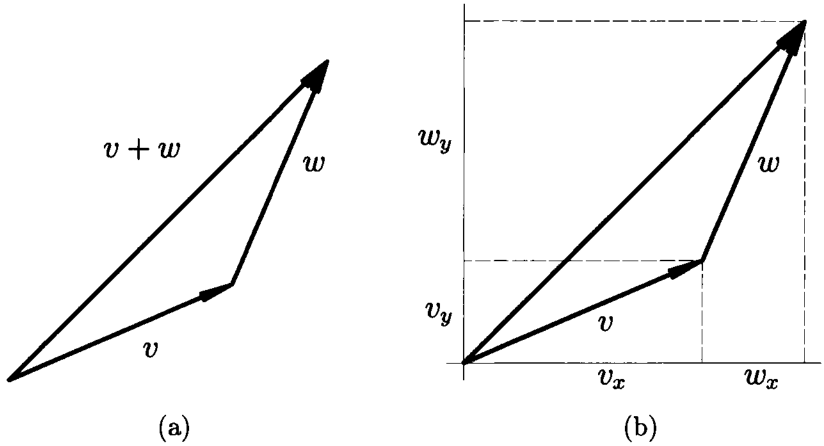


Figure 1.1. The geometric meaning of adding $\boldsymbol{v}$ and $\boldsymbol{w}$ in (a), and vector addition in (b).

Geometrically, vector addition can be interpreted as positioning $\boldsymbol{v}$ arbitrarily in space and then placing $\boldsymbol{w}$ so its tail is at the same position as the head of $\boldsymbol{v}$. Then the vector with its tail in the same position as the tail of $\boldsymbol{v}$ and its head in the same position as the head of $\boldsymbol{w}$ is the sum of the vectors $\boldsymbol{v}$ and $\boldsymbol{w}$, see Figure 1.1 (a).

To derive a quantitative formula, consider fixed formulations of the vectors $\boldsymbol{v} = (x_v, y_v, z_v)$ and $\boldsymbol{w} = (x_w, y_w, z_w)$. We wish to determine a method of finding the coordinate representation for $\boldsymbol{s} = \boldsymbol{v} + \boldsymbol{w}$, that is (x_s, y_s, z_s).

We know that $\boldsymbol{s} - \boldsymbol{v} = \boldsymbol{w}$. Given the positions, we know that the change of position in the x-direction must be x_w. Similarly for y and z. But the tail of $\boldsymbol{w}$ in this position is at (x_v, y_v, z_v). Thus its head must be x_w units over, or have an x coordinate of $x_v + x_w$. Similarly, the y and z coordinates of the head must be at $y_v + y_w$ and $z_v + z_w$, respectively. But the head of $\boldsymbol{w}$ in this position is at the same place as the head of $\boldsymbol{s}$ when $\boldsymbol{s}$ starts at the origin. Thus, in general $\boldsymbol{v} + \boldsymbol{w} = (x_v + x_w, y_v + y_w, z_v + z_w)$, as shown in Figure 1.1 (b).

A similar result may be derived for an oblique basis in $\boldsymbol{R}^2$ or $\boldsymbol{R}^3$. We look at the $\boldsymbol{R}^2$ case. Let $\boldsymbol{v}$ and $\boldsymbol{w}$ be two vectors in the plane that do not have the same direction, that is, $\boldsymbol{v} \neq c\boldsymbol{w}$ for all $c \in \boldsymbol{R}^1$. Then $\boldsymbol{v}$ and $\boldsymbol{w}$ are linearly independent. Consider $\boldsymbol{S} = span\{\boldsymbol{v}, \boldsymbol{w}\} = \left\{ a\boldsymbol{v} + b\boldsymbol{w} : a, b \in \boldsymbol{R}^1 \right\}$. First represent them as fixed vectors, $\boldsymbol{v} = (x_v, y_v)$ and $\boldsymbol{w} = (x_w, y_w)$. Suppose $\boldsymbol{u}$ is any vector, then it has a fixed representation $\boldsymbol{u} = (x_u, y_u)$. What vectors $\boldsymbol{u}$ in the plane are also in $\boldsymbol{S}$?

If we suppose $\{\boldsymbol{e_1}, \boldsymbol{e_2}\}$ is an orthonormal basis of $\boldsymbol{R}^2$ with $\boldsymbol{v} = x_v\boldsymbol{e_1} + y_v\boldsymbol{e_2}$ and $\boldsymbol{w} = x_w\boldsymbol{e_1} + y_w\boldsymbol{e_2}$ then $[\boldsymbol{v}\ \boldsymbol{w}] = [\boldsymbol{e_1}\ \boldsymbol{e_2}]A$ where

$$A = \begin{bmatrix} x_v & x_w \\ y_v & y_w \end{bmatrix}.$$

Suppose an arbitrary element $\boldsymbol{u} \in \boldsymbol{R}^2$ can be written $\boldsymbol{u} = x_u\boldsymbol{e_1} + y_u\boldsymbol{e_2}$. We want to know when there exists a_u, $b_u \in \boldsymbol{R}^1$ such that $\boldsymbol{u} = a_u\boldsymbol{v} + b_u\boldsymbol{w}$. That is, when there is a solution to

$$\begin{bmatrix} \boldsymbol{e_1} & \boldsymbol{e_2} \end{bmatrix} \begin{bmatrix} x_u \\ y_u \end{bmatrix} = \boldsymbol{u} = \begin{bmatrix} \boldsymbol{v} & \boldsymbol{w} \end{bmatrix} \begin{bmatrix} a_u \\ b_u \end{bmatrix}.$$

But since $[\boldsymbol{v}\ \boldsymbol{w}] = [\boldsymbol{e_1}\ \boldsymbol{e_2}]A$, this question is equivalent to asking when

$$A \begin{bmatrix} a_u \\ b_u \end{bmatrix} = \begin{bmatrix} x_u \\ y_u \end{bmatrix}$$

can be solved for unknowns a_u, b_u.

By Theorem 1.28, if $\det A \neq 0$, A^{-1} exists and the system can be solved.

Setting $[a_u\ b_u]^T = A^{-1}[x_u\ y_u]^T$ solves the system. Thus, for any $\boldsymbol{u}$, there exists $a_u, b_u \in \boldsymbol{R}^1$ such that $\boldsymbol{u} = a_u\boldsymbol{v} + b_u\boldsymbol{w}$. The pair (a_u, b_u) are the *coordinate values* in the oblique v-w coordinate system for the vector $\boldsymbol{u}$.

Consider vectors $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. Suppose they are placed so that their tails meet at the point O. Let θ denote the angle between the vectors.

We use the law of cosines to find the particular realization for the inner product in $\boldsymbol{R}^3$.

$$\|\boldsymbol{v_2} - \boldsymbol{v_1}\|^2 = \|\boldsymbol{v_2}\|^2 + \|\boldsymbol{v_1}\|^2 - 2\|\boldsymbol{v_2}\|\,\|\boldsymbol{v_1}\| \cos\theta.$$

Expanding the left side one gets:

$$\begin{aligned} \|\boldsymbol{v_2} - \boldsymbol{v_1}\|^2 &= \langle \boldsymbol{v_2} - \boldsymbol{v_1}, \boldsymbol{v_2} - \boldsymbol{v_1} \rangle \\ &= \langle \boldsymbol{v_2}, \boldsymbol{v_2} \rangle - \langle \boldsymbol{v_2}, \boldsymbol{v_1} \rangle - \langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle + \langle \boldsymbol{v_1}, \boldsymbol{v_1} \rangle \\ &= \|\boldsymbol{v_2}\|^2 - 2\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle + \|\boldsymbol{v_1}\|^2. \end{aligned}$$

Thus one has

$$\|\boldsymbol{v_2}\|^2 - 2\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle + \|\boldsymbol{v_1}\|^2 = \|\boldsymbol{v_2}\|^2 + \|\boldsymbol{v_1}\|^2 - 2\|\boldsymbol{v_2}\|\,\|\boldsymbol{v_1}\| \cos\theta$$

and

$$\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle = \|\boldsymbol{v_1}\|\,\|\boldsymbol{v_2}\| \cos\theta.$$

By Corollary 1.15, $\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle = x_1x_2 + y_1y_2 + z_1z_2$. Is there a geometric interpretation to this view of the inner product? Using the result gives the following theorem:

Theorem 1.16. *The* directed length *of the projection of* $\boldsymbol{v_1}$ *onto the direction of* $\boldsymbol{v_2}$ *is* $\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle / \|\boldsymbol{v_2}\|$.

Proof: Place the vectors so that their tails meet at point O. Drop a perpendicular from the head of $\boldsymbol{v_1}$ onto the direction of $\boldsymbol{v_2}$. From trigonometry it is known that, if θ is the angle between the vectors, then the directed length of $\boldsymbol{v_1}$ in the direction of $\boldsymbol{v_2}$ is given by the projection of

$$\begin{aligned} \boldsymbol{v_1} &= \|\boldsymbol{v_1}\| \cos\theta \\ &= \frac{\langle \boldsymbol{v_1}, \boldsymbol{v_2} \rangle}{\|\boldsymbol{v_2}\|}. \end{aligned}$$

Note that if the angle between the vectors is greater than 90 degrees, then the directed length of the projection is considered negative. ■

We can use the simple geometric knowledge that three points determine a plane, or two vectors determine a plane to give us more complicated information.

Several operations have been defined on elements in arbitrary vector spaces. Reviewing them, we see,

- Addition: $\boldsymbol{V} \times \boldsymbol{V} \to \boldsymbol{V}$;
- Scalar multiplication: $\boldsymbol{R}^1 \times \boldsymbol{V} \to \boldsymbol{V}$;
- Inner product: $\boldsymbol{V} \times \boldsymbol{V} \to \boldsymbol{R}^1$.

Another operation, the *cross product* can be defined for the special case when the vector space is $\boldsymbol{R}^3$. The cross product of $\boldsymbol{v}$ and $\boldsymbol{w}$, $\boldsymbol{v} \times \boldsymbol{w}$, can intuitively be defined as a vector that is perpendicular to both $\boldsymbol{v}$ and $\boldsymbol{w}$, with orientation prescribed by the *right hand rule* and a magnitude prescribed by a rule which depends on the magnitude of $\boldsymbol{v}$, the magnitude of $\boldsymbol{w}$ and the angle between them. That is, point the right hand in the direction of $\boldsymbol{v}$. Then move it to the direction of $\boldsymbol{w}$ in a continuous rotational movement. The direction that the right thumb points is the direction of the cross product vector. That leaves one degree of freedom. That freedom can be restricted by requiring that $\boldsymbol{e_1} \times \boldsymbol{e_2} = \boldsymbol{e_3}$.

The formula for the coefficients can be derived from the knowledge that if $\boldsymbol{u} = \boldsymbol{v} \times \boldsymbol{w}$ then $\langle \boldsymbol{u}, \boldsymbol{v} \rangle = 0$ and $\langle \boldsymbol{u}, \boldsymbol{w} \rangle = 0$, since $\boldsymbol{u}$ is perpendicular to both $\boldsymbol{v}$ and $\boldsymbol{w}$.

If we let $\boldsymbol{e_1}$ be the unit vector in the x direction, $\boldsymbol{e_2}$ be the unit vector in the y direction, and $\boldsymbol{e_3}$ be the unit vector in the z direction, then if $\boldsymbol{u} = (x_u, y_u, z_u)$, $\boldsymbol{v} = (x_v, y_v, z_v)$, and $\boldsymbol{w} = (x_w, y_w, z_w)$, then $\boldsymbol{i} = x_i\boldsymbol{e_1} + y_i\boldsymbol{e_2} + z_i\boldsymbol{e_3}$, $i \in \{u, v, w\}$.

The inner product equations now can be written

$$\begin{aligned}
\langle \boldsymbol{v}, \boldsymbol{u} \rangle &= x_v x_u + y_v y_u + z_v z_u &= 0, \\
\langle \boldsymbol{w}, \boldsymbol{u} \rangle &= x_w x_u + y_w y_u + z_w z_u &= 0.
\end{aligned}$$

Since the system consists of two linear equations in three unknowns, it still has an undefined degree of freedom. Rewrite the equations in terms of x_u as known and both y_u and z_u as unknown:

$$\begin{aligned}
-x_v x_u &= y_v y_u + z_v z_u, \\
-x_w x_u &= y_w y_u + z_w z_u.
\end{aligned}$$

Then solve for y_u and z_u to get

$$\begin{aligned}
y_u &= x_u \frac{z_v x_w - x_v z_w}{y_v z_w - z_v y_w}, \\
z_u &= x_u \frac{y_w x_v - y_v x_w}{y_v z_w - z_v y_w}.
\end{aligned}$$

so

$$\begin{aligned}
\boldsymbol{u} &= \frac{x_u}{y_v z_w - z_v y_w} \left(y_v z_w - z_v y_w, z_v x_w - x_v z_w, y_w x_v - y_v x_w \right) \\
&= c \left(y_v z_w - z_v y_w, z_v x_w - x_v z_w, y_w x_v - y_v x_w \right).
\end{aligned}$$

Now, the last degree of freedom is set by the magnitude and "right hand rule" orientation requirement. Since c must be the same constant for all $\boldsymbol{v}$, $\boldsymbol{w} \in \boldsymbol{R}^3$, one can choose simple cases to determine c. Let $\boldsymbol{v} = \boldsymbol{e_1}$ and $\boldsymbol{w} = \boldsymbol{e_2}$. Since $\boldsymbol{u} = \boldsymbol{e_3} = c(0, 0, 1)$, c must equal 1.

Definition 1.17. *For $\boldsymbol{v}$, $\boldsymbol{w} \in \boldsymbol{R}^3$, the* cross product *operator defines a new vector, $\boldsymbol{v} \times \boldsymbol{w}$, as $\boldsymbol{v} \times \boldsymbol{w} = \left(y_v z_w - z_v y_w, z_v x_w - x_v z_w, y_w x_v - y_v x_w \right)$. The cross product is not a commutative operation.*

Lemma 1.18. $\|\boldsymbol{v} \times \boldsymbol{w}\|^2 = \|\boldsymbol{v}\|^2 \|\boldsymbol{w}\|^2 - \left\|(\boldsymbol{v}, \boldsymbol{w})\right\|^2$.

Proof: It is left as an exercise. ■

Theorem 1.19. $\|\boldsymbol{v} \times \boldsymbol{w}\| = \|\boldsymbol{v}\| \, \|\boldsymbol{w}\| \, |\sin\theta|$.

Proof: Since $\left\|(\boldsymbol{v}, \boldsymbol{w})\right\|^2 = \|\boldsymbol{v}\|^2 \, \|\boldsymbol{w}\|^2 \cos^2\theta$, Lemma 1.18 gives,

$$\begin{aligned}\|\boldsymbol{v}\times\boldsymbol{w}\|^2 &= \|\boldsymbol{v}\|^2\|\boldsymbol{w}\|^2 - \|\boldsymbol{v}\|^2\|\boldsymbol{w}\|^2\cos^2\theta \\ &= \|\boldsymbol{v}\|^2\|\boldsymbol{w}\|^2(1-\cos^2\theta) \\ &= \|\boldsymbol{v}\|^2\|\boldsymbol{w}\|^2(\sin^2\theta).\end{aligned}$$

Since $0 \le \sin\theta$ for $0 \le \theta \le \pi$, $\|\boldsymbol{v}\times\boldsymbol{w}\| = \|\boldsymbol{v}\|\,\|\boldsymbol{w}\|\sin\theta$. ■

Thus, the cross product of two vectors $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$ is the vector with magnitude equal to the area of the parallelogram described when the tails of the vectors meet at a point O, with direction perpendicular to the plane defined by $\boldsymbol{v_1}$ and $\boldsymbol{v_2}$. The orientation of this perpendicular is determined by the right hand rule. The following properties are a direct consequence of the definition:

$$\begin{aligned}\boldsymbol{u}\times\boldsymbol{v} &= -\boldsymbol{v}\times\boldsymbol{u} \\ \boldsymbol{u}\times\boldsymbol{u} &= 0 \\ \boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w}) &\neq (\boldsymbol{u}\times\boldsymbol{v})\times\boldsymbol{w}.\end{aligned}$$

The example $(\boldsymbol{e_1}\times\boldsymbol{e_1})\times\boldsymbol{e_2} =$ undefined, but $\boldsymbol{e_1}\times(\boldsymbol{e_1}\times\boldsymbol{e_2}) = \boldsymbol{e_1}\times\boldsymbol{e_3} = -\boldsymbol{e_2}$.

To compute the cross product for $\boldsymbol{v_i} = (x_i, y_i, z_i)$, $i = 1, 2$, determinants are commonly used:

$$\boldsymbol{v} = \boldsymbol{v_1}\times\boldsymbol{v_2} = \begin{vmatrix} \boldsymbol{e_1} & \boldsymbol{e_2} & \boldsymbol{e_3} \\ x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \end{vmatrix}.$$

We define the *triple scalar product* of three vectors $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{w}$ as

$$\langle\boldsymbol{u},(\boldsymbol{v}\times\boldsymbol{w})\rangle = \boldsymbol{u}\cdot(\boldsymbol{v}\times\boldsymbol{w}) = \begin{vmatrix} u_x & u_y & u_z \\ v_x & v_y & v_z \\ w_x & w_y & w_z \end{vmatrix}.$$

Note that $\langle\boldsymbol{u},\boldsymbol{v}\times\boldsymbol{w}\rangle = \langle\boldsymbol{v}\times\boldsymbol{w},\boldsymbol{u}\rangle = -\langle\boldsymbol{v},\boldsymbol{u}\times\boldsymbol{w}\rangle$.

Finally, the *triple vector product* can be decomposed in terms of inner products as

$$\boldsymbol{u}\times(\boldsymbol{v}\times\boldsymbol{w}) = (\boldsymbol{u},\boldsymbol{w})\boldsymbol{v} - (\boldsymbol{u},\boldsymbol{v})\boldsymbol{w}.$$

1.2 Linear Transformations

Definition 1.20. *Suppose $\boldsymbol{X}$ and $\boldsymbol{Y}$ are vector spaces and T is a function, $T : \boldsymbol{X}\to\boldsymbol{Y}$ such that*

$$T(r_1\boldsymbol{x_1} + r_2\boldsymbol{x_2}) = r_1T(\boldsymbol{x_1}) + r_2T(\boldsymbol{x_2}),$$

then T is called a linear transformation *from $\boldsymbol{X}$ to $\boldsymbol{Y}$, or a* linear operator.

Example 1.21. The following are examples of linear transformations:

1. $P : \mathbb{R}^3 \to \mathbb{R}^3$ by $P\big((x, y, z)\big) = (x, y, 0)$. Then P is called the *orthogonal projection* from $\mathbb{R}^3$ to the x-y plane. Orthogonal projections to the x-z plane and y-z plane are defined analogously.

   This is a linear transformation since

$$\begin{aligned} P\Big(r_1(x_1, y_1, z_1) + r_2(x_2, y_2, z_2)\Big) &= (r_1x_1 + r_2x_2, r_1y_1 + r_2y_2, 0) \\ &= (r_1x_1, r_1y_1, 0) + (r_2x_2, r_2y_2, 0) \\ &= r_1(x_1, y_1, 0) + r_2(x_2, y_2, 0) \\ &= r_1P\Big((x_1, y_1, z_1)\Big) + r_2P\Big((x_2, y_2, z_2)\Big). \end{aligned}$$

   The proof of the properties about P relies on the vector space properties of $\mathbb{R}^2$ and $\mathbb{R}^3$. Such proofs are typical of showing an operator is a linear transformation.

2. $D[f] = f'$, the derivative operator.

3. $I[f] = \int_a^b f(t)dt$.

4. $T_{j,x^*} : C^{(n)}[a, b] \to \mathbb{R}^1$ by $T_{j,x^*}[f] = f^{(j)}(x^*)$, for $j = 0, \ldots, n$: point evaluation of the j^{th} derivative at a particular point x^*.

5. $T : C^{(n+1)}[a, b] \to P_n$, the polynomials of degree less than or equal to n, by $T[f] = f(x_0) + f'(x_0)(x - x_0) + \frac{f^{(2)}(x_0)}{2!}(x - x_0)^2 + \cdots + \frac{f^{(n)}(x_0)}{n!}(x - x_0)^n$.

   Note that T takes a function to a function and we can write $\big(T[f]\big)(x)$ or $T[f](x)$ to evaluate that function at a point x. □

Definition 1.22. *A linear transformation whose range is $\mathbb{R}^1$ is called a* linear functional.

Definition 1.23. *Suppose S and T are two linear transformations from $\boldsymbol{V}$ to $\boldsymbol{W}$ and $r \in \mathbb{R}^1$. Define addition by $(S + T)(v) = S(v) + T(v)$ and define scalar multiplication as $(rS)(v) = r\big(S(v)\big)$. Problem 11 shows that $(S + T)$ and rS are both linear transformations, and that the set of linear transformations from $\boldsymbol{V}$ to $\boldsymbol{W}$, $L(\boldsymbol{V}, \boldsymbol{W})$, is a vector space with these operations defined.*

1.3 Review of Matrix Properties

Definition 1.24. *Consider a rectangular array of numbers arranged as follows:*

$$A = (a_{ij}) = \begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ & & \vdots & \\ a_{m,1} & a_{m,2} & \cdots & a_{m,n} \end{bmatrix}$$

where $\boldsymbol{a_j} = [\ a_{1,j} \quad a_{2,j} \quad \ldots \quad a_{m,j}\ ]^T$*, a vector in* $\mathbb{R}^m$*. The array is called an* $m \times n$ matrix*, and* $\boldsymbol{M_{m,n}}$ *denotes the set of all matrices with* n *vectors from* $\mathbb{R}^m$*. If* $m = n$ *we say that the matrices in* $\boldsymbol{M_{n,n}} = \boldsymbol{M_n}$ *are square.*

Definition 1.25. *The* transpose *of* A *is* $(a_{j,i})$ *and is denoted by* A^T*.*

Denote by $\boldsymbol{e_j}$ the column vector $(0, \ldots, 0, \delta_{i,j}, 0, \ldots, 0)^T$ consisting of all zeros except a 1 in the j^{th} position.

If A and B are two matrices in $\boldsymbol{M_{m,n}}$ we can define the following operations:

1. Addition: $+ : \boldsymbol{M_{m,n}} \times \boldsymbol{M_{m,n}} \to \boldsymbol{M_{m,n}}$. Denote by $A + B$ the matrix equal to $(a_{i,j} + b_{i,j})$, the matrix whose elements are the sum elementwise of elements of A and B.

2. Scalar Multiplication: $\mathbb{R}^1 \times \boldsymbol{M_{m,n}} \to \boldsymbol{M_{m,n}}$. If $r \in \mathbb{R}^1$ then $(rA) = (ra_{i,j})$, the matrix whose elements are multiplied elementwise by the scalar r.

3. Zero element: Let $Z = (0)$ be the matrix such that $z_{i,j} = 0$, for all i, j. Then $A + Z = A$ for all matrices A, and Z is the identity under addition. It can be shown easily that the rest of the properties hold for $\boldsymbol{M_{m,n}}$ to be called a vector space.

4. Define another operation: $\cdot : \boldsymbol{M_{m,n}} \times \boldsymbol{M_{n,k}} \to \boldsymbol{M_{m,k}}$. For $A \in \boldsymbol{M_{m,n}}$ and $B \in \boldsymbol{M_{n,k}}$ $C \in \boldsymbol{M_{m,k}}$ is defined by $C = AB = (c_{i,j})$ where $c_{i,j} = \sum_{p=1}^{n} a_{i,p} b_{p,j}$. The matrix C is called the product of A and B. Note that the dimensions must match to be able to multiply A and B.

Suppose $m = n = p$ and let $I = (\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$, then $AI = A = IA$, and I is called a multiplicative identity. Suppose I_2 is another multiplicative identity. $I = II_2 = I_2$; the left side occurring since I_2 is an identity, and the right side occurring since I is an identity. This proves that the multiplicative identity is unique.

Definition 1.26. *If $A \in \boldsymbol{M_n}$, and there exists a matrix B such that $AB = BA = I$, then B is called the* inverse *to A (conversely, A is the inverse to B).*

If A and B are in $\boldsymbol{M_n}$, does $AB \stackrel{?}{=} BA$? Namely is matrix multiplication a commutative operation? The answer is negative, in general. Let

$$A = \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix} \quad \text{and} \quad B = \begin{bmatrix} 1 & 3 \\ 1 & 1 \end{bmatrix}$$

then

$$AB = \begin{bmatrix} 4 & 3 \\ 2 & 1 \end{bmatrix} \quad \text{and} \quad BA = \begin{bmatrix} 1 & 3 \\ 2 & 4 \end{bmatrix}.$$

Definition 1.27. *Let $A = [\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n] \in M_n$. The* determinant *is a functional $|\ | : \boldsymbol{M_n} \to \mathbb{R}^1$ written $|A| = D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n)$ which is completely defined by the following conditions:*

1. $D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a_n}) = D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i + \boldsymbol{a}_j, \ldots, \boldsymbol{a}_n)$, *for* $i \neq j$;
2. $D(\boldsymbol{a}_1, \ldots, r\boldsymbol{a}_i, \ldots, \boldsymbol{a}_n) = rD(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a}_n)$;
3. $D(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_i, \ldots, \boldsymbol{e_n}) = 1$.

Define the i-j^{th} *cofactor* as $A^*_{i,j} = (-1)^{i+j}|A_{i,j}|$ where $A_{i,j}$ is the $(n-1) \times (n-1)$ matrix that omits the i^{th} row and the j^{th} column of A.

Properties of the determinant:

1. $|A| = |A^T|$;
2. Define B as A with any two rows interchanged. Then, $|A| = -|B|$;
3. If $\boldsymbol{a_i} = \boldsymbol{a_j}$ for $i \neq j$, then $|A| = 0$;
4. $D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a_n}) = D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i + r\boldsymbol{a}_j, \ldots, \boldsymbol{a}_n)$, for $i \neq j$;
5. $D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i, \ldots, \boldsymbol{a_n}) + D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}'_{\boldsymbol{i}}, \ldots, \boldsymbol{a_n}) =$
   $D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_i + \boldsymbol{a}'_{\boldsymbol{i}}, \ldots, \boldsymbol{a_n})$;
6. Fix j, then $|A| = \sum_{i=1}^n a_{i,j}A^*_{i,j}$ or fix i, then $|A| = \sum_{j=1}^n a_{i,j}A^*_{i,j}$.

Since property (1) holds, the properties (2) through (5) hold also when the term *row* is substituted for *column* and *column* for *row*.

There is a geometric interpretation of the determinant as well. Consider either the rows or the columns of a matrix as vectors in $\mathbb{R}^n$. Then the determinant is the area of the n-dimensional hyper-parallelopiped generated by those vector *edges.*

For linear equations $\sum_{j=1}^{n} a_{i,j}x_j = b_i$ for $i = 1, \ldots, n$, where the $a_{i,j}$, and b_i are known, and the x_j, $j = 1, \ldots, n$ are unknown, it is necessary, sometimes, to determine if the system has no solutions, a unique solution, or many solutions. And, if there is a unique solution, how it can be found. The above system of equations can readily be posed as a matrix problem:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ & & \vdots & \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix} = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

$$A \qquad\qquad X \quad = \quad B$$

where $A = (a_{i,j})$, is the $n \times n$ matrix of coefficients, X is the $n \times 1$ matrix of unknowns, and B is the $n \times 1$ matrix of equation values.

The solution to this problem is given by $A^{-1}AX = IX = X = A^{-1}B$, when such an inverse exists. Does the inverse exist? These questions are answered by

Theorem 1.28. Cramer's Rule. *If $|A| \neq 0$ then the solution to the above system has a unique solution given by*

$$x_r = \frac{\sum_{i=1}^{n} A^*{}_{i,r} b_i}{|A|}, \qquad r = 1, \ldots, n$$

or, the homogeneous system with $b_j = 0$, $j = 1, \ldots, n$ possesses a nontrivial solution if and only if $|A| = 0$.

Proof: We shall treat existence only, here. If the $\boldsymbol{a_j}$ are considered as n column vectors over $\mathbb{R}^n$, then the system of equations can be written

$$x_1\boldsymbol{a_1} + x_2\boldsymbol{a_2} + \cdots + x_n\boldsymbol{a_n} = \boldsymbol{b}.$$

The homogeneous case is defined as the case when $\boldsymbol{b} = 0$. A solution to the nonhomogeneous system is equivalent to $\boldsymbol{b}$ being in the span of $\boldsymbol{a_j}$, $j = 1, \ldots, n$. A nontrivial solution to the homogeneous case is equivalent to the vectors $\boldsymbol{a_j}$, $j = 1, \ldots, n$, being dependent.

If $\boldsymbol{a_j} = \mathbf{0}$ for any j then clearly every vector in the span can be written in an infinite number of ways, so the vectors are dependent. Hence, $\boldsymbol{a_j} \neq \mathbf{0}$, $j = 1, \ldots, n$. Now, consider the homogeneous problem and suppose it has a nontrivial solution $\boldsymbol{x} = (x_1, \ldots, x_n)$. Suppose, without loss of generality, that $x_1 \neq 0$. Then, it is true that

$$\boldsymbol{a_1} + c_2' \boldsymbol{a_2} + \cdots + c_n' \boldsymbol{a_n} = 0$$

for new scalars c_i', $i = 2, \ldots, n$, not all zero. But

$$\begin{aligned} D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) &= D(\boldsymbol{a_1} + c_2'\boldsymbol{a_2} + \cdots + c_n'\boldsymbol{a_n}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) \\ &= D(\mathbf{0}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) \\ &= 0D(\mathbf{0}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) \\ &= 0. \end{aligned}$$

Thus, if the columns are dependent vectors, then the determinant is zero. So if the homogeneous system has a nontrivial solution, the determinant is zero.

Now, what vectors $\boldsymbol{b}$ can be written as unique linear combinations of the vectors $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$? We shall show the result using a proof by induction on the size of the system of equations. First, if $n = 1$, namely, $a_{1,1}x_1 = b_1$, for $b_1 \neq 0$, has a solution if and only if $|A| = a_{1,1} \neq 0$. If $b_1 = 0$ then a nontrivial solution results if and only if $|A| = 0$. Now suppose it is true that if there are k equations in k unknowns, for all $k < n$, the theorem is true, and consider the $n \times n$ case.

If the set of n column vectors is independent over an n dimensional space and hence forms a basis,

$$\boldsymbol{e_1} = r_1\boldsymbol{a_1} + r_2\boldsymbol{a_2} + \cdots + r_n\boldsymbol{a_n},$$

for some new collection of coefficients r_i. Then, without loss of generality, suppose $r_1 \neq 0$:

$$\begin{aligned} r_1 D(\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) &= D(r_1\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n) \\ &= D(r_1\boldsymbol{a_1} + r_2\boldsymbol{a_2} + \cdots + r_n\boldsymbol{a_n}, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) \\ &= D(\boldsymbol{e}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n) \\ &= \sum_{j=1}^{n} \boldsymbol{e}_{1,j} {A^*}_{1,j} \\ &= {A^*}_{1,1}. \end{aligned}$$

$A^*_{1,1}$ is a $n-1 \times n-1$ determinant. $|A| = 0$ only if $A^*_{1,1} = 0$. Let $\boldsymbol{a}_j^*$ denote the j^{th} column vector of A with the first element omitted. Then

$A^*_{1,1} = D(\boldsymbol{a}^*_2, \ldots, \boldsymbol{a}^*_n)$. If $A^*_{1,1} = 0$ then by the induction hypothesis, there exist scalars $x_2, \ldots, x_n$ not all zero, making an $n-1$ vector $\boldsymbol{x}^*$, such that

$$x_2\boldsymbol{a}^*_2 + \cdots + x_n\boldsymbol{a}^*_n = 0.$$

Let $x_1 = -(x_2a_{1,2} + \cdots + x_na_{1,n})$. Then

$$x_1\boldsymbol{e_1} + x_2\boldsymbol{a_2} + \cdots + x_n\boldsymbol{a_n} = 0,$$

which means that $\boldsymbol{e}_1, \boldsymbol{a}_2, \ldots, \boldsymbol{a}_n$ is a dependent set. But since $-r_1\boldsymbol{a_1} = -\boldsymbol{e_1} + r_2\boldsymbol{a_2} + \cdots + r_n\boldsymbol{a_n}$, that implies that $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ is a dependent set. Thus, if $\boldsymbol{a}_1, \ldots, \boldsymbol{a}_n$ are independent, then $|A| \neq 0$. We already showed that if they are dependent then $|A| = 0$.

That contradicts the hypothesis. Hence $A^*_{1,1} \neq 0$ and $|A| \neq 0$. This shows that linear independence of the columns is equivalent to a nonzero determinant and implies a basis of $\boldsymbol{R}^n$ which means that the column vector on the right can be written as a linear combination of the columns on the left.

The constructive part of the proof appears in many advanced calculus books. ■

1.4 Barycentric Coordinates

One mechanism for escaping coordinate system dependence is to develop a method for specifying arbitrary points in the plane, or in $\boldsymbol{R}^3$ as combinations of points (vectors) which have some meaning to the problem at hand. Section 3.7 develops an application of this for representing conic sections. We here develop the simplest forms and properties for barycentric coordinates.

We know that $n+1$ points P_i, $i = 0, \ldots, n$, in $\boldsymbol{R}^n$ can be used to form the vectors $\boldsymbol{r_i} = P_i - P_0$, $i = 1, \ldots, n$. If the vectors $\{\boldsymbol{r_i}\}$ form a basis of $\boldsymbol{R}^n$, then the points $\{P_i\}$ are said to be in *general position.*

Suppose there are two points P_0 and P_1 and a point T on the line through P_0 and P_1. If $P_0 \neq P_1$, then these points determine a line which can be thought of as a transformation of $\boldsymbol{R}^1$. Clearly $\boldsymbol{r_1} = P_1 - P_0$ forms a basis for this one-dimensional subspace, and hence all points on the line can be written as a combination of P_0 and P_1.

Considering $\boldsymbol{R}^1$, suppose P_0, P_1, and T are just real numbers, with $P_0 < T$ and $P_0 < P_1$. Then considered as vectors,

$$\frac{T - P_0}{\|T - P_0\|} = \frac{P_1 - P_0}{\|P_1 - P_0\|},$$

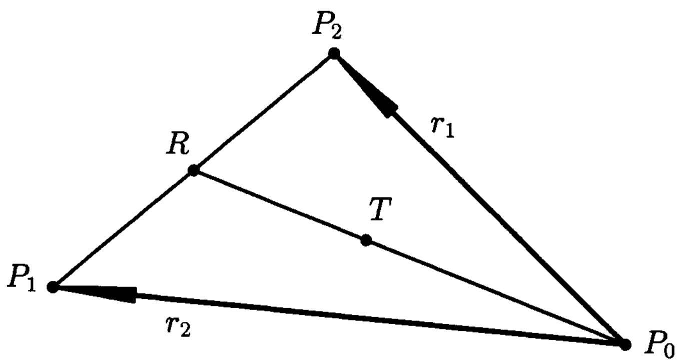


Figure 1.2. Finding barycentric coordinates.

which, setting $\lambda = \|T - P_0\| \ / \ \|P_1 - P_0\|$ simplifies to

$$T = (1 - \lambda)P_0 + \lambda P_1. \tag{1.1}$$

When $T < P_0$, Equation 1.1 holds true if $\lambda = -\|T - P_0\| \ / \ \|P_1 - P_0\|$. Here, the magnitudes of the vectors are just their absolute values. However, equation 1.1 holds true for finding the barycentric coordinates of a point in a one-dimensional subspace of $\mathbb{R}^n$with respect to two other points. Hence T is a convex combination of P_0 and P_1. Then $(1 - \lambda)$ and λ are called the *barycentric coordinates* of T with respect to P_0 and P_1. If $P_0 < P_1 < T$, then the ratio λ above is greater than 1. It is still true that $T = (1 - \lambda)P_0 + \lambda P_1$, however, the coefficient of P_0 is now a negative number. Analogously, if $T < P_0$, λ is negative. If P_0 and P_1 are two points in $\mathbb{R}^n$, the exact same results hold since there exists a translation followed by a rotation which will take the line through P_0 and P_1 into the x-axis.

Now, in $\mathbb{R}^2$, suppose the three points P_0, P_1, and P_2 are in general position. Further, suppose that the point T is in the interior of the triangle formed by the three points. Draw a line from one of the points, say P_0, through the point T until it intersects the edge P_1P_2. Call that point R. (See Figure 1.2.)

We suppose that $P_i = (x_i, y_i)$, $T = (x_t, y_t)$, and $R = (x_r, y_r)$. Since R is on the line connecting P_1 and P_2, there exists a real number α such that $R = (1 - \alpha)P_1 + \alpha P_2$. Now, since T is on the line segment connecting R and P_0, there exists a real number β, such that $T = (1 - \beta)R + \beta P_0$. Putting the two equations together,

$$\begin{aligned}
T &= (1 - \beta)\big[(1 - \alpha)P_1 + \alpha P_2\big] + \beta P_0 \\
&= \beta P_0 + (1 - \beta)(1 - \alpha)P_1 + (1 - \beta)\alpha P_2 \\
&= \lambda_0 P_0 + \lambda_1 P_1 + \lambda_2 P_2;
\end{aligned}$$

$$\begin{aligned}\lambda_0 + \lambda_1 + \lambda_2 &= \beta + (1-\beta)(1-\alpha) + (1-\beta)\alpha \\ &= \beta + (1-\beta)\big[(1-\alpha+\alpha)\big] \\ &= \beta + (1-\beta) \\ &= 1.\end{aligned}$$

Thus, the sum of the coefficients is 1. These coefficients, λ_i, $i = 0, 1, 2$, are called the *barycentric coordinates* with respect to P_0, P_1, and P_2 and depend linearly on T. If T is inside or on the boundaries of the triangle formed by P_0, P_1, and P_2, then $0 \le \alpha \le 1$ and $0 \le \beta \le 1$, so $0 \le \lambda_i \le 1$, for $i = 0, \ldots, 2$.

We now want to give a geometric interpretation of barycentric coordinates as ratios of areas.

Since R is also on the line through P_0 and T, $R = \delta T + (1-\delta)P_0$ as well as $R = (1-\alpha)P_1 + \alpha P_2$. We first set the two equations equal,

$$P_0 - P_1 = \alpha(P_2 - P_1) + \delta(P_0 - T),$$

and then break the x and y components of this new vector equation apart.

$$\begin{aligned}x_0 - x_1 &= \alpha(x_2 - x_1) + \delta(x_0 - x_t), \\ y_0 - y_1 &= \alpha(y_2 - y_1) + \delta(y_0 - y_t).\end{aligned}$$

Using Cramer's Rule to solve this system yields

$$\begin{aligned}\alpha &= \frac{\begin{vmatrix} x_0 - x_1 & x_0 - x_t \\ y_0 - y_1 & y_0 - y_t \end{vmatrix}}{\begin{vmatrix} x_2 - x_1 & x_0 - x_t \\ y_2 - y_1 & y_0 - y_t \end{vmatrix}} \\ &= \frac{\|(P_0 - P_1) \times (P_0 - T)\|}{\|(P_2 - P_1) \times (P_0 - T)\|}\end{aligned}$$

with the natural extended notion of the cross products over vectors in $\boldsymbol{R}^3$, considering vector $V = (v_x, v_y)$ as $V = (v_x, v_y, 0)$, and

$$\begin{aligned}\delta &= \frac{\begin{vmatrix} x_2 - x_1 & x_0 - x_1 \\ y_2 - y_1 & y_0 - y_1 \end{vmatrix}}{\begin{vmatrix} x_2 - x_1 & x_0 - x_t \\ y_2 - y_1 & y_0 - y_t \end{vmatrix}} \\ &= \frac{\|(P_2 - P_1) \times (P_0 - P_1)\|}{\|(P_2 - P_1) \times (P_0 - T)\|}.\end{aligned}$$

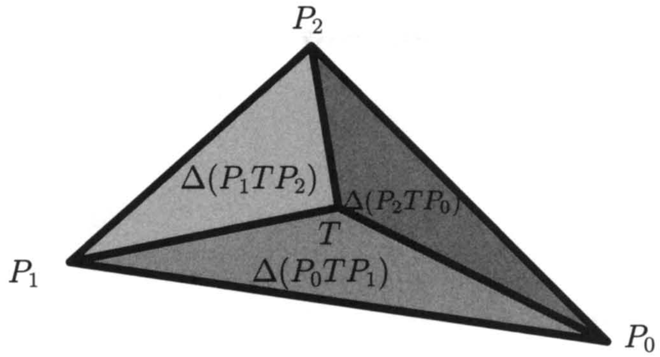


Figure 1.3. Barycentric coordinates as ratios of areas.

Since $R = P_0 + \delta(T - P_0)$, $\delta > 0$, and

$$T = \frac{1}{\delta}(R - P_0) + P_0,$$

then

$$T = \left(1 - \frac{1}{\delta}\right) P_0 + \frac{1-\alpha}{\delta} P_1 + \frac{\alpha}{\delta} P_2.$$

We finish by solving for these three coefficients in terms of the coordinates of the points. We will use the property, shown in the exercises relating cross products to areas of related parallelograms and triangles.

The coefficient of P_2 is

$$\begin{aligned} \frac{\alpha}{\delta} &= \frac{\|(P_0 - P_1) \times (P_0 - T)\|}{\|(P_2 - P_1) \times (P_0 - P_1)\|} \\ &= \frac{\text{area } \Delta(P_0TP_1)}{\text{area } \Delta(P_0P_1P_2)}. \end{aligned}$$

Applying the same equations, but reverting to the determinant values gives the coefficient of P_0 as

$$\begin{aligned} 1 - \frac{1}{\delta} &= 1 - \frac{\|(P_2 - P_1) \times (P_0 - T)\|}{\|(P_2 - P_1) \times (P_0 - P_1)\|} \\ &= \frac{\|(P_2 - P_1) \times (T - P_1)\|}{\|(P_2 - P_1) \times (P_0 - P_1)\|} \\ &= \frac{\text{area } \Delta(P_1TP_2)}{\text{area } \Delta(P_0P_1P_2)}. \end{aligned}$$

We have used the identity $P_2 - P_1 = (P_2 - T) + (T - P_1)$ to get to the final

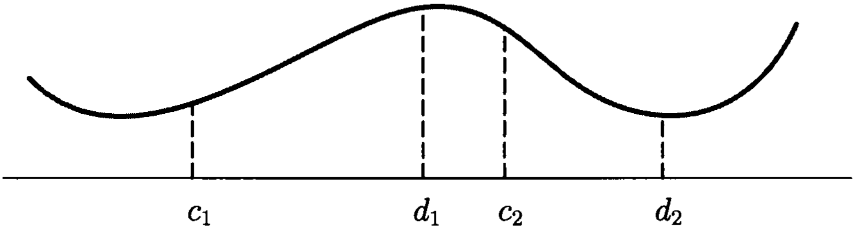


Figure 1.4. The function is increasing on (c_1, d_1) and is decreasing on (c_2, d_2).

identity. Analogously, the coefficient of P_1 is

$$\frac{1-\alpha}{\delta} = \frac{\text{area } \Delta(P_0 T P_2)}{\text{area } \Delta(P_0 P_1 P_2)}.$$

Thus, it is shown that the barycentric coordinates for a point within a triangle are the ratios of the area of the subtriangle opposite the vertex to the area of the whole triangle.

The analogous result is true for a point within a tetrahedron (in $\mathbb{R}^3$), that is, the four barycentric coordinates are the ratios of the volume of the *opposite* subtetrahedron to the volume of the whole tetrahedron.

1.5 Functions

Definition 1.29. *A function $f(x)$ is called* increasing (non-decreasing) *on an interval (c, d) if for all $u, v \in (c, d)$, $u < v$ implies $f(u) < f(v)$ $(f(u) \leq f(v))$.*

Definition 1.30. *A function $f(x)$ is called* decreasing (non-increasing) *on an interval (c, d) if for all $u, v \in (c, d)$, $u < v$ implies $f(u) > f(v)$ $(f(u) \geq f(v))$.*

Theorem 1.31. *Suppose $f(x) \in C^{(1)}(c, d)$. If $f'(x) > 0$ for $x \in (c, d)$, then $f(x)$ is increasing on (c, d). If $f'(x) < 0$ for $x \in (c, d)$, then $f(x)$ is decreasing on (c, d).*

Definition 1.32. *A* (local) maximum *for a function $f \in C^{(0)}$ occurs at a point x_0 if there exists $\epsilon > 0$ so that $f(x_0) \geq f(x)$ for all $x \neq x_0$ such that $|x - x_0| < \epsilon$. A* (local) minimum *to f is defined analogously.*

Definition 1.33. *The* extremal points *of a function are the ordered abscissa-ordinate pairs at which maxima or minima occur.*

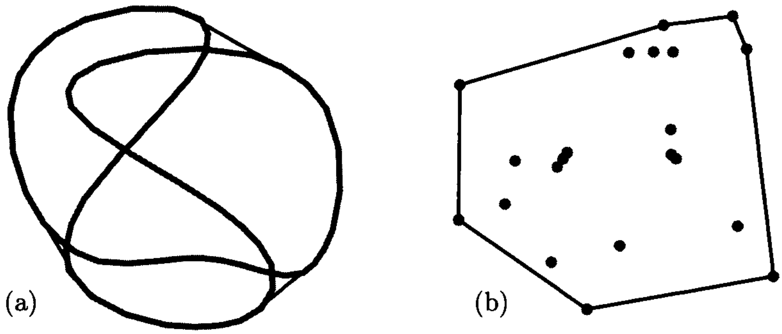


Figure 1.5. The convex hulls of (a) a continuous curve; (b) a discrete set of points.

Lemma 1.34. *Suppose a function f is piecewise $C^{(1)}$. The extremal points of a function f might occur for only the following values of x:*

- *$x = a$ and $x = b$, that is the interval endpoints,*
- *values of x for which $f'(x) = 0$,*
- *values of x for which $f'(x)$ does not exist.*

Definition 1.35. *A subset of $\mathbf{R}^3$, X is called* convex *if for all $x_1, x_2 \in X$, $(1-t)x_1 + tx_2 \in X$, for $t \in [0,1]$. That is the line segment connecting x_1 and x_2 lies entirely within the set X.*

Definition 1.36. *The* convex hull *of a set X is the smallest convex set containing X.*

If the set X is a finite set of points, the convex hull can be found by finding the line segment connecting each pair of points in the set (an n^2 operation count for the naive algorithm), and then finding out which ones form the boundary. There are more efficient algorithms, $O(n \log n)$, for finding convex hulls[68, 27].

Definition 1.37. *A function $f(x)$ is called* convex *on $[c,d]$ if for all $u, v \in [c,d]$, $f(\frac{u+v}{2}) \leq \frac{f(u)+f(v)}{2}$.*

Definition 1.38. *A function $f(x)$ is called* concave *on (c,d) if for all $u, v \in (c,d)$, $f(\frac{u+v}{2}) \geq \frac{f(u)+f(v)}{2}$.*

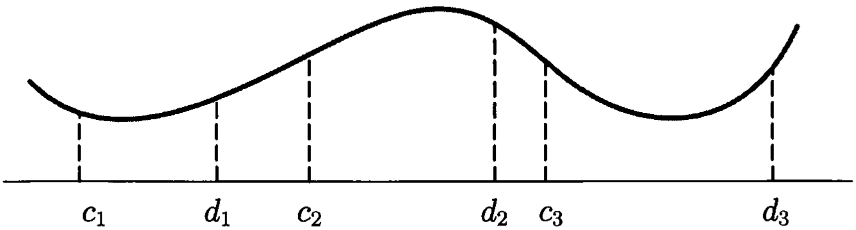


Figure 1.6. The function is convex on $(c_1, d_1) \cup (c_3, d_3)$ and concave on (c_2, d_2).

Theorem 1.39. *Suppose $f(x) \in C^{(2)}$. If $f''(x) > 0$ for $x \in (c, d)$, then $f(x)$ is convex on (c, d). If $f''(x) < 0$ for $x \in (c, d)$, then $f(x)$ is concave on (c, d).*

The implications of these results are that the signs and the values of the first and second derivatives yield important information about the shape of the curve.

1.5.1 Equations of Lines

In the plane, the implicit equation for a line is $ax + by + c = 0$. Using inner product notation yields $\langle (a, b, c), (x, y, 1) \rangle = 0$. Given a slope m which is not infinite, and a point (x_1, y_1) on the line, one has $(y - y_1) = m(x - x_1)$. It can be written more generally as, $a(x - x_1) + b(y - y_1) = \langle (a, b), (x - x_1, y - y_1) \rangle = 0$. To use parametric equations, a line can be represented using a direction vector $\boldsymbol{m}$ and one point (vector from origin) $\boldsymbol{p}$ that the line passes through. Thus $L(t) = t\boldsymbol{m} + \boldsymbol{p}$. Or one may use two points, $\boldsymbol{p_1}$ and $\boldsymbol{p_2}$, to yield $L(t) = t(\boldsymbol{p_2} - \boldsymbol{p_1}) + \boldsymbol{p_1} = (1 - t)\boldsymbol{p_1} + t\boldsymbol{p_2}$. This last form is called the *blending function* formulation. The derivations for these parametric lines occur in the introductory section on parametric functions.

What is the equation of a vector perpendicular to a line in $\mathbb{R}^2$? A unit vector that is perpendicular to a line, curve, or surface is called a *normal.* Whenever $b \neq 0$, the slope of the line $ax + by + c = 0$ is $-a/b$. It is clear that a perpendicular will have slope b/a (whenever $a \neq 0$). The vector (a, b) has that direction and, hence a vector with direction (a, b), having any length and any position, is perpendicular to the original line. We can look at this problem in a slightly different way.

Any point on the line $ax+by+c = 0$ must satisfy $\langle (a, b, c), (x, y, 1) \rangle = 0$. Now consider the situation where the line goes through the origin, that is, when $c = 0$. Then we can write $\langle (a, b), (x, y) \rangle = 0$, so the vector (a, b), is perpendicular to all points on the line. The case for $c \neq 0$ results simply in translates of the line but effects nether the slope nor the normal.

Example 1.40. Find the equation of a line through two points, $P_1 = (x_1, y_1)$ and $P_2 = (x_2, y_2)$.

The equation of a line is $ax + by + c = 0$ where we must determine a, b, and c.

P_1 on the line means $\langle (a, b, c), (x_1, y_1, 1) \rangle = 0$, and
P_2 on the line means $\langle (a, b, c), (x_2, y_2, 1) \rangle = 0$.

This means that considered as "3-space vectors" the vectors $(x_1, y_1, 1)$ and $(x_2, y_2, 1)$ must both be perpendicular to the vector (a, b, c). Thus, we can set $(a, b, c) = (x_1, y_1, 1) \times (x_2, y_2, 1)$. Furthermore, any multiple of (a, b, c) also works! □

Example 1.41. Find the intersection point between the two lines $a_1x + b_1y + c_1 = 0$ and $a_2x + b_2y + c_2 = 0$.
Denote the intersection point by $I = (x_I, y_I)$.

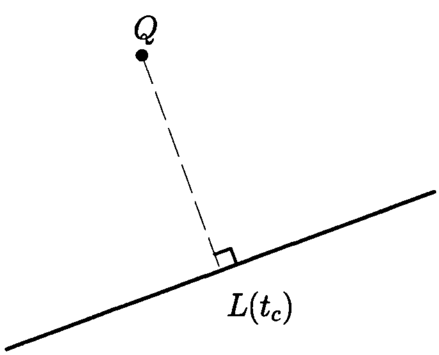
$a_1x_I + b_1y_I + c_1 = 0 = \langle (a_1, b_1, c_1), (x_I, y_I, 1) \rangle$ since I is on the first line, and
$a_2x_I + b_2y_I + c_2 = 0 = \langle (a_2, b_2, c_2), (x_I, y_I, 1) \rangle$ since I is on the second line.

Using this "modified" three space notation, the "point" $(x_I, y_I, 1)$ must be orthogonal to both (a_1, b_1, c_1) and (a_2, b_2, c_2). Hence, it lies along the "three space vector" $Q = (a_1, b_1, c_1) \times (a_2, b_2, c_2)$ which is perpendicular to both. The third coordinate of Q is $a_1b_2 - a_2b_1$, however, not 1.

To solve this let $Q' = Q/(a_1b_2 - a_2b_1)$. Q' has its third coordinate equal to 1, $(a_1, b_1, c_1) \cdot Q' = 0$, and $(a_2, b_2, c_2) \cdot Q' = 0$. Thus the x-coordinate of Q' is x_I, and the y-coordinate of Q' is y_I. □

Example 1.42. Find the distance from a point Q to a parametric line $L(t) = \boldsymbol{p} + t\boldsymbol{d}$.

Finding the distance is equivalent to finding the magnitude of the vector perpendicular to the line through the point.



Let $L(t_c)$ denote the point on the line closest to Q. We wish to discover the value of t_c. Consider $L(t_c) - \boldsymbol{Q}$. This is a vector perpendicular to the line L through the point Q. Since it is perpendicular to L,

$$0 = \big(L(t_c) - \boldsymbol{Q}\big) \cdot \boldsymbol{d} = (\boldsymbol{p}, \boldsymbol{d}) + t_c(\boldsymbol{d}, \boldsymbol{d}) - (\boldsymbol{Q}, \boldsymbol{d}).$$

So,

$$(\boldsymbol{Q}, \boldsymbol{d}) = (\boldsymbol{p}, \boldsymbol{d}) + t_c(\boldsymbol{d}, \boldsymbol{d}),$$

and

$$t_c = \frac{(\boldsymbol{Q}, \boldsymbol{d}) - (\boldsymbol{p}, \boldsymbol{d})}{(\boldsymbol{d}, \boldsymbol{d})}.$$

The point $L(t_c)$ is now known, Q is known, so the distance from the point to the line is just the distance between these two points. □

1.5.2 Equations of Planes

The explicit equation for a plane not perpendicular to the x-y plane is $ax + by + d = z$. The implicit equation, which can be used for any plane in $\boldsymbol{R}^3$ is $Ax+By+Cz+D = 0$. While it seems as if there are four degrees of freedom, that is not true. The same plane is specified by $(rA)x + (rB)y + (rC)z + (rD) = 0$ as is specified by $Ax + By + Cz + D = 0$, for any $r \neq 0$. Since a plane has three degrees of freedom, any three independent constraints specify a unique plane. Some of the more commonly used specifications are: three points, one point and one "normal" vector, and two direction vectors and one point.

The inner product formulation developed for specifying a line can be generalized to specify planar characteristics. If (x, y, z) is on the plane, then $\langle (A, B, C, D), (x, y, z, 1)\rangle = 0$. If the plane goes through the origin, $D = 0$, and the vector (A, B, C) is perpendicular to all points in the plane. If the plane is simply translated, its orientation is unchanged, so the vectors perpendicular to it will remain unchanged. Thus, (A, B, C) is perpendicular to the plane with equation $\langle (A, B, C, D), (x, y, z, 1)\rangle = 0$.

Suppose a perpendicular direction (A, B, C) to the plane is specified. Given a point $P_1 = (x_1, y_1, z_1)$, let $x' = x - x_1$, $y' = y - y_1$, and $z' = z - z_1$. This can be seen as a translation of the coordinate axes which puts the point P_1 at the origin of the new coordinate system. The equation for the plane in that new system is $A'x' + B'y' + C'z' + D' = 0$. If that point is on the plane, $D' = 0$ by the discussion above. Further, a translation does not change directions, being parallel to the original axes, so $A' = A$, $B' = B$ and $C' = C$. Thus, one has $\langle (A, B, C), (x - x_1, y - y_1, z - z_1)\rangle = 0$

for the equation of the plane. This is modified to $\langle (A, B, C), (x, y, z)\rangle - \langle (A, B, C), (x_1, y_1, z_1)\rangle = 0$, and D is known.

To find the plane equation when two direction vectors $\boldsymbol{v}$, $\boldsymbol{w}$ for the plane, and a point in the plane are specified, one can find the normal (perpendicular) direction $(A, B, C) = \boldsymbol{v} \times \boldsymbol{w}$. This problem then reduces to the previous case.

If one wants to specify a plane in a parametric formulation, one needs two direction vectors and a point in the plane. Consider $P(s, t) = P_0 + s\boldsymbol{u} + t\boldsymbol{v}$. The points specified as the "head" of these vectors are on the surface: P_0, $P_0 + \boldsymbol{u}$, and $P_0 + \boldsymbol{v}$. If $P_0 = 0$, it is clear that this is simply a plane spanned by $\boldsymbol{u}$ and $\boldsymbol{v}$. Hence, $P(s, t)$ is a plane translated away from the origin.

If three points P_0, P_1, and P_2 are specified, one can set $\boldsymbol{u} = P_1 - P_0$, and $\boldsymbol{v} = P_2 - P_0$.

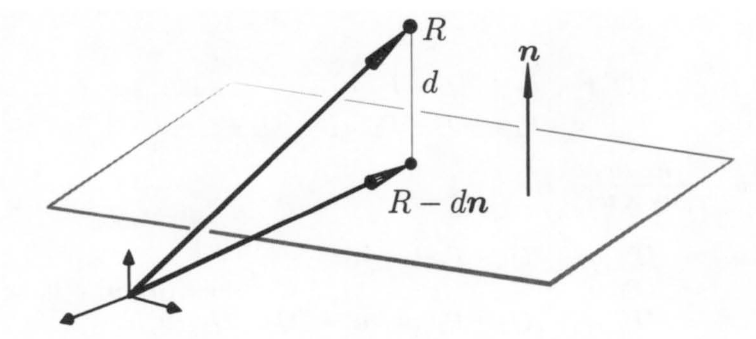
Example 1.43. Find the angle between two planes.

It will be shown in Exercise 5 that finding the angle, θ, between two planes is equivalent to finding the angle between the two plane normals, $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$. Hence, the solution is

$$\theta = \cos^{-1}\left(\frac{\boldsymbol{n_1} \cdot \boldsymbol{n_2}}{\|\boldsymbol{n_1}\|\ \|\boldsymbol{n_2}\|}\right).$$

□

Example 1.44. What is the distance from a point R to a plane?



Let $\boldsymbol{n}$ be a unit normal to the plane, and $P_0 = (x_0, y_0, z_0)$ be any point in the plane. The plane equation, then is,

$$\begin{aligned} 0 &= \boldsymbol{n} \cdot (x - x_0, y - y_0, z - z_0) \\ &= \boldsymbol{n} \cdot \left[(x, y, z) - (x_0, y_0, z_0)\right]. \end{aligned}$$

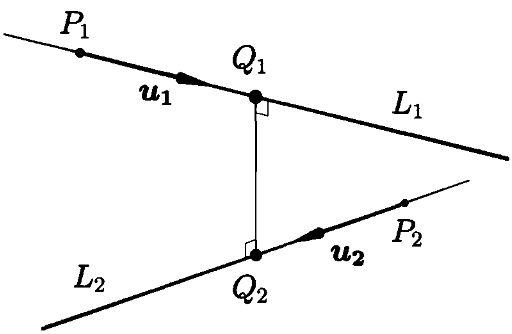
Suppose d is the unknown distance from R to the plane. Then, $\boldsymbol{R} - d\boldsymbol{n}$ is a vector from the origin whose head is a point in the plane and

$$\begin{aligned} 0 &= \boldsymbol{n} \cdot [\boldsymbol{R} - d\boldsymbol{n} - \boldsymbol{P}_0] \\ &= \boldsymbol{n} \cdot \boldsymbol{R} - \boldsymbol{n} \cdot \boldsymbol{P}_0 - d\boldsymbol{n} \cdot \boldsymbol{n} \\ &= \boldsymbol{n} \cdot \boldsymbol{R} - \boldsymbol{n} \cdot \boldsymbol{P}_0 - d, \end{aligned}$$

so $d = \boldsymbol{n} \cdot \boldsymbol{R} - \boldsymbol{n} \cdot \boldsymbol{P}_0$. □

Example 1.45. Find the common perpendicular to two 3-space skew lines L_1 and L_2.

Suppose L_1 has direction vector $\boldsymbol{u_1}$ through point P_1 and L_2 has direction vector $\boldsymbol{u_2}$ through point P_2.



Then

$$\begin{aligned} Q_1 &= P_1 + \|Q_1 - P_1\|\boldsymbol{u_1} \\ &= P_2 + \|Q_2 - P_2\|\boldsymbol{u_2} - \|Q_2 - Q_1\|\boldsymbol{u}, \end{aligned}$$

where $\boldsymbol{u} = \frac{\boldsymbol{u_1} \times \boldsymbol{u_2}}{\|\boldsymbol{u_1} \times \boldsymbol{u_2}\|}$, so

$$\begin{aligned} \langle Q_1, \boldsymbol{u} \rangle &= \langle P_1, \boldsymbol{u} \rangle + \|Q_1 - P_1\| \langle \boldsymbol{u_1}, \boldsymbol{u} \rangle \\ &= \langle P_1, \boldsymbol{u} \rangle & \text{since } \langle \boldsymbol{u_1}, \boldsymbol{u} \rangle = 0, \text{ and} \\ &= \langle P_2, \boldsymbol{u} \rangle + \|Q_2 - P_2\| \langle \boldsymbol{u_2}, \boldsymbol{u} \rangle - \|Q_2 - Q_1\| \langle \boldsymbol{u}, \boldsymbol{u} \rangle \\ &= \langle P_2, \boldsymbol{u} \rangle - \|Q_2 - Q_1\| & \text{since } \langle \boldsymbol{u_2}, \boldsymbol{u} \rangle = 0, \\ & & \text{and } \langle \boldsymbol{u}, \boldsymbol{u} \rangle = 1. \end{aligned}$$

Thus $\langle P_2 - P_1, \boldsymbol{u} \rangle = \|Q_2 - Q_1\|$, or $\|Q_2 - Q_1\| = \left\langle P_2 - P_1, \frac{\boldsymbol{u_1} \times \boldsymbol{u_2}}{\|\boldsymbol{u_1} \times \boldsymbol{u_2}\|} \right\rangle$.

□

Example 1.46. Find the common intersection $\boldsymbol{r}$ of the three planes with normals $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{w}$.

By the equations of the three planes

$$\begin{aligned}\langle \boldsymbol{r}, \boldsymbol{u} \rangle &= p_u, \\ \langle \boldsymbol{r}, \boldsymbol{v} \rangle &= p_v, \\ \langle \boldsymbol{r}, \boldsymbol{w} \rangle &= p_w.\end{aligned}$$

This yields just three linear equations in three unknowns. □

This section concludes with some food for thought. Consider $\boldsymbol{E} = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ such that $\langle \boldsymbol{e_i}, \boldsymbol{e_j} \rangle = \delta_{i,j}$ for i, $j = 1$, 2, 3, and ordered so that $\boldsymbol{e_1} \times \boldsymbol{e_2} = \boldsymbol{e_3}$, $\boldsymbol{e_2} \times \boldsymbol{e_3} = \boldsymbol{e_1}$, and $\boldsymbol{e_3} \times \boldsymbol{e_1} = \boldsymbol{e_2}$. This does not require that $\boldsymbol{e_1}$ be a unit vector in the direction of x, nor is an analogous constraint placed upon $\boldsymbol{e_2}$ or $\boldsymbol{e_3}$.

1. Is $\boldsymbol{E}$ a basis for $\boldsymbol{R}^3$?
2. For an arbitrary $\boldsymbol{v} \in \boldsymbol{R}^3$, consider $\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{e_1} \rangle \boldsymbol{e_1} + \langle \boldsymbol{v}, \boldsymbol{e_2} \rangle \boldsymbol{e_2} + \langle \boldsymbol{v}, \boldsymbol{e_3} \rangle \boldsymbol{e_3}$. Is $\boldsymbol{w} = \boldsymbol{v}$?
3. If, instead, one has three unit vectors $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}\}$, which form a basis for $\boldsymbol{R}^3$, but are not orthonormal, and $\boldsymbol{v} \in \boldsymbol{R}^3$, is $\boldsymbol{v}$ equal to $\langle \boldsymbol{v}, \boldsymbol{v_1} \rangle \boldsymbol{v_1} + \langle \boldsymbol{v}, \boldsymbol{v_2} \rangle \boldsymbol{v_2} + \langle \boldsymbol{v}, \boldsymbol{v_3} \rangle \boldsymbol{v_3}$?
4. Under what conditions on the basis is such a decomposition true for all elements of $\boldsymbol{R}^3$?

These problems are left as reinforcement exercises for the reader.

1.5.3 Polynomials

Definition 1.47. *For arbitrary complex numbers $a_0, \ldots, a_n$, with $a_n \neq 0$, a* polynomial, *p_n of degree n over the complex numbers is defined as a function of the form:*

$$p_n(z) = a_n z^n + a_{n-1} z^{n-1} + \cdots + a_1 z + a_0.$$

For the general complex polynomial, one has

Theorem 1.48. The Fundamental Theorem of Algebra. *If $n > 0$ and p_n is defined over the complex domain with complex coefficients and*

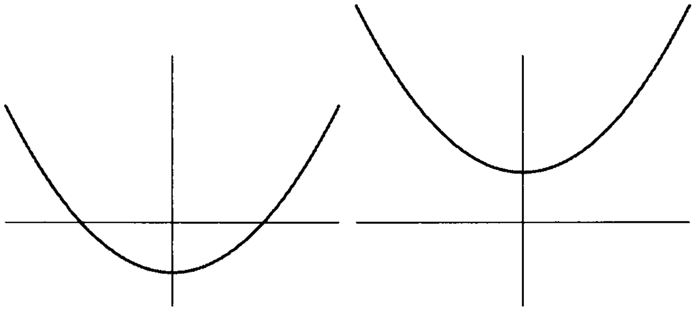


Figure 1.7. $p(x) = x^2 - 1$ $q(x) = x^2 + 1$

$a_n \neq 0$, then p_n has exactly n complex, not necessarily distinct, roots. Thus, one can write

$$p_n(z) = a_n(z - z_1)(z - z_2) \cdots (z - z_n),$$

where the values $z_1, \ldots, z_n$ are the zeros of the polynomial p_n.

If all the coefficients are real and the domain is restricted to the reals, then by Theorem 1.48 there are n roots possible, although the values may not be real.

Example 1.49. Consider the function $p(x) = x^2 - 1$. It has roots $\{1, -1\}$. The function $q(x) = x^2 + 1$, however has no real roots. Its complex roots are $\{i, -i\}$, where i is the square root of -1. □

The roots of a polynomial do not uniquely define that polynomial. For example, both $f(x) = (x - 1)$ and $g(x) = 3(x - 1)$ are straight lines that have roots at $x = 1$, but they have different slopes. In fact the coefficient of x is the slope of the line. Therefore, knowing the root and the slope uniquely defines the line. Interpolation with polynomials is based on the fundamental theorem of algebra.

An immediate corollary is

Corollary 1.50. *If a polynomial p_n of degree n vanishes (has roots) at more than n distinct points, then $p_n \equiv 0$, that is, p_n is identically zero.*

An immediate result is that the behavior of an n^{th} degree polynomial is determined by its function values at $n + 1$ points, since if two polynomials $p(x)$ and $q(x)$ of degree n agree at $n + 1$ points, then their difference is a polynomial of degree n with $n + 1$ zeros. The difference must, by the corollary, be identically zero, and $p(x) = q(x)$.

Within the scope of this book, the coefficients a_0, ..., a_n and the domain (and hence the range) will be real numbers.

Definition 1.51. *The* space of polynomials *of degree* n, $\mathcal{P}_n$, *is defined as*

$$\mathcal{P}_n = \left\{ p(x) = a_0 + a_1 x^1 + \cdots + a_n x^n : a_i \in \boldsymbol{R},\ i = 0, \ldots, n \right\}.$$

Definition 1.52. *A* bivariate polynomial of degree n *is a bivariate function* $f(u, v)$ *such that*

$$f(u, v) = \sum_{k=0}^{n} \sum_{i=0}^{k} p_{i,k} u^i v^{k-i}.$$

A bivariate polynomial is called bilinear, biquadratic, *or* bicubic *if the highest power in each of the variables is* 1, 2, *or* 3, *respectively.*

It is clear from the definition that $\mathcal{P}_n$ is a vector space in which each polynomial is a vector. This type of space is called a function space. Further it is clear from the definition of a polynomial that the set $\{1, x, x^2, \ldots, x^n\}$ spans the space. We ask whether this set forms a basis. That is, are the functions 1, x, x^2, ..., x^n linearly independent? We shall use straightforward reasoning. Suppose they are not. Then there are coefficients c_i, $i = 0$, ..., n not all zero so that

$$0 \equiv c_0 + c_1 x + \cdots + c_n x^n.$$

That is, the polynomial on the right evaluates to zero at all values of x. Now suppose c_k is the coefficient with the lowest order subscript which is nonzero. Then differentiating both sides k times gives

$$0 \equiv k! c_k + (k+1)k \ldots 2 c_{k+1} x + \cdots + n(n-1) \cdots (n-k+1) c_n x^{n-k}.$$

Evaluating the polynomial on the right at $x = 0$ yields that $k! c_k = 0$, or that $c_k = 0$. This contradicts the hypothesis, so 0 cannot be represented as a nontrivial polynomial and the powers of x are independent functions. Thus, $\{1, x, \ldots, x^n\}$ forms a basis for the space of polynomials, and $\mathcal{P}_n$ has dimension $n + 1$, and we have shown,

Lemma 1.53. *The dimension of* $\mathcal{P}_n$, *the space of polynomials of degree* n, *is* $n + 1$.

A more concise form equation for the fundamental theorem can be derived as follows. Suppose both $p(a) = 0$ and $p'(a) = 0$. Can we tell anything more about the form of the polynomial $p(x)$? Since $p(a) = 0$, $p(x) = (x-a)p_1(x)$. But since $p'(a) = 0$, $p'(x) = (x-a)p_2(x)$, by the fundamental theorem of algebra. But from the first form on $p(x)$,

$$\begin{aligned} p'(x) &= p_1(x) + (x-a)p_1'(x) \\ &= (x-a)p_2(x). \end{aligned}$$

We see, then that $p_1(x) = (x-a)q(x)$, and $p(x) = (x-a)^2 q(x)$. Following this line of reasoning gives:

Theorem 1.54. *Suppose $p(x)$ is a polynomial of degree n with distinct real roots $x_1, \ldots, x_k$, and suppose $p^{(j)}(x_i) = 0$, $j = 0, \ldots, s_i$, $i = 1, \ldots, k$. Then*

$$p(x) = (x-x_1)^{s_1+1}(x-x_2)^{s_2+1} \cdots (x-x_k)^{s_k+1} q(x)$$

where $q(x)$ is of degree $n - k - \sum_{i=1}^{k} s_i$.

Proof: The proof is by induction. We saw above that if a is a root of both $p(x)$ and $p'(x)$, then $p(x) = (x-a)^2 q_1(x)$. Suppose we have shown that if a is a root of $p^{(j)}(x)$, $j = 0, \ldots, m$, then $p(x) = (x-a)^{m+1}q(x)$. Suppose now that in addition, a is a root of $p^{(m+1)}(x)$. We know that $p'(x) = (x-a)^{m+1}r(x)$, since a is a root of $[p']^{(j)}(x)$, $j = 0, \ldots, m$, but we also know

$$\begin{aligned} p'(x) &= (m+1)(x-a)^m q(x) + (x-a)^{m+1}q'(x) \\ &= (x-a)^m \left[(m+1)q(x) + (x-a)q'(x)\right] \\ &= (x-a)^m (x-a) r(x). \end{aligned}$$

Since the last two lines are equal, for $x \neq a$, we see that $(x-a)r(x) = (m+1)q(x) + (x-a)q'(x)$, and so $q(x) = (x-a)z(x)$, and the result is proved for a single root.

Now, if there is more than one root, we then apply the theorem to the polynomial $z(x)$ that is the remainder at a different value and get a corresponding decomposition. After applying it to all k distinct roots, the result is proved. ∎

1.5.4 Rational Functions

Definition 1.55. *A function $f(x)$ is called a* rational function *if $f(x) = p(x)/q(x)$, where both p and q are polynomials.*

Example 1.56. A simple rational function can be constructed as the quotient of two linear polynomials, that is, $f(x) = (ax+b)/(cx+d)$. □

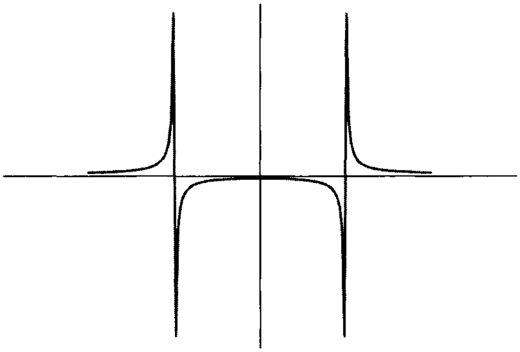


Figure 1.8. A simple rational function $(x^2+1)/(x^2-1)$.

The properties of the rational are not completely defined by the separate properties of the numerator and denominator taken separately. While the roots of the numerator may be zeros of the rational function and the roots of the denominator may be poles (infinite asymptotes) of the rational function, this may also not occur. That is, these are necessary but not sufficient conditions. If a root is common to both numerator and denominator then the number of repetitions in each may decide the final root/pole configuration.

Example 1.57. The equation

$$\frac{(x-1)(x+1)}{(x-1)(x-2)} = 0$$

has a root at $x = -1$ and a pole at $x = 2$, and is not defined at $x = 1$. However it is possible to define a function

$$f(x) = \begin{cases} \frac{(x-1)(x+1)}{(x-1)(x-2)} & x \neq 1, \\ -2 & x = 1, \end{cases}$$

which is continuous at $x = 1$.

The function $g(x) = \frac{(x-1)^2}{(x-1)(x-2)}$ has a root at $x = 1$ and a pole at $x = 2$. □

The rationals also have the feature that if $f(x) = p(x)/q(x)$, then $f(x) = ap(x)/aq(x)$ as well for all real numbers $a \neq 0$. This can lead to confusion over the number of points needed to uniquely specify a rational function.

1.6 Parametric or Vector Functions

Definition 1.58. *A subset U of $\boldsymbol{R}^2$ is called open if for every point $(a,b) \in U$ there exists an $\epsilon > 0$ such that, if $(x-a)^2+(y-b)^2 < \epsilon$, then $(x,y) \in U$. That is, there is a boundaryless (open) disk around each point contained entirely in the set U.*

Definition 1.59. *Suppose a vector basis for $\boldsymbol{R}^i$ is $\boldsymbol{e}^j$, $j = 1, \ldots, i$. Let U be an open subset of $\boldsymbol{R}^i$, $i = 1, 2$ and let function $\boldsymbol{f}(\boldsymbol{x}) : U \to \boldsymbol{R}^j$, where $j = 1, 2, 3$. This can always be written in vector notation as $\boldsymbol{f}(\boldsymbol{x}) = \sum_{k=1}^{j} f_k(\boldsymbol{x})\boldsymbol{e}^k = (f_1(\boldsymbol{x}), f_2(\boldsymbol{x}), \ldots, f_j(\boldsymbol{x}))$. The functions f_k are called the* coordinate functions *of the vector function (f).*

Definition 1.60. *A vector function $\boldsymbol{f}$ is called* continuous *at $\boldsymbol{x}^0 = (x_1^0, x_2^0)$ if for every $\epsilon > 0$ there exists $\delta > 0$ such that if $\boldsymbol{x} = (x_1, x_2)$ then $\|\boldsymbol{f}(\boldsymbol{x}) - \boldsymbol{f}(\boldsymbol{x}^0)\| < \epsilon$ whenever $\|\boldsymbol{x} - \boldsymbol{x}^0\| < \delta$, where $\|\boldsymbol{x} - \boldsymbol{x}^0\| = \sqrt{(x_1 - x_1^0)^2 + (x_2 - x_2^0)^2}$.*

It is clear that a vector function is continuous if and only if its coordinate functions are continuous.

Suppose that $\boldsymbol{f} : U \to \boldsymbol{R}^3$, and further, suppose that we can define functions $x_1 = \theta_1(t)$ and $x_2 = \theta_2(t)$ where $\theta_1, \theta_2 : \mathrm{I} \to U$, where I is an interval in $\boldsymbol{R}^1$. Then the function $\gamma(t) = \boldsymbol{f}(x_1, x_2) = \boldsymbol{f}(x_1(t), x_2(t)) = \boldsymbol{f}(\theta_1(t), \theta_2(t))$ is a space curve whose image lies in the image of the function $\boldsymbol{f}$.

Definition 1.61. *Consider the set $U_{2,c} = \{(x_1, c) : (x_1, c) \in U\}$, and $U_{1,k} = \{(k, x_2) : (k, x_2) \in U\}$. Consider the space curve, $\gamma_c(x_1)$ defined by $\gamma_c(x_1) = \boldsymbol{f}(x_1, c)$, a curve in the image $\boldsymbol{f}$. Each constant c defines a distinct curve. Analogously, for a constant k, $\phi_k(x_2) = \boldsymbol{f}(k, x_2)$ defines a curve in the image $\boldsymbol{f}$, where each constant k defines a distinct curve. Each γ_c curve is has domain parallel to the x_1 axis in the $x_1 - x_2$ plane traces out a curve on the surface given by the image of $\boldsymbol{f}$. Analogously, each ϕ_k has a domain parallel to the x_2 axis and traces out a curve on the surface. The set of curves formed by the γ_c and ϕ_k on the surface is called a* curvilinear coordinate system.

Unlike curves for which derivatives can be defined only in one direction, a surface has an infinite number of curves through a point, and thus the meaning of derivative must be adapted.

Definition 1.62. *Suppose that $\boldsymbol{f}$ is defined on an open set U in $\mathbb{R}^2$, and let $\boldsymbol{x}^0$ be some point in U, with $\boldsymbol{u}^0$ a nonzero vector in U. The* directional derivative *of $\boldsymbol{f}$ at $\boldsymbol{x}^0$ in the direction of $\boldsymbol{u}^0$ is the vector*

$$D_{\boldsymbol{u}^0}\boldsymbol{f}(\boldsymbol{x}^0) = \lim_{h\to 0} \frac{\boldsymbol{f}(\boldsymbol{x}^0 + h\boldsymbol{u}^0) - \boldsymbol{f}(\boldsymbol{x}^0)}{h}$$

whenever that limit exists.

This definition has the same effect as defining a univariate derivative of a space curve of a variable h where $\gamma(h) = \boldsymbol{f}\big(\boldsymbol{x}^0 + h\boldsymbol{u}^0\big)$. This curve lies in the surface $\boldsymbol{f}$ and has as domain an open interval around zero. In general, if for each point $\boldsymbol{x} \in U$, $D_{\boldsymbol{u}^0}\boldsymbol{f}(\boldsymbol{x})$ exists, then $\boldsymbol{f}$ is said to have a directional derivative in the direction $\boldsymbol{u}^0$ in U.

In particular when $\boldsymbol{u}^0 = (1,0)$, the result is a derivative with respect to the first direction in the curvilinear coordinate system, and when $\boldsymbol{u}^0 = (0,1)$, the result is a derivative with respect to the second.

Written $\frac{\partial \boldsymbol{f}}{\partial x_1}$ and $\frac{\partial \boldsymbol{f}}{\partial x_2}$, respectively, these derivatives are called the *partial derivatives* of the function $\boldsymbol{f}$ with respect to the first and second coordinates, respectively.

The world of multivariate functions is much more complicated than that of univariate functions. Since the directional derivative depends only on the values of the function along an open line interval near a point, a function $\boldsymbol{f}$ can have derivatives in every direction at a point $\boldsymbol{x}^0$, but not be continuous at $\boldsymbol{x}^0$. Remember, the continuity depends on the actions of the function in a two-dimensional neighborhood of the point.

Theorem 1.63. *A vector function $\boldsymbol{f} = (f_1, f_2, f_3)$ is said to be of class $C^{(k)}$ if each coordinate function is of class $C^{(k)}$. A bivariate scalar function f is of class $C^{(k)}$ if all partial derivatives of order less than or equal to k exist and are continuous independent of the order of differentiation .*

Suppose that it is desired to perform a change of variables, a *reparametrization*, as it is called.

Theorem 1.64. Chain Rule. *For a parametric function $\boldsymbol{f}$, defined as $\boldsymbol{f}(x_1,x_2) = (f_1(x_1,x_2), f_2(x_1,x_2), f_3(x_1,x_2)) \in C^{(1)}$. If $x_1 = x_1(t)$ and $x_2 = x_2(t)$ and $x_1(t)$, $x_2(t) \in C^{(1)}$ then*

$$\begin{aligned}
\frac{d\boldsymbol{f}}{dt} &= \frac{\partial \boldsymbol{f}}{\partial x_1}\frac{dx_1}{dt} + \frac{\partial \boldsymbol{f}}{\partial x_2}\frac{dx_2}{dt} \\
&= \left(\frac{\partial f_1}{\partial x_1}, \frac{\partial f_2}{\partial x_1}, \frac{\partial f_3}{\partial x_1}\right)\frac{dx_1}{dt} + \left(\frac{\partial f_1}{\partial x_2}, \frac{\partial f_2}{\partial x_2}, \frac{\partial f_3}{\partial x_2}\right)\frac{dx_2}{dt}
\end{aligned}$$

$$= \begin{bmatrix} \frac{dx_1}{dt} & \frac{dx_2}{dt} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix}.$$

If $x_1 = x_1(u_1, u_2)$ *and* $x_2 = x_2(u_1, u_2)$ *and* $x_1(u_1, u_2)$, $x_2(u_1, u_2) \in C^{(1)}$ *then*

$$\begin{aligned} \frac{\partial f}{\partial u_1} &= \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial u_1} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial u_1} \\ &= \left(\frac{\partial f_1}{\partial x_1}, \frac{\partial f_2}{\partial x_1}, \frac{\partial f_3}{\partial x_1}\right)\frac{\partial x_1}{\partial u_1} + \left(\frac{\partial f_1}{\partial x_2}, \frac{\partial f_2}{\partial x_2}, \frac{\partial f_3}{\partial x_2}\right)\frac{\partial x_2}{\partial u_1} \\ &= \begin{bmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_2}{\partial u_1} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix}. \end{aligned}$$

Analogously,

$$\frac{\partial f}{\partial u_2} = \begin{bmatrix} \frac{\partial x_1}{\partial u_2} & \frac{\partial x_2}{\partial u_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix}.$$

The fundamentally important matrix

$$\begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix}$$

is called the *Jacobian matrix.* Further,

$$\begin{bmatrix} \frac{\partial f}{\partial u_1} \\ \frac{\partial f}{\partial u_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_2}{\partial u_1} \\ \frac{\partial x_1}{\partial u_2} & \frac{\partial x_2}{\partial u_2} \end{bmatrix} \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \frac{\partial f_3}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \frac{\partial f_3}{\partial x_2} \end{bmatrix}$$

and the matrix

$$\begin{bmatrix} \frac{\partial x_1}{\partial u_1} & \frac{\partial x_2}{\partial u_1} \\ \frac{\partial x_1}{\partial u_2} & \frac{\partial x_2}{\partial u_2} \end{bmatrix}$$

is called the *Jacobian matrix of the reparametrization.* Its determinant is called the *Jacobian,* $\frac{\partial(x_1,x_2)}{\partial(u_1,u_2)}$, *of the transformation.*

1.6.1 Function Characteristics

In order to facilitate later proofs, we shall state without complete proofs some of the following theorems from calculus:

Theorem 1.65. Rolle's Theorem. *Suppose $f \in C[a,b]$ and that $f'(x)$ exists at each point of (a,b) and $f(a) = f(b)$. Then there exists a point ζ, $a < \zeta < b$ such that $f'(\zeta) = 0$.*

Proof: If $f(x) \equiv f(a)$, then f is a constant function and $f' \equiv 0$. Now, suppose there exists x such that $f(x) > f(a)$. The fact that f is continuous on $[a,b]$ implies that f achieves a maximum value, say at ζ, in (a,b). Since f' exists for all points in (a,b) that means that $f'(\zeta) = 0$. ■

Theorem 1.66. Mean Value Theorem. *Let $f \in C[a,b]$ such that $f'(x)$ exists at each point of (a,b). Then there exists a point ζ, $a < \zeta < b$ such that $(b-a)f'(\zeta) = f(b) - f(a)$.*

Proof: Consider the function $g(x) = f(x) + (b-x)\frac{f(b)-f(a)}{b-a}$. Then $g(a) = f(a) + \big(f(b) - f(a)\big) = f(b)$ and $g(b) = f(b)$. By Rolle's theorem, there exists ζ such that $g'(\zeta) = 0$. But, $g'(x) = f'(x) - \frac{f(b)-f(a)}{b-a}$ so at ζ, $f'(\zeta) = \frac{f(b)-f(a)}{b-a}$. ■

Theorem 1.67. Generalized Rolle's Theorem. *For $2 \leq n$, let $f \in C[a,b]$ such that $f^{(n-1)}$ exists for each point of (a,b). If there exists $a \leq x_1 < x_2 < \cdots < x_n \leq b$ such that $f(x_1) = f(x_2) = \cdots = f(x_n)$, then there exists ζ, $x_1 < \zeta < x_n$ such that $f^{(n-1)}(\zeta) = 0$.*

Proof: Apply Rolle's theorem $n-1$ times. ■

Theorem 1.68. Taylor's Theorem. *For $f \in C^{(n+1)}[a,b]$ then for all x_0, $x \in [a,b]$,*

$$\begin{aligned} f(x) &= f(x_0) + f'(x_0)(x-x_0) + \frac{f''(x_0)}{2}(x-x_0)^2 + \cdots \\ &\quad + \frac{f^{(n)}(x_0)}{n!}(x-x_0)^n + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n\,dt. \end{aligned}$$

The use of Taylor's theorem requires explicit evaluation at only one point, and other knowledge of the $(n+1)^{st}$ derivative to bound the integral. The n^{th} degree polynomial part of the expansion is called the *Taylor* *polynomial approximation about x_0 to f of degree n* and the integral part is called the *remainder*.

Proof: From the fundamental theorem of calculus, $f(x) - f(x_0) = \int_{x_0}^{x} f'(t)\,dt$. Rewriting yields

$$f(x) = f(x_0) + \int_{x_0}^{x} f'(t)dt.$$

We integrate by parts with $u = f'(t)$, $du = f''(t)dt$ and $dv = dt$, $v = (t-x)$ to get

$$\begin{aligned} f(x) &= f(x_0) + \left[f'(t)(t-x)\right]\Big|_{x_0}^{x} - \int_{x_0}^{x} (t-x)f''(t)dt \\ &= f(x_0) + f'(x_0)(x-x_0) + \int_{x_0}^{x} (x-t)f''(t)dt. \end{aligned}$$

Integrating by parts again we see

$$\int_{x_0}^{x} (x-t)f''(t)dt = f''(x_0)\frac{(x-x_0)^2}{2!} + \int_{x_0}^{x} \frac{(x-t)^2}{2} f^{(3)}(t)dt$$

so

$$f(x) = f(x_0) + f'(x_0)(x-x_0) + f''(x_0)\frac{(x-x_0)^2}{2!} + \int_{x_0}^{x} \frac{(x-t)^2}{2} f^{(3)}(t)dt.$$

Using the induction hypothesis we have

$$\begin{aligned} f(x) &= \sum_{j=0}^{n-1} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \int_{x_0}^{x} \frac{f^{(n)}(t)}{(n-1)!}(x-t)^{n-1}dt \\ &= \sum_{j=0}^{n-1} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j \\ &\quad + \frac{1}{n!}\left(-(x-t)^n f^{(n)}(t)\,\Big|_{t=x_0}^{x} + \int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n dt\right) \\ &= \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \frac{1}{n!}\int_{x_0}^{x} f^{(n+1)}(t)(x-t)^n dt \\ &= \sum_{j=0}^{n} \frac{f^{(j)}(x_0)}{j!}(x-x_0)^j + \frac{1}{n!}\int_{x_0}^{b} f^{(n+1)}(t)(x-t)_+^n dt. \end{aligned}$$

Notice that the integrals in the last two lines have different beginning and ending points, and so the plus function notation is used. ■

Definition 1.69. *The function* $(x-t)_+^n$ *defined by*

$$(x-t)_+^n = \begin{cases} (x-t)^n & \text{if } x > t \\ 0 & \text{otherwise} \end{cases}$$

is called the n^{th} *degree* plus function.

Note that by this definition, $(0)_+^0 = 0$. For $n = 0$, it is clear that $(x-t)_+^0 = 1$,if $t < x$. That is, the function is continuous and is a single polynomial, namely the constant 1 for t values less than x, and it is the constant 0 for values of t greater than or equal to x. Thus, for $n = 0$, the function is discontinuous at $x = t$, i.e., in $C^{(-1)}$. Considered as a function of t for fixed x, it is *right continuous.* That is, $\lim_{t\to x^+}(x-t)_+^0 = 0 = (x-x)_+^0$. Considered as a function of x, the function is *left continuous* since $\lim_{x\to t^-}(x-t)_+^0 = 0 = (t-t)_+^0$ but $\lim_{x\to t^+}(x-t)_+^0 = 1$.

Next consider $f(t) = (x-t)_+$, always as a function of t. For $t < x$, $f(t) = (x-t)$, a polynomial, and is continuous and differentiable. For $t > x$, $f(t) = 0$, and this is also continuous and differentiable. Now at $t = x$, the function is continuous since $0 = f(x) = \lim_{t\to x^+}(x-t)$. However, the function does not have a derivative at $t = x$, since the derivative on the left is the zero function, and the derivative on the right is 1. Thus $(x-t)_+$ is continous and continuously differentiable everywhere but at $t = x$, where the derivative is right continuous.

Now, using induction, suppose that $(x-t)_+^{n-1}$ is in $C^{(n-2)}$ at $t = x$. Then, for $t \neq x$,

$$\frac{d}{dx}(x-t)_+^n = n(x-t)_+^{n-1}.$$

Since the right hand side is $C^{(n-2)}$, then the derivative clearly exists at $t = x$, and we find:

Lemma 1.70. *For positive integers n, $(x-t)_+^n$ is contained in $C^{(n-1)}$, and as a function of t, its n^{th} derivative is continuous everywhere except $t = x$ where it is right continuous. As a function of x, its n^{th} derivative is continuous everywhere except at $x = t$, where it is left continuous.*

Exercises

1. Suppose $\boldsymbol{V}$ is a finite-dimensional vector space and $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$ are two bases for $\boldsymbol{V}$. Show that $\boldsymbol{B_1}$ and $\boldsymbol{B_2}$ *must* have the same number of elements.

2. Is $C^{(0)}[a, b]$ a finite-dimensional vector space? If so prove it, if not explain why not.

3. Show that the definition of scalar multiplication of an $\boldsymbol{R}^3$ vector $\boldsymbol{v}$ by a scalar $c > 0$ makes a new vector $\boldsymbol{w}$ having the same direction as $\boldsymbol{v}$ with a magnitude c times as long. What happens when c is negative?

4. For vectors $\boldsymbol{v}$, $\boldsymbol{w} \in \boldsymbol{R}^3$ show that $\|\boldsymbol{v} \times \boldsymbol{w}\|^2 = \|\boldsymbol{v}\|^2 \|\boldsymbol{w}\|^2 - \|(\boldsymbol{v}, \boldsymbol{w})\|^2$.

5. If $\boldsymbol{P_1}$ and $\boldsymbol{P_2}$ are two planes with normals $\boldsymbol{n_1}$ and $\boldsymbol{n_2}$, respectively, show that finding the angle between the two planes is equivalent to finding the angle between the two plane normals.

6. Consider $\boldsymbol{E} = \{\boldsymbol{e_1}, \boldsymbol{e_2}, \boldsymbol{e_3}\}$ such that $\langle \boldsymbol{e_i}, \boldsymbol{e_j} \rangle = \delta_{i,j}$ for i, $j = 1$, 2, 3, and ordered so that $\boldsymbol{e_1} \times \boldsymbol{e_2} = \boldsymbol{ecd_3}$, $\boldsymbol{e_2} \times \boldsymbol{e_3} = \boldsymbol{e_1}$, and $\boldsymbol{e_3} \times \boldsymbol{e_1} = \boldsymbol{e_2}$. This does not require that $\boldsymbol{e_1}$ be a unit vector in the direction of x, nor is an analogous constraint placed on $\boldsymbol{e_2}$ or $\boldsymbol{e_3}$.

   (a) Is $\boldsymbol{E}$ a basis for $\boldsymbol{R}^3$?

   (b) For an arbitrary $\boldsymbol{v} \in \boldsymbol{R}^3$, consider $\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{e_1} \rangle \boldsymbol{e_1} + \langle \boldsymbol{v}, \boldsymbol{e_2} \rangle \boldsymbol{e_2} + \langle \boldsymbol{v}, \boldsymbol{e_3} \rangle \boldsymbol{e_3}$. Is $\boldsymbol{w} = \boldsymbol{v}$?

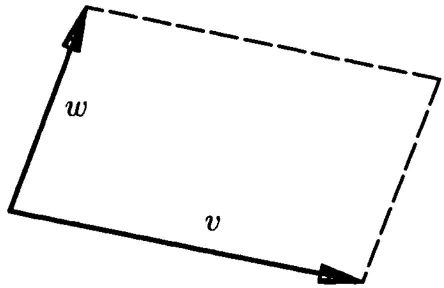
   (c) If instead one has three unit vectors $\{\boldsymbol{v_1}, \boldsymbol{v_2}, \boldsymbol{v_3}\}$, which form a basis for $\boldsymbol{R}^3$, but which is not orthonormal, and $\boldsymbol{v} \in \boldsymbol{R}^3$, if $\boldsymbol{w} = \langle \boldsymbol{v}, \boldsymbol{v_1} \rangle \boldsymbol{v_1} + \langle \boldsymbol{v}, \boldsymbol{v_2} \rangle \boldsymbol{v_2} + \langle \boldsymbol{v}, \boldsymbol{v_3} \rangle \boldsymbol{v_3}$, is $\boldsymbol{v} = \boldsymbol{w}$?

   (d) Under what conditions on the basis is such a decomposition true for all elements of $\boldsymbol{R}^3$?

   Give a proof for your response when it indicates something is true or not true all the time, and counterexamples when your response indicates that something contrary can occur.

7. Prove Lemma 1.13. That is, show that $\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \sum_{i=1}^{n} r_{v,i} r_{w,i}$, for $r_{v,i}$, $r_{w,i}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ defined as in the lemma.

8. Show that if $\boldsymbol{v}$ and $\boldsymbol{w}$ are vectors then $\|\boldsymbol{v} \times \boldsymbol{w}\|$ is equal to the area of the parallelogram formed with sides $\boldsymbol{v}$ and $\boldsymbol{w}$:



9. Show that if $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$ are vectors, then $|\langle \boldsymbol{u}, \boldsymbol{v} \times \boldsymbol{w}\rangle|$ is the volume of the parallelopiped formed with edges $\boldsymbol{u}$, $\boldsymbol{v}$, and $\boldsymbol{w}$.

10. (a) Is $T\big((x,y)\big) = (x+a, y+b)$ a linear transformation? Prove your answer.

    (b) Is $T\big((x,y,w)\big) = (x/w, y/w)$ a linear transformation? Prove your answer.

11. Suppose S and T are two linear transformations from $\boldsymbol{V}$ to $\boldsymbol{W}$ and $r \in \boldsymbol{R}$. Define $Q(v) = S(v) + T(v)$ and $R(v) = r\big(S(v)\big)$. Show that Q and R are both linear transformations. They are also written $Q = S + T$ and $R = r \cdot S$.

    Further prove that $L(\boldsymbol{V}, \boldsymbol{W})$, the set of linear transformations from $\boldsymbol{V}$ to $\boldsymbol{W}$, is a vector space with $+$ and $\cdot$ defined as above.

12. Prove that composition of two linear transformations, $S \cdot T(\boldsymbol{u}) = S(T(\boldsymbol{u}))$, is a linear transformation. Also, show by counterexample that composition of linear transformations is not a commutative operation.

13. Let $\{\boldsymbol{e}_1, \boldsymbol{e}_2, \boldsymbol{e}_3\}$ be a basis for $\boldsymbol{R}^3$. Given two vectors $\boldsymbol{v}, \boldsymbol{w} \in \boldsymbol{R}^3$, show that $\boldsymbol{v} = \boldsymbol{w}$ if and only if

$$\langle \boldsymbol{v}, \boldsymbol{e}_i\rangle = \langle \boldsymbol{w}, \boldsymbol{e}_i\rangle, \qquad \text{for i = 1,2,3.}$$

14. Let $\{\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3\}$ be a basis for $\boldsymbol{R}^3$, and let $\boldsymbol{w} \in \boldsymbol{R}^3$ be arbitrary. Find and justify necessary and sufficient conditions on the basis so that

$$\boldsymbol{w} = \langle \boldsymbol{w}, \boldsymbol{u}_1\rangle \boldsymbol{u}_1 + \langle \boldsymbol{w}, \boldsymbol{u}_2\rangle \boldsymbol{u}_2 + \langle \boldsymbol{w}, \boldsymbol{u}_3\rangle \boldsymbol{u}_3.$$