Interpolation and Approximation by Polynomials

George M. Phillips

Springer

Canadian Mathematical Society
Société mathématique du Canada

Springer
New York
Berlin
Heidelberg
Hong Kong
London
Milan
Paris
Tokyo

CMS Books in Mathematics

Ouvrages de mathématiques de la SMC

1 HERMAN/KUČERA/ŠIMŠA Equations and Inequalities

2 ARNOLD Abelian Groups and Representations of Finite Partially Ordered Sets

3 BORWEIN/LEWIS Convex Analysis and Nonlinear Optimization

4 LEVIN/LUBINSKY Orthogonal Polynomials for Exponential Weights

5 KANE Reflection Groups and Invariant Theory

6 PHILLIPS Two Millennia of Mathematics

7 DEUTSCH Best Approximation in Inner Product Spaces

8 FABIAN ET AL. Functional Analysis and Infinite-Dimensional Geometry

9 KŘÍŽEK/LUCA/SOMER 17 Lectures on Fermat Numbers

10 BORWEIN Computational Excursions in Analysis and Number Theory

11 REED/SALES (Editors) Recent Advances in Algorithms and Combinatorics

12 HERMAN/KUČERA/ŠIMŠA Counting and Configurations

13 NAZARETH Differentiable Optimization and Equation Solving

14 PHILLIPS Interpolation and Approximation by Polynomials

George M. Phillips

Interpolation and Approximation by Polynomials

With 22 Illustrations

Springer

George M. Phillips
Mathematical Institute
University of St. Andrews
St. Andrews KY16 9SS
Scotland

Mathematics Subject Classification (2000): 00A05, 01A05

Library of Congress Cataloging-in-Publication Data
Phillips, G.M. (George McArtney)
Interpolation and approximation by polynomials / George M. Phillips.
p. cm. — (CMS books in mathematics ; 14)
Includes bibliographical references and index.
ISBN 0-387-00215-4 (alk. paper)
1. Numerical analysis. 2. Approximation theory. I. Title. II. Series.
QA297 .P518 2003
519—dc21 2002042735

ISBN 0-387-00215-4 Printed on acid-free paper.

Printed in the United States of America.

9 8 7 6 5 4 3 2 1 SPIN 10903674

Typesetting: Pages created by the author using a Springer TEX macro package.

www.springer-ny.com

Springer-Verlag New York Berlin Heidelberg
A member of BertelsmannSpringer Science+Business Media GmbH

To Rona

Elle est à toi, cette chanson.

This page intentionally left blank

Preface

This book is intended as a course in numerical analysis and approximation theory for advanced undergraduate students or graduate students, and as a reference work for those who lecture or research in this area. Its title pays homage to *Interpolation and Approximation* by Philip J. Davis, published in 1963 by Blaisdell and reprinted by Dover in 1976. My book is less general than Philip Davis's much respected classic, as the qualification "by polynomials" in its title suggests, and it is pitched at a less advanced level.

I believe that no one book can fully cover all the material that *could* appear in a book entitled *Interpolation and Approximation by Polynomials*. Nevertheless, I have tried to cover most of the main topics. I hope that my readers will share my enthusiasm for this exciting and fascinating area of mathematics, and that, by working through this book, some will be encouraged to read more widely and pursue research in the subject. Since my book is concerned with polynomials, it is written in the language of classical analysis and the only prerequisites are introductory courses in analysis and linear algebra.

In deciding whether to include a topic in any book or course of lectures, I always ask myself, Is the proposed item mathematically *interesting*? Paradoxically, utility is a useless guide. For instance, why should we discuss interpolation nowadays? Who uses it? Indeed, how many make direct use of numerical integration, orthogonal polynomials, Bernstein polynomials, or techniques for computing various best approximations? Perhaps the most serious *users* of mathematics are the relatively small number who construct, and the rather larger number who apply, specialist mathematical packages, including those for evaluating standard functions, solving systems of linear

equations, carrying out integrations, solving differential equations, drawing surfaces with the aid of CAGD (computer-aided geometrical design) techniques, and so on. However, it is all too easy to make use of such packages without understanding the mathematics on which they are based, or their limitations, and so obtain poor, or even meaningless, results. Many years ago someone asked my advice on a mathematical calculation that he was finding difficult. I gently pointed out that his result was invalid because the series he was summing was divergent. He responded honestly that, faced with an infinite series, his strategy was always to compute the sum of the first hundred terms.

There are many connections between the various chapters and sections in this book. I have sought to emphasize these interconnections to encourage a deeper understanding of the subject. The first topic is interpolation, from its precalculus origins to the insights and advances made in the twentieth century with the study of Lebesgue constants. Unusually, this account of interpolation also pursues the direct construction of the interpolating polynomial by solving the system of linear equations involving the Vandermonde matrix. How could we dream of despising a study of interpolation, when it is so much at the centre of the development of the calculus? Our understanding of the interpolating polynomial leads us naturally to a study of integration rules, and an understanding of Gaussian integration rules requires knowledge of orthogonal polynomials, which are at the very heart of classical approximation theory. The chapter on numerical integration also includes an account of the Euler–Maclaurin formula, in which we can use a series to estimate an integral or vice versa, and the justification of this powerful formula involves some particularly interesting mathematics, with a forward reference to splines. The chapter devoted to orthogonal polynomials is concerned with best approximation, and concentrates on the Legendre polynomials and least squares approximations, and on the Chebyshev polynomials, whose minimax property leads us on to minimax approximations.

One chapter is devoted to Peano kernel theory, which was developed in the late nineteenth century and provides a special calculus for creating and justifying error terms for various approximations, including those generated by integration rules. This account of Peano kernel theory is rather more extensive than that usually given in a textbook, and I include a derivation of the error term for the Euler–Maclaurin formula. The following chapter extends the topic of interpolation to several variables, with most attention devoted to interpolation in two variables. It contains a most elegant generalization of Newton's divided difference formula plus error term to a triangular set of points, and discusses interpolation formulas for various sets of points in a triangle. The latter topic contains material that was first published in the late twentieth century and is justified by geometry dating from the fourth century AD, towards the very end of the golden millennium of Greek mathematics, and by methods belonging to projective geometry,

using homogeneous coordinates. Mathematics certainly does not have to be new to be relevant. This chapter contains much material that has not appeared before in a textbook at any level.

There is a chapter on polynomial splines, where we split the interval on which we wish to approximate a function into subintervals. The approximating function consists of a sequence of polynomials, one on each subinterval, that connect together *smoothly*. The simplest and least smooth example of this is a polygonal arc. Although we can detect some ideas in earlier times that remind us of splines, this is a topic that truly belongs to the twentieth century. It is a good example of exciting, relatively new mathematics that worthily stands alongside the best mathematics of any age. Bernstein polynomials, the subject of the penultimate chapter, date from the early twentieth century. Their creation was inspired by the famous theorem stated by Weierstrass towards the end of the nineteenth century that a continuous function on a finite interval of the real line can be approximated by a polynomial with any given precision over the whole interval. Polynomials are simple mathematical objects that are easy to evaluate, differentiate, and integrate, and Weierstrass's theorem justifies their importance in approximation theory.

Several of the processes discussed in this book have special cases where a function is evaluated at equal intervals, and we can scale the variable so that the function is evaluated at the integers. For example, in finite difference methods for interpolation the interpolated function is evaluated at equal intervals, and the same is true of the integrand in the Newton–Cotes integration rules. Equal intervals occur also in the Bernstein polynomials and the uniform B-splines. In our study of these four topics, we also discuss processes in which the function is evaluated at intervals whose lengths are in geometric progression. These can be scaled so that the function is evaluated on the q-integers. Over twenty years ago I was asked to referee a paper by the distinguished mathematician I. J. Schoenberg (1903–1990), who is best known for his pioneering work on splines. Subsequently I had a letter from the editor of the journal saying that Professor Schoenberg wished to know the name of the anonymous referee. Over the following few years I had a correspondence with Professor Schoenberg which I still value very much. His wonderful enthusiasm for mathematics continued into his eighties. Indeed, of his 174 published papers and books, 56 appeared after his retirement in 1973. The above-mentioned paper by Iso Schoenberg was the chief influence on the work done by S. L. Lee and me in applying q-integers to interpolation on triangular regions. The q-integer motif was continued in joint work with Zeynep Koçak on splines and then in my work on the Bernstein polynomials, in which I was joined by Tim Goodman and Halil Oruç. The latter work nicely illustrates variation-diminishing ideas, which take us into the relatively new area of CAGD.

The inclusion of a few rather minor topics in whose development I have been directly involved may cause some eyebrows to be raised. But I trust

I will be forgiven, since my intentions are honourable: I hope that by including a relatively small amount of material on topics in which I have carried out research I can encourage all students of mathematics, and especially those thinking of doing research, to know and understand that the discovery of new ideas in mathematics is not the sole preserve of the most outstanding mathematicians.

I was born in Aberdeen, Scotland. Soon after I began lecturing at the University of St Andrews I learned that the famous Scottish mathematician James Gregory (1638–1675) had also lectured there. Gregory was born in Drumoak, near Aberdeen, and was educated at Aberdeen Grammar School, as I was. Moreover, Gregory's year of birth was exactly three hundred years before mine. There the similarities end, most fortunately for me, in view of the woeful brevity of Gregory's life. James Gregory obtained many important results in the early development of the calculus, including his discovery of the series for the inverse tangent. Indeed, H. W. Turnbull [54], who carried out a most rigorous study of Gregory's publications and unpublished manuscripts and letters, argues that Gregory's mastery of what we call *Maclaurin* series and *Taylor* series, after Colin Maclaurin (1698–1746) and Brook Taylor (1685–1731), entitles him to much more recognition than he has received. Much ahead of his time, Gregory made a bold attempt at showing the transcendental nature of the numbers π and e, results that were completed only at the end of the nineteenth century by C. L .F. Lindemann (1852–1939).

On completing this book it is a pleasure to renew my thanks to my old friend Peter Taylor, from whom I have learned much about numerical analysis. Our meeting as colleagues at the University of Southampton in 1963 was the start of a most valued friendship combined with a fruitful collaboration in mathematics. Our book *Theory and Applications of Numerical Analysis* was first published in 1973 and is still in print as I write this. It has been translated into Chinese and Farsi (Persian), and a second edition was published in 1996. Now, in the fortieth year of our friendship, I am very grateful to Peter Taylor for his most helpful comments on the manuscript of this book. At an earlier stage in the preparation of the manuscript, I particularly remember discussions with Peter concerning the divided difference form for interpolation on a triangular region. This is one of the few significant mathematical conversations I can recall sharing that did not involve writing anything down. We were sitting on a park bench in the Meadows in Edinburgh before attending a concert by the Edinburgh Quartet in the Queen's Hall.

It is also a pleasure to thank my much respected colleague József Szabados, who read the first draft of the manuscript of this book on behalf of my publisher. I am extremely grateful to him for the great care he took in pursuing this task, and for the wisdom of his remarks and suggestions. As a result, I believe I have been able to make some substantial improvements in the text. However, I am solely responsible for the final form of this book.

A few years ago I invited my good friend and former student Halil Oruç to join me in writing a book on approximation theory. We were both very disappointed that he was unable to do this, due to pressure of other work. Different parents produce different children. Indeed, children of the same parents are usually rather different. Therefore, although Halil and I would surely have produced a rather different book together, I hope that he will approve of this one.

This book is my second contribution to the series *CMS Books in Mathematics*, and it is a pleasure to thank the editors, Jonathan and Peter Borwein, for their support and encouragement. I also wish to acknowledge the fine work of those members of the staff of Springer, New York who have been involved with the production of this book. I am especially grateful to the copyeditor, David Kramer. I worked through his suggestions, page by page, with an ever increasing respect for the great care he devoted to his task. I also wish to thank my friend David Griffiths for his help with one of the items in the Bibliography. I must also mention David's book *Learning LaTeX*, written jointly with Desmond J. Higham and published by SIAM. It has been my guide as I prepared this text in LaTeX.

In *Two Millennia of Mathematics*, my first contribution to the CMS series, I expressed my thanks to my early teachers and lecturers and to the many mathematicians, from many countries, who have influenced me and helped me. I will not repeat that lengthy list here, having put it on record so recently. However, having mentioned I. J. Schoenberg, let me write down also, in the order in which I met them, the names of three other approximation theorists, Philip Davis, Ward Cheney, and Ted Rivlin. Their most elegantly written and authoritative books on approximation theory inspired me and taught me a great deal. I would also like to mention the name of my good friend Lev Brutman (1939–2001), whose work I quote in the section on Lebesgue constants. I was his guest in Israel, and he was mine in Scotland, and we corresponded regularly for several years. I admired very much his fine mathematical achievements and his language skills in Russian, Hebrew, and English. He loved to read poetry in these three languages. Lev's favourite poet was Alexander Pushkin (1799–1837), and he also admired Robert Burns (1759–1796), whose work he began reading in Russian translation during his early years in Moscow.

I dedicated my Ph.D. thesis to my dear parents, Betty McArtney Phillips (1910–1961) and George Phillips (1911–1961). With the same measure of seriousness, love, and gratitude, I dedicate this book to my wife, Rona.

George M. Phillips
Crail, Scotland

This page intentionally left blank

Contents

1
Univariate Interpolation

1.1 Introduction

Given the values of a function $f(x)$ at two distinct values of x, say x_0 and x_1, we could approximate f by a linear function p that satisfies the conditions

$$p(x_0) = f(x_0) \qquad \text{and} \qquad p(x_1) = f(x_1). \tag{1.1}$$

It is geometrically obvious that such a p exists and is unique. (See Figure 1.1.) We call p a linear interpolating polynomial. We may then evaluate $p(x)$ for a value of x other than x_0 or x_1 and use it as an approximation for $f(x)$. This process is called *linear interpolation*.

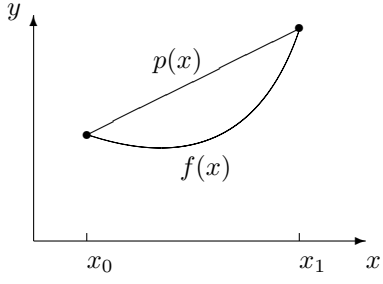

FIGURE 1.1. Linear interpolation. The curve $y = f(x)$ is approximated by the straight line $y = p(x)$.

We can construct the linear interpolating polynomial directly, writing $p(x) = ax + b$ and using the above two conditions in (1.1) to give two linear equations to determine a and b. On solving these equations, we obtain

$$p(x) = \frac{x_1 f(x_0) - x_0 f(x_1)}{x_1 - x_0} + x\left(\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right). \tag{1.2}$$

This can also be expressed in the Lagrange symmetric form

$$p(x) = \left(\frac{x - x_1}{x_0 - x_1}\right) f(x_0) + \left(\frac{x - x_0}{x_1 - x_0}\right) f(x_1), \tag{1.3}$$

or in Newton's divided difference form

$$p(x) = f(x_0) + (x - x_0)\left(\frac{f(x_1) - f(x_0)}{x_1 - x_0}\right), \tag{1.4}$$

to which we will return later. Observe that if we write $x_1 = x_0 + h$ in (1.4), the limit of $p(x)$ as $h \to 0$ gives the first two terms of the Taylor series for f, assuming that f is differentiable.

It is convenient to denote the set of all polynomials of degree at most n by P_n. Given the values of a function $f(x)$ at $n+1$ distinct values of x, say $x_0, x_1, \ldots, x_n$, can we find a polynomial $p_n \in P_n$, say

$$p_n(x) = a_0 + a_1 x + a_2 x^2 + \cdots + a_n x^n,$$

such that $p_n(x_j) = f(x_j)$, for $j = 0, 1, \ldots, n$? This means that we require

$$a_0 + a_1 x_j + a_2 x_j^2 + \cdots + a_n x_j^n = f(x_j), \quad 0 \le j \le n, \tag{1.5}$$

giving a system of $n+1$ linear equations to determine the $n+1$ unknowns $a_0, a_1, \ldots, a_n$. These equations, which may be written in the form

$$\begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ \vdots \\ a_n \end{bmatrix} = \begin{bmatrix} f(x_0) \\ f(x_1) \\ \vdots \\ f(x_n) \end{bmatrix}, \tag{1.6}$$

have a unique solution if the matrix

$$\mathbf{V} = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}, \tag{1.7}$$

called the *Vandermonde* matrix, is nonsingular. It is not hard to verify (see Problem 1.1.1) that the determinant of $\mathbf{V}$ is given by

$$\det \mathbf{V} = \prod_{i>j} (x_i - x_j), \tag{1.8}$$

where the product is taken over all i and j such that $0 \le j < i \le n$. For example, when $n = 2$,

$$\det \mathbf{V} = (x_1 - x_0)(x_2 - x_0)(x_2 - x_1).$$

Since the abscissas $x_0, x_1, \ldots, x_n$ are distinct, it is clear from (1.8) that $\det \mathbf{V}$ is nonzero. Thus the Vandermonde matrix $\mathbf{V}$ is nonsingular, and the system of linear equations (1.6) has a unique solution. We conclude that given a function f defined on a set of distinct points $x_0, x_1, \ldots, x_n$, there is a *unique* polynomial $p_n \in P_n$ such that $p_n(x_j) = f(x_j)$ for $j = 0, 1, \ldots, n$. This is called the *interpolating polynomial*. Note that the degree of p_n may be less than n. For example, if all $n+1$ points $(x_j, f(x_j))$ lie on a straight line, then the interpolating polynomial will be of degree 1 or 0, the latter case occurring when all the $f(x_j)$ are equal.

It is not necessary to solve the above system of equations (1.6), whose matrix is the Vandermonde matrix, because the interpolating polynomial p_n can easily be constructed by other means, as we will now show. However, since the solution of the Vandermonde system is an interesting problem in its own right, we will return to it in Section 1.2.

Instead of using the monomials 1, x, $x^2, \ldots, x^n$ as a *basis* for the polynomials of degree at most n, let us use the *fundamental polynomials* L_0, $L_1, \ldots, L_n$, where

$$L_i(x) = \prod_{j \ne i} \left(\frac{x - x_j}{x_i - x_j} \right), \tag{1.9}$$

and the product is taken over all j between 0 and n, but excluding $j = i$. It follows from this definition that $L_i(x)$ takes the value 1 at $x = x_i$ and is zero at all n other abscissas x_j, with $j \ne i$. Each polynomial $L_i(x)$ is of degree n. For example,

$$L_0(x) = \frac{(x - x_1) \cdots (x - x_n)}{(x_0 - x_1) \cdots (x_0 - x_n)},$$

and we see that $L_0(x_0) = 1$ and $L_0(x_j) = 0$ for $1 \le j \le n$. Thus $f(x_i)L_i(x)$ has the value $f(x_i)$ at $x = x_i$ and is zero at the other abscissas. We can therefore express the interpolating polynomial $p_n(x)$ very simply in terms of the fundamental polynomials $L_i(x)$ as

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x), \tag{1.10}$$

for the polynomial on the right of (1.10) is of degree at most n and takes the appropriate value at each abscissa $x_0, x_1, \ldots, x_n$. We call (1.10) the *Lagrange form* of the interpolating polynomial. It is named after J. L. Lagrange (1736–1813) and generalizes the linear interpolating polynomial given in the form (1.3).

As we have seen, an interpolating polynomial p_n for a given function f is constructed by using the values of f at certain abscissas $x_0, \ldots, x_n$. We can then evaluate p_n at some point x distinct from all the x_j, and use this as an approximation for $f(x)$. This process is called *interpolation*, and later in this section we will comment on the accuracy of this process. Another application involving an interpolating polynomial p_n for a given function f is to integrate p_n over some appropriate interval $[a, b]$, and use this as an approximation to the integral of f over $[a, b]$. We will pursue this application in Chapter 3.

Isaac Newton (1642–1727) found a particularly elegant way of constructing the interpolating polynomial. Instead of using the monomials 1, x, $x^2, \ldots, x^n$ or the fundamental polynomials L_i, defined above, as a basis for the polynomials of degree at most n, he used the polynomials $\pi_0, \pi_1, \ldots, \pi_n$, where

$$\pi_i(x) = \begin{cases} 1, & i = 0, \\ (x - x_0)(x - x_1) \cdots (x - x_{i-1}), & 1 \le i \le n. \end{cases} \tag{1.11}$$

The interpolating polynomial $p_n \in P_n$, which assumes the same values as the function f at x_0, $x_1, \ldots, x_n$, is then written in the form

$$p_n(x) = a_0\pi_0(x) + a_1\pi_1(x) + \cdots + a_n\pi_n(x). \tag{1.12}$$

We may determine the coefficients a_j by setting

$$p_n(x_j) = f(x_j), \quad 0 \le j \le n,$$

giving the system of linear equations

$$a_0\pi_0(x_j) + a_1\pi_1(x_j) + \cdots + a_j\pi_j(x_j) = f(x_j), \tag{1.13}$$

for $0 \le j \le n$. Note that only $a_0, \ldots, a_j$ appear in (1.13), because $\pi_i(x_j) = 0$ when $i > j$. The system of equations (1.13) has the matrix form

$$\mathbf{Ma} = \mathbf{f}, \tag{1.14}$$

say, where $\mathbf{a} = [a_0, \ldots, a_n]^T$, $\mathbf{f} = [f(x_0), \ldots, f(x_n)]^T$, and the matrix $\mathbf{M}$ is

$$\mathbf{M} = \begin{bmatrix} \pi_0(x_0) & 0 & 0 & \cdots & 0 \\ \pi_0(x_1) & \pi_1(x_1) & 0 & \cdots & 0 \\ \pi_0(x_2) & \pi_1(x_2) & \pi_2(x_2) & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \pi_0(x_n) & \pi_1(x_n) & \pi_2(x_n) & \cdots & \pi_n(x_n) \end{bmatrix}. \tag{1.15}$$

The matrix $\mathbf{M}$, which we will call the Newton matrix, is said to be *lower triangular*, and we will have more to say about lower triangular matrices

in the next section. If we evaluate the determinant of $\mathbf{M}$ via the first row, we readily obtain

$$\det \mathbf{M} = \pi_0(x_0)\,\pi_1(x_1)\cdots\pi_n(x_n). \tag{1.16}$$

If the $n+1$ abscissas $x_0, x_1, \ldots, x_n$ are all distinct, it is clear from (1.16) that $\det \mathbf{M} \neq 0$, and so the linear system (1.14) has a unique solution. The fact that the matrix $\mathbf{M}$ is lower triangular is crucial to the success of this approach to the interpolation problem, since we can solve the linear system (1.14) by *forward substitution*. We obtain a_0 immediately from the first equation, and then a_1 from the second. In general, we determine a_j from the $(j+1)$th equation, and we can see that a_j depends only on the values of x_0 up to x_j and $f(x_0)$ up to $f(x_j)$. In particular, we obtain

$$a_0 = f(x_0) \qquad \text{and} \qquad a_1 = \frac{f(x_1) - f(x_0)}{x_1 - x_0}. \tag{1.17}$$

We will write

$$a_j = f[x_0, x_1, \ldots, x_j], \quad 0 \leq j \leq n, \tag{1.18}$$

to emphasize its dependence on f and $x_0, x_1, \ldots, x_j$, and refer to a_j as a jth divided difference. The form of the expression for a_1 in (1.17) above and the recurrence relation (1.22) below show why the term *divided difference* is appropriate. Thus we may write (1.12) in the form

$$p_n(x) = f[x_0]\pi_0(x) + f[x_0, x_1]\pi_1(x) + \cdots + f[x_0, x_1, \ldots, x_n]\pi_n(x), \tag{1.19}$$

which is Newton's *divided difference formula* for the interpolating polynomial. Observe that $f[x_0] = f(x_0)$. We write $f[x_0]$ in (1.19) rather than $f(x_0)$ for the sake of harmony of notation. The formula (1.4), which we gave earlier for the linear interpolating polynomial, is the special case of (1.19) with $n = 1$. Note that since we can interpolate on any set of distinct abscissas, we can define a divided difference with respect to any set of distinct abscissas. Later in this chapter we will find it more appropriate to use another notation for divided differences, where we write

$$[x_0, x_1, \ldots, x_j]f \tag{1.20}$$

instead of $f[x_0, x_1, \ldots, x_j]$. In (1.20) we regard $[x_0, x_1, \ldots, x_j]$ as an *operator* that acts on the function f. We now show that the divided difference $f[x_0, x_1, \ldots, x_n]$ is a symmetric function of its arguments x_0, $x_1, \ldots, x_n$.

Theorem 1.1.1 The divided difference $f[x_0, x_1, \ldots, x_n]$ can be expressed as the following symmetric sum of multiples of $f(x_j)$,

$$f[x_0, x_1, \ldots, x_n] = \sum_{r=0}^{n} \frac{f(x_r)}{\prod_{j \neq r}(x_r - x_j)}, \tag{1.21}$$

where in the above product of n factors, r remains fixed and j takes all values from 0 to n, excluding r.

x_0	$f[x_0]$			
		$f[x_0, x_1]$		
x_1	$f[x_1]$		$f[x_0, x_1, x_2]$	
		$f[x_1, x_2]$		$f[x_0, x_1, x_2, x_3]$
x_2	$f[x_2]$		$f[x_1, x_2, x_3]$	
		$f[x_2, x_3]$		$f[x_1, x_2, x_3, x_4]$
x_3	$f[x_3]$		$f[x_2, x_3, x_4]$	
		$f[x_3, x_4]$		
x_4	$f[x_4]$			

TABLE 1.1. A systematic scheme for calculating divided differences.

Proof. Since the interpolating polynomial is unique, the polynomials $p_n(x)$ in (1.10) and (1.19) are the same. If we equate the coefficients of x^n in (1.10) and (1.19), we obtain (1.21) immediately. ■

It is clear from the symmetric form (1.21) that each divided difference $f[x_0, x_1, \ldots, x_n]$ is indeed a symmetric function of its arguments, meaning that it is unchanged if we rearrange the x_j in any order. For example, we have

$$f[x_0, x_1, x_2] = \frac{f(x_0)}{(x_0 - x_1)(x_0 - x_2)} + \frac{f(x_1)}{(x_1 - x_0)(x_1 - x_2)} + \frac{f(x_2)}{(x_2 - x_0)(x_2 - x_1)},$$

and we can see that $f[x_0, x_1, x_2]$ is equal to $f[x_1, x_2, x_0]$, and to each of the four other expressions obtained by permuting the x_j.

We can use the symmetric form (1.21) to show that

$$f[x_0, x_1, \ldots, x_n] = \frac{f[x_1, x_2, \ldots, x_n] - f[x_0, x_1, \ldots, x_{n-1}]}{x_n - x_0}. \tag{1.22}$$

For we can replace both divided differences on the right of (1.22) by their respective symmetric forms and collect the terms in $f(x_0), f(x_1)$, and so on, showing that this gives the symmetric form for the divided difference $f[x_0, x_1, \ldots, x_n]$. By repeatedly applying the relation (1.22) systematically, we can build up a table of divided differences, as depicted in Table 1.1.

Example 1.1.1 Given the table of values

x	0	1	2	3
$f(x)$	6	−3	−6	9

let us derive the interpolating polynomial for f of degree at most 3 by using Newton's divided difference form (1.19). We first construct the following divided difference table, like the model given in Table 1.1, but with one fewer entry in each column.

$$\begin{array}{cc|ccc} 0 & 6 & & & \\ & & -9 & & \\ 1 & -3 & & 3 & \\ & & -3 & & 2 \\ 2 & -6 & & 9 & \\ & & 15 & & \\ 3 & 9 & & & \end{array}$$

The values of the x_j and $f(x_j)$ are given in the first two columns, as in Table 1.1. To evaluate the divided difference formula (1.19), we require only the first numbers in columns 2, 3, 4, and 5, together with the first three numbers in the first column, giving the interpolating polynomial

$$p_3(x) = 6 - 9x + 3x(x-1) + 2x(x-1)(x-2). \tag{1.23}$$

Let us rearrange the order of the four pairs $(x_j, f(x_j))$ in the above table and recompute the divided difference table. For example, the table

$$\begin{array}{cc|ccc} 3 & 9 & & & \\ & & 1 & & \\ 0 & 6 & & 7 & \\ & & -6 & & 2 \\ 2 & -6 & & 3 & \\ & & -3 & & \\ 1 & -3 & & & \end{array}$$

yields the interpolating polynomial

$$p_3(x) = 9 + (x-3) + 7(x-3)x + 2(x-3)x(x-2), \tag{1.24}$$

and we can easily check that the polynomials in (1.23) and (1.24) are, as we expect, the same. Both may be expressed in the standard form

$$p_3(x) = 2x^3 - 3x^2 - 8x + 6. \qquad \blacksquare$$

The function f in Example 1.1.1 above is defined at only four points, and we have no incentive to evaluate its interpolating polynomial at any other point. In the following example we begin with a function that is defined on an interval (in fact, on the whole real line), construct an interpolating polynomial based on the values of the function at five points, and estimate the value of the function at some other point of our choice by evaluating the interpolating polynomial in place of the function.

Example 1.1.2 Let us construct the interpolating polynomial $p_4(x)$ for the function 2^x based on the points -2, -1, 0, 1, and 2, and hence estimate $2^{1/2} = \sqrt{2}$ by evaluating $p_4(\frac{1}{2})$.

For $x = -2$, -1, 0, 1, and 2, we have $2^x = \frac{1}{4}$, $\frac{1}{2}$, 1, 2, and 4, respectively. Then, following the method discussed above, we find that the forward difference form of the interpolating polynomial is

$$p_4(x) = \frac{1}{4} + \frac{1}{4}(x+2) + \frac{1}{8}(x+2)(x+1) + \frac{1}{24}(x+2)(x+1)x + \frac{1}{96}(x+2)(x+1)x(x-1).$$

On evaluating $p_4(x)$ at $x = \frac{1}{2}$, we obtain

$$\frac{723}{512} \approx 1.4121.$$

Since $\sqrt{2} \approx 1.4142$, this interpolation method has provided an approximation whose error is in the third digit after the decimal point. ■

In the above example, we were able to determine how close the interpolated value $p_4(\frac{1}{2})$ is to $2^{1/2}$. What can we say, in general, about the accuracy of interpolation? The following theorem gives at least a partial answer to this question.

Theorem 1.1.2 Let x and the abscissas $x_0, x_1, \ldots, x_n$ be contained in an interval $[a, b]$ on which f and its first n derivatives are continuous, and let $f^{(n+1)}$ exist in the open interval (a, b). Then there exists $\xi_x \in (a, b)$, which depends on x, such that

$$f(x) - p_n(x) = (x - x_0)(x - x_1)\cdots(x - x_n)\frac{f^{(n+1)}(\xi_x)}{(n+1)!}. \tag{1.25}$$

Proof. The proof makes repeated use of Rolle's theorem, which simply says that between any two zeros of a differentiable function there must be at least one zero of its derivative. (See any text on analysis, for example, Howie [27].) Consider the function

$$G(x) = f(x) - p_n(x) - \frac{(x - x_0)\cdots(x - x_n)}{(\alpha - x_0)\cdots(\alpha - x_n)} \cdot (f(\alpha) - p_n(\alpha)), \tag{1.26}$$

where α is any point in the interval $[a, b]$ that is distinct from all of the abscissas x_0, $x_1, \ldots, x_n$. We note that G has at least $n + 2$ zeros, at α and all the $n + 1$ interpolating abscissas x_j. We then argue from Rolle's theorem that G' must have at least $n + 1$ zeros. By repeatedly applying Rolle's theorem, we argue that G'' has at least n zeros (if $n \geq 1$), $G^{(3)}$ has at least $n - 1$ zeros (if $n \geq 2$), and finally that $G^{(n+1)}$ has at least one zero, say at $x = \xi_\alpha$. Thus, on differentiating (1.26) $n+1$ times and putting $x = \xi_\alpha$, we obtain

$$0 = f^{(n+1)}(\xi_\alpha) - \frac{(n+1)!(f(\alpha) - p_n(\alpha))}{(\alpha - x_0)\cdots(\alpha - x_n)}.$$

To complete the proof we rearrange the last equation to give an expression for $f(\alpha) - p_n(\alpha)$, and then replace α by x. ■

The above expression for the interpolation error is obviously of limited use, since it requires the evaluation of the $(n+1)$th-order derivative $f^{(n+1)}$ at ξ_x, and in general, we do not even know the value of ξ_x. We note that there can also be an error in evaluating $p_n(x)$ if there are rounding errors in the values of $f(x_j)$. Despite these shortcomings, Theorem 1.1.2 is valuable because, as we will see later, it both provides a useful comparison with Taylor's theorem and is helpful in establishing a connection between divided differences and derivatives.

Example 1.1.3 Let us apply (1.25) to estimate the error in the interpolation carried out in Example 1.1.2. In this case, $f(x) = 2^x$, $x = \frac{1}{2}$, and $n = 4$. We have

$$\frac{d}{dx}\, 2^x = 2^x \log 2 \qquad \text{and} \qquad \frac{d^5}{dx^5}\, 2^x = 2^x (\log 2)^5,$$

where log denotes the natural logarithm, so that $\log 2 \approx 0.693147$. Using (1.25), we find that the error of interpolation in Example 1.1.2 is

$$\frac{5}{2} \cdot \frac{3}{2} \cdot \frac{1}{2} \cdot \frac{-1}{2} \cdot \frac{-3}{2} \cdot \frac{2^\xi (\log 2)^5}{5!},$$

where $\frac{1}{4} < \xi < 4$. On inserting the two extreme values of ξ into the above estimate, we find that the error of interpolation lies between 0.0004 and 0.0076. This is consistent with the *known* error, which is approximately 0.0021. ■

Our next example shows how the error estimate (1.25) can be used to estimate the error of linear interpolation for any function f whose second derivative can be evaluated.

Example 1.1.4 Suppose we have a table of values of $\sin x$, tabulated at intervals of 0.01. What is the maximum error incurred by using linear interpolation between two consecutive entries in this table? From (1.25) the error of linear interpolation between two points x_0 and x_1 is

$$f(x) - p_1(x) = (x - x_0)(x - x_1)\frac{f''(\xi_x)}{2!}. \tag{1.27}$$

For any function f such that $|f''(x)| \le M$ on $[x_0, x_1]$, we can verify (see Problem 1.1.8) that

$$|f(x) - p_1(x)| \le \frac{1}{8} M h^2, \tag{1.28}$$

where $h = x_1 - x_0$. In particular, for $f(x) = \sin x$, we have $f'(x) = \cos x$ and $f''(x) = -\sin x$. Thus we can take $M = 1$ in (1.28), and with $h = 0.01$, the

error in linear interpolation is not greater than $\frac{1}{8} \cdot 10^{-4}$. It would therefore be appropriate for the entries in this table, spaced at intervals of 0.01, to be given to 4 decimal places. This was well understood in the era, now long gone, when such tables were in everyday use, and one finds in published four-figure tables of the function $\sin x$ that the entries are tabulated at intervals of 0.01. ■

We will now derive an alternative error term for the interpolating polynomial that has the merit of being applicable to *all* functions and not merely to those that possess high-order derivatives. This error term involves a divided difference rather than an $(n+1)$th derivative, as in (1.25). We begin by using (1.22) to express the divided difference $f[x, x_0, x_1, \ldots, x_n]$ in terms of $f[x_0, x_1, \ldots, x_n]$ and $f[x, x_0, x_1, \ldots, x_{n-1}]$. On rearranging this, we obtain

$$f[x, x_0, \ldots, x_{n-1}] = f[x_0, \ldots, x_n] + (x - x_n) f[x, x_0, \ldots, x_n]. \tag{1.29}$$

Similarly, we have

$$f[x] = f[x_0] + (x - x_0) f[x, x_0]. \tag{1.30}$$

On the right side of (1.30) we now replace $f[x, x_0]$, using (1.29) with $n = 1$, to give

$$f[x] = f[x_0] + (x - x_0) f[x_0, x_1] + (x - x_0)(x - x_1) f[x, x_0, x_1], \tag{1.31}$$

and we note that (1.31) may be expressed as

$$f(x) = p_1(x) + (x - x_0)(x - x_1) f[x, x_0, x_1],$$

where $p_1(x)$ is the interpolating polynomial for f based on the two abscissas x_0 and x_1. We can continue, replacing $f[x, x_0, x_1]$ in (1.31), using (1.29) with $n = 2$. Continuing in this way, we finally obtain

$$f(x) = p_n(x) + (x - x_0) \cdots (x - x_n) f[x, x_0, x_1, \ldots, x_n]. \tag{1.32}$$

On comparing (1.32) and (1.25), we see that if the conditions of Theorem 1.1.2 hold, then there exists a number ξ_x such that

$$f[x, x_0, x_1, \ldots, x_n] = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}.$$

Since this holds for any x belonging to an interval $[a, b]$ that contains all the abscissas x_j, and within which f satisfies the conditions of Theorem 1.1.2, we can replace n by $n - 1$, put $x = x_n$, and obtain

$$f[x_0, x_1, \ldots, x_n] = \frac{f^{(n)}(\xi)}{n!}, \tag{1.33}$$

where $\xi \in (x_0, x_n)$. Thus an nth-order divided difference, which involves $n+1$ parameters, behaves like a multiple of an nth-order derivative. If we now return to Newton's divided difference formula (1.19) and let every x_j tend to x_0, then in view of (1.33), we obtain the limiting form

$$p_n(x) = f(x_0) + (x - x_0)\frac{f'(x_0)}{1!} + \cdots + (x - x_0)^n \frac{f^{(n)}(x_0)}{n!}. \tag{1.34}$$

This is the Taylor polynomial of degree n, consisting of the first $n+1$ terms of the Taylor series for f, and if we take the limiting form of the error formula (1.25), we obtain

$$f(x) = \sum_{r=0}^{n} (x - x_0)^r \frac{f^{(r)}(x_0)}{r!} + (x - x_0)^{n+1} \frac{f^{(n+1)}(\eta_x)}{(n+1)!}, \tag{1.35}$$

where η_x lies between x and x_0.

In Chapter 2 we will consider another way of measuring how well an interpolating polynomial p_n approximates f by studying a function, called a Lebesgue function, that does not depend on f but is derived from the interpolating abscissas.

There is a very elegant iterative process, called the Neville–Aitken algorithm, that evaluates the interpolating polynomial p_n for a function f on a given set of distinct abscissas $X = \{x_0, x_1, \ldots, x_n\}$. In this process, which is named after E. H. Neville (1889–1961) and A. C. Aitken (1895–1967), the value of $p_n(x)$ is the final number obtained from a sequence of $\frac{1}{2}n(n+1)$ similar calculations. Each of these calculations is like the simple process of linear interpolation, cast in the form

$$p_1(x) = \frac{(x - x_0)f(x_1) - (x - x_1)f(x_0)}{x_1 - x_0}, \tag{1.36}$$

and every such calculation evaluates an interpolating polynomial for f at some subset of the abscissas in X. Consider the following theorem:

Theorem 1.1.3 Let us define $p_0^{[i]} = f(x_i)$ for $0 \le i \le n$, and then, for each k such that $0 \le k \le n-1$, we recursively define

$$p_{k+1}^{[i]}(x) = \frac{(x - x_i)p_k^{[i+1]}(x) - (x - x_{i+k+1})p_k^{[i]}(x)}{x_{i+k+1} - x_i}, \tag{1.37}$$

for $0 \le i \le n-k-1$. Then each $p_k^{[i]}$ is the interpolating polynomial for the function f based on the abscissas $x_i, x_{i+1}, \ldots, x_{i+k}$. In particular, $p_n^{[0]}$ is the interpolating polynomial for the function f based on the abscissas $x_0, x_1, \ldots, x_n$.

Proof. We use induction on k. By definition, each $p_0^{[i]}(x)$ is a polynomial of degree zero with the constant value $f(x_i)$. Let us assume that

$x - x_0$	$p_0^{[0]}(x)$			
		$p_1^{[0]}(x)$		
$x - x_1$	$p_0^{[1]}(x)$		$p_2^{[0]}(x)$	
		$p_1^{[1]}(x)$		$p_3^{[0]}(x)$
$x - x_2$	$p_0^{[2]}(x)$		$p_2^{[1]}(x)$	
		$p_1^{[2]}(x)$		
$x - x_3$	$p_0^{[3]}(x))$			

TABLE 1.2. The quantities computed in the Neville–Aitken algorithm.

for some $k \geq 0$, each $p_k^{[i]}(x)$ is in P_k and interpolates $f(x)$ on the abscissas $x_i, x_{i+1}, \ldots, x_{i+k}$. (This statement holds for $k = 0$ and all i such that $0 \leq i \leq n$.) Then we can verify from (1.37) that if $p_k^{[i]}(x)$ and $p_k^{[i+1]}(x)$ both have the same value C for a given value of x, then $p_{k+1}^{[i]}(x)$ also has the value C. Therefore, since both $p_k^{[i]}(x)$ and $p_k^{[i+1]}(x)$ take the value $f(x_j)$ for $i+1 \leq j \leq i+k$, so also does $p_{k+1}^{[i]}(x)$. If we set $x = x_i$ and $x = x_{i+k+1}$ in (1.37), we can verify that $p_{k+1}^{[i]}(x)$ takes the values $f(x_i)$ and $f(x_{i+k+1})$, respectively. Thus each $p_{k+1}^{[i]}(x)$ interpolates f on the abscissas $x_i, \ldots, x_{i+k+1}$, and it follows from (1.37) that $p_{k+1}^{[i]}(x)$ is in P_{k+1}. This completes the proof by induction. ■

The following algorithm gives a precise formulation of the Neville–Aitken process.

Algorithm 1.1.1 (Neville–Aitken)
input: $x_0, \ldots, x_n$, $f(x_0), \ldots, f(x_n)$, and x
  for $i = 0$ **to** n
    $p_0^{[i]} := f(x_i)$
    $t_i := x - x_i$
  next i
  for $k = 0$ **to** $n - 1$
    for $i = 0$ **to** $n - k - 1$
      $p_{k+1}^{[i]} := \left(t_i p_k^{[i+1]} - t_{i+k+1} p_k^{[i]}\right) / (t_i - t_{i+k+1})$
    next i
  next k
output: $p_n^{[0]} = p_n(x)$ ■

If we carry out the Neville–Aitken algorithm by hand, it is helpful to write down the values of $p_k^{[i]}$ in a triangular array, as illustrated in Table 1.2 for the case $n = 3$. If we are implementing the algorithm on a computer, this triangular array helps us visualize the algorithm. Note that although the algorithm could be implemented algebraically, either by hand or with

the aid of a symbolic mathematics program, such as Maple, it is usually applied arithmetically, and then the algorithm must be applied separately to compute each required value of $p_n(x)$. If we need to evaluate $p_n(x)$ for many values of x (for example, in order to draw its graph), it would be more efficient to use Newton's divided difference formula, since the divided differences need be evaluated only once.

Example 1.1.5 Let us apply the Neville–Aitken algorithm to evaluate $p_3\left(\frac{3}{2}\right)$, for the function that is tabulated in Example 1.1.1. In the first column of the following table, we have the numbers $x - x_i$, with $x = \frac{3}{2}$.

$$\begin{array}{cc|ccc} x - x_i & f(x_i) & & & \\ \frac{3}{2} & 6 & & & \\ & & -\frac{15}{2} & & \\ \frac{1}{2} & -3 & & -\frac{21}{4} & \\ & & -\frac{9}{2} & & -6 \\ -\frac{1}{2} & -6 & & -\frac{27}{4} & \\ & & -\frac{27}{2} & & \\ -\frac{3}{2} & 9 & & & \end{array}$$

The numbers in the above table correspond to those in Table 1.2. We obtain $p_3\left(\frac{3}{2}\right) = -6$, which agrees with the result obtained by evaluating the polynomial $p_3(x)$ defined in (1.23). ■

Let us consider the interpolating polynomial for a function f on the $2n+2$ abscissas $x_0, x_1, \ldots, x_n$ and $x_0 + h, x_1 + h, \ldots, x_n + h$, and let f' exist on an interval containing all those abscissas. This interpolating polynomial is of the form

$$p_{2n+1}(x) = \sum_{i=0}^{n} [f(x_i)\alpha_i(x;h) + f(x_i + h)\beta_i(x;h)],$$

say, where α_i and β_i are polynomials in x that depend on h. The polynomial p_{2n+1} can be rearranged in the form

$$\sum_{i=0}^{n} f(x_i)[\alpha_i(x;h) + \beta_i(x;h)] + h \sum_{i=0}^{n} \left(\frac{f(x_i + h) - f(x_i)}{h} \right) \beta_i(x;h).$$

We then let $h \to 0$, to give

$$p_{2n+1}(x) = \sum_{i=0}^{n} [f(x_i)u_i(x) + f'(x_i)v_i(x)], \tag{1.38}$$

say. We can show that

$$u_i(x) = [1 - 2L_i'(x_i)(x - x_i)](L_i(x))^2, \tag{1.39}$$

$$v_i(x) = (x - x_i)(L_i(x))^2, \tag{1.40}$$

where L_i is the fundamental polynomial on the abscissas $x_0, x_1, \ldots, x_n$, as defined in (1.9). The polynomial p_{2n+1} is called the Hermite interpolating polynomial for f on the $n+1$ abscissas $x_0, x_1, \ldots, x_n$, named after C. Hermite (1822–1901). Since $L_i \in P_n$, it is clear from (1.39) and (1.40) that u_i and v_i belong to P_{2n+1}, and it then follows from (1.38) that $p_{2n+1} \in P_{2n+1}$. The derivation of (1.40) is easily completed by writing

$$v_i(x) = \lim_{h \to 0} h\beta_i(x; h).$$

The derivation of (1.39) from

$$u_i(x) = \lim_{h \to 0} [\alpha_i(x; h) + \beta_i(x; h)]$$

takes a little more work. It is, however, straightforward to verify (1.39) and (1.40) by checking that

$$u_i(x_j) = \delta_{i,j}, \quad u_i'(x_j) = 0, \quad v_i(x_j) = 0, \quad v_i'(x_j) = \delta_{i,j},$$

for all i and j, where $\delta_{i,j}$ is the Kronecker delta function, which has the value 1 for $i = j$ and is zero otherwise.

We can easily derive an error term for Hermite interpolation. Let us begin with the error term (1.25) and choose interpolating abscissas $x_0, x_1, \ldots, x_n$ and $x_0 + h, x_1 + h, \ldots, x_n + h$. Then, if $f^{(2n+2)}$ exists in some open interval (a, b) that contains all the interpolating abscissas, we let $h \to 0$ and obtain the error term

$$f(x) - p_{2n+1}(x) = (x - x_0)^2 (x - x_1)^2 \cdots (x - x_n)^2 \frac{f^{2n+2}(\eta_x)}{(2n+2)!}, \tag{1.41}$$

where $\eta_x \in (a, b)$.

Example 1.1.6 Let us obtain the Hermite interpolating polynomial (1.38) for $\sin \pi x$ with interpolating points 0, $\frac{1}{2}$, and 1. With a little work, we find that (1.38) simplifies to give

$$p_5(x) = (16 - 4\pi)x^2(1 - x)^2 + \pi x(1 - x).$$

From (1.41) the error of this approximation is of the form

$$x^2 \left(x - \frac{1}{2}\right)^2 (x - 1)^2 \frac{f^{(6)}(\eta_x)}{6!},$$

where $f(x) = \sin \pi x$, and we find that the maximum modulus of the polynomial $x^2 \left(x - \frac{1}{2}\right)^2 (x - 1)^2$ on $[0, 1]$ is $\frac{1}{432}$, attained at $x = \frac{1}{2} \pm \frac{\sqrt{3}}{6}$. Thus

$$\max_{0 \le x \le 1} |\sin \pi x - p_5(x)| \le \frac{1}{432} \frac{\pi^6}{6!} < 0.0031. \qquad \blacksquare$$

Problem 1.1.1 Consider the Vandermonde matrix $\mathbf{V}$ in (1.7). One term in the expansion of $\det \mathbf{V}$ is the product of the elements on the main diagonal,

$$x_1 x_2^2 x_3^3 \cdots x_n^n,$$

which has total degree

$$1 + 2 + \cdots + n = \frac{1}{2} n(n+1).$$

Show that $\det \mathbf{V}$ is a polynomial in the variables $x_0, x_1, \ldots, x_n$ of total degree $\frac{1}{2}n(n+1)$. If $x_i = x_j$ for any i and j, observe that $\det \mathbf{V} = 0$ and deduce that $x_i - x_j$ is a factor of $\det \mathbf{V}$. Note that there are $\frac{1}{2}n(n+1)$ factors of this form and deduce that

$$\det \mathbf{V} = C \prod_{i>j} (x_i - x_j),$$

where C is a constant, since the right side of the latter equation is also of total degree $\frac{1}{2}n(n+1)$. Verify that the choice $C = 1$ gives the correct coefficient for the term $x_1 x_2^2 x_3^3 \cdots x_n^n$ on both sides, thus verifying (1.8).

Problem 1.1.2 Show that the fundamental polynomials L_i satisfy the identities

$$L_0(x) + L_1(x) + \cdots + L_n(x) = 1 \tag{1.42}$$

for all $n \geq 0$ and

$$x_0 L_0(x) + x_1 L_1(x) + \cdots + x_n L_n(x) = x$$

for $n \geq 1$. Can you find any other identities of this kind?

Problem 1.1.3 Show that the fundamental polynomial $L_i(x)$ can be expressed in the form

$$L_i(x) = \frac{w(x)}{(x - x_i) w'(x_i)},$$

where $w(x) = (x - x_0)(x - x_1) \cdots (x - x_n)$. By differentiating the above expression for $L_i(x)$ and using L'Hospital's rule, show further that

$$L_i'(x_i) = \frac{1}{2} \frac{w''(x_i)}{w'(x_i)}.$$

Problem 1.1.4 Show that

$$\det \mathbf{V} = \det \mathbf{M},$$

where $\mathbf{V}$ and $\mathbf{M}$ are defined by (1.7) and (1.15), respectively.

Problem 1.1.5 Verify the recurrence relation (1.22) for divided differences.

Problem 1.1.6 Using Newton's divided difference form of the interpolating polynomial for the function $\sin \pi x$ based on the five points 0, $\pm\frac{1}{6}$, and $\pm\frac{1}{2}$, obtain the approximation 91/128 for $\sin(\pi/4)$.

Problem 1.1.7 Apply (1.25) to the function $\sin \pi x$ and thus show that the estimate obtained in Problem 1.1.6 for $\sin(\pi/4)$ is too large, by an amount not greater than $\pi^5/73728 \approx 0.004$. Compare this error estimate with the actual error.

Problem 1.1.8 Verify that the derivative of the function $(x-x_0)(x-x_1)$ has only one zero, at the midpoint of $[x_0, x_1]$. Hence show that

$$\max_{x_0 \le x \le x_1} |(x-x_0)(x-x_1)| = \frac{1}{4}(x_1-x_0)^2$$

and thus derive the inequality (1.28).

Problem 1.1.9 Apply the Neville–Aitken algorithm to evaluate $p_4(x)$ at $x = \frac{1}{2}$ for the function 2^x based on the points -2, -1, 0, 1, and 2, and check that your result agrees with that obtained in Example 1.1.2.

1.2 The Vandermonde Equations

Newton's solution of the interpolating problem by using divided differences makes the direct solution of the Vandermonde equations (1.6) unnecessary. Nevertheless, we will show in this section that the solution of these equations is not nearly as difficult as one might suppose. We include this material for its intrinsic interest, while emphasizing that it is not a recommended practical method for constructing the interpolating polynomial. We begin with definitions concerning certain symmetric polynomials in several variables.

Definition 1.2.1 The *elementary symmetric function* $\sigma_r(x_0, x_1, \ldots, x_n)$, for $r \ge 1$, is the sum of all products of r distinct variables chosen from the set $\{x_0, x_1, \ldots, x_n\}$, and we define $\sigma_0(x_0, x_1, \ldots, x_n) = 1$. ■

For example,

$$\sigma_2(x_0, x_1, x_2) = x_0x_1 + x_0x_2 + x_1x_2.$$

As a consequence of Definition 1.2.1, we have

$$\sigma_r(x_0, x_1, \ldots, x_n) = 0 \quad \text{if} \quad r > n+1. \tag{1.43}$$

Definition 1.2.2 The *complete symmetric function* $\tau_r(x_0, x_1, \ldots, x_n)$ is the sum of all products of the variables $x_0, x_1, \ldots, x_n$ of total degree r, for $r \ge 1$, and we define $\tau_0(x_0, x_1, \ldots, x_n) = 1$. ■

For example,

$$\tau_2(x_0, x_1, x_2) = x_0^2 + x_1^2 + x_2^2 + x_0x_1 + x_0x_2 + x_1x_2,$$
$$\tau_3(x_0, x_1) = x_0^3 + x_0^2x_1 + x_0x_1^2 + x_1^3.$$

Note that $\tau_r(x_0, x_1, \ldots, x_n)$ contains all of the terms that are contained in $\sigma_r(x_0, x_1, \ldots, x_n)$, together with other terms (if $r > 1$ and $n > 0$) in which at least one x_j occurs to a power greater than one. It follows immediately from Definition 1.2.1 that

$$(1 + x_0x) \cdots (1 + x_nx) = \sum_{r=0}^{n+1} \sigma_r(x_0, \ldots, x_n)x^r, \tag{1.44}$$

so that $(1+x_0x) \cdots (1+x_nx)$ is the *generating function* for the elementary symmetric functions. Likewise, it follows from Definition 1.2.2 that

$$\frac{1}{(1 - x_0x) \cdots (1 - x_nx)} = \prod_{j=0}^{n} \sum_{r=0}^{\infty} x_j^r\, x^r = \sum_{r=0}^{\infty} \tau_r(x_0, \ldots, x_n)x^r,$$

so that $(1 - x_0x)^{-1} \cdots (1 - x_nx)^{-1}$ is the generating function for the complete symmetric functions. By equating coefficients of x^r in the identity

$$\frac{1}{(1 - x_0x) \cdots (1 - x_nx)} - \frac{1}{(1 - x_0x) \cdots (1 - x_{n-1}x)}$$
$$= \frac{x_nx}{(1 - x_0x) \cdots (1 - x_nx)},$$

we deduce that

$$\tau_r(x_0, \ldots, x_n) - \tau_r(x_0, \ldots, x_{n-1}) = x_n\, \tau_{r-1}(x_0, \ldots, x_n) \tag{1.45}$$

for $r \geq 1$. Before taking the next step, let us remember that each τ_r, being a *symmetric* function, is unchanged if we permute its arguments x_j. Then, by interchanging x_0 and x_n in the recurrence relation (1.45), we obtain

$$\tau_r(x_0, \ldots, x_n) - \tau_r(x_1, \ldots, x_n) = x_0\, \tau_{r-1}(x_0, \ldots, x_n). \tag{1.46}$$

If we now subtract (1.46) from (1.45) and divide by $x_n - x_0$, we find that $\tau_{r-1}(x_0, \ldots, x_n)$ is expressed in the *divided difference* form

$$\tau_{r-1}(x_0, \ldots, x_n) = \frac{\tau_r(x_1, \ldots, x_n) - \tau_r(x_0, \ldots, x_{n-1})}{x_n - x_0}. \tag{1.47}$$

Theorem 1.2.1 For any positive integer m and nonnegative integer i,

$$\tau_{m-n}(x_i, \ldots, x_{n+i}) = f[x_i, \ldots, x_{n+i}], \tag{1.48}$$

where $f(x) = x^m$ and $0 \leq n \leq m$. It is worth repeating this result in words: The complete symmetric function of $n + 1$ variables of order $m - n$ is an nth-order divided difference of the monomial x^m.

Proof. We prove this by induction on n. Since $\tau_m(x_i) = x_i^m = f[x_i]$, we see that (1.48) holds for the given value of m and $n = 0$. Now let us assume that (1.47) holds for a value of n such that $0 \leq n < m$. Using this assumption and the divided difference relation (1.47), we deduce that

$$\begin{aligned}\tau_{m-n-1}(x_i, \ldots, x_{n+i+1}) &= \frac{\tau_{m-n}(x_{i+1}, \ldots, x_{n+i+1}) - \tau_{m-n}(x_i, \ldots, x_{n+i})}{x_{n+i+1} - x_i} \\ &= \frac{f[x_{i+1}, \ldots, x_{n+i+1}] - f[x_i, \ldots, x_{n+i}]}{x_{n+i+1} - x_0},\end{aligned}$$

and from the recurrence relation for divided differences (1.22) it follows that

$$\tau_{m-n-1}(x_0, \ldots, x_{n+1}) = f[x_i, \ldots, x_{n+i+1}],$$

showing that (1.48) holds for $n + 1$. This completes the proof. ■

Example 1.2.1 To illustrate (1.48) for $m = 4$ and $n = 2$, we compute the appropriate first-order divided differences as follows:

$$\begin{array}{ll|l} x_i & x_i^4 & \\ & & x_i^3 + x_i^2 x_{i+1} + x_i x_{i+1}^2 + x_{i+1}^3 \\ x_{i+1} & x_{i+1}^4 & \\ & & x_{i+1}^3 + x_{i+1}^2 x_{i+2} + x_{i+1} x_{i+2}^2 + x_{i+2}^3 \\ x_{i+2} & x_{i+2}^4 & \end{array}$$

and we can verify that the second divided difference simplifies to give $\tau_2(x_i, x_{i+1}, x_{i+2})$, in agreement with (1.48). ■

We now require some definitions concerning matrices.

Definition 1.2.3 Given a square matrix $\mathbf{A}$ of order n, its *principal submatrix* of order m, for $1 \leq m \leq n$, is the matrix formed from the first m rows and columns of $\mathbf{A}$. ■

Definition 1.2.4 Given a square matrix $\mathbf{A}$ of order n, its *principal minor* of order m, for $1 \leq m \leq n$, is the determinant of the matrix formed from the first m rows and columns of $\mathbf{A}$. ■

Definition 1.2.5 A square matrix $\mathbf{A} = (a_{ij})$ is said to be *lower triangular* if $a_{ij} = 0$ for $i < j$. ■

Definition 1.2.6 A square matrix $\mathbf{A} = (a_{ij})$ is said to be *upper triangular* if $a_{ij} = 0$ for $i > j$. ■

We now state and prove a result concerning the *factorization* of a square matrix as a product of a lower and an upper triangular matrix.

Theorem 1.2.2 If all principal minors of a square matrix $\mathbf{A}$ are nonzero, then $\mathbf{A}$ may be factorized uniquely in the form $\mathbf{A} = \mathbf{LU}$, where $\mathbf{L}$ is lower triangular with units on the main diagonal and $\mathbf{U}$ is upper triangular.

Proof. We will prove this by using induction on n, the order of the matrix $\mathbf{A}$. The result is obviously valid for $n = 1$. Let us assume that it is valid for some $n \geq 1$. Then consider any $(n+1) \times (n+1)$ matrix $\mathbf{A}_{n+1}$ whose principal minors are all nonzero. We can write this in block form as

$$\mathbf{A}_{n+1} = \begin{bmatrix} \mathbf{A}_n & \mathbf{b}_n \\ \mathbf{c}_n^T & a_{n+1,n+1} \end{bmatrix}, \tag{1.49}$$

where $\mathbf{A}_n$ is a matrix consisting of the first n rows and columns of $\mathbf{A}_{n+1}$, $\mathbf{b}_n$ is a column vector consisting of the first n elements in the last column of $\mathbf{A}_{n+1}$, $\mathbf{c}_n^T$ is a row vector consisting of the first n elements in the last row of $\mathbf{A}_{n+1}$, and $a_{n+1,n+1}$ is the element in the last row and column of $\mathbf{A}_{n+1}$. It follows from our assumption about $\mathbf{A}_{n+1}$ that all principal minors of its submatrix $\mathbf{A}_n$ are nonzero. Thus, from our inductive hypothesis, we can express $\mathbf{A}_n = \mathbf{L}_n\mathbf{U}_n$, say, where $\mathbf{L}_n$ is lower triangular with units on the diagonal and $\mathbf{U}_n$ is upper triangular. We observe that since $\mathbf{A}_n$ is nonsingular, so are its factors $\mathbf{L}_n$ and $\mathbf{U}_n$. The expression of $\mathbf{A}_{n+1}$ in block form given in (1.49) suggests how we should now proceed: We should try to express $\mathbf{A}_{n+1}$ as a product, of an appropriate form $\mathbf{L}_{n+1}\mathbf{U}_{n+1}$, where the first n rows and columns of $\mathbf{L}_{n+1}$ and $\mathbf{U}_{n+1}$ are the matrices $\mathbf{L}_n$ and $\mathbf{U}_n$, respectively. Let us therefore write

$$\mathbf{A}_{n+1} = \mathbf{L}_{n+1}\mathbf{U}_{n+1} = \begin{bmatrix} \mathbf{L}_n & \mathbf{0} \\ \mathbf{d}_n^T & 1 \end{bmatrix} \begin{bmatrix} \mathbf{U}_n & \mathbf{e}_n \\ \mathbf{0}^T & u_{n+1,n+1} \end{bmatrix}, \tag{1.50}$$

where the zero column vector $\mathbf{0}$, the row vector $\mathbf{d}_n^T$, the column vector $\mathbf{e}_n$, and the zero row vector $\mathbf{0}^T$ all have n elements, and $u_{n+1,n+1}$ is the element in the last row and column of $\mathbf{U}_{n+1}$. Our next task is to determine values of the vectors $\mathbf{d}_n^T$ and $\mathbf{e}_n$, and the scalar $u_{n+1,n+1}$ for which the above factorization is valid. We proceed by multiplying the above two matrices in block form. (Note that this multiplication process obeys the same rules as ordinary matrix multiplication, as if the blocks were all scalars.) We thus obtain

$$\mathbf{A}_{n+1} = \begin{bmatrix} \mathbf{L}_n\mathbf{U}_n & \mathbf{L}_n\mathbf{e}_n \\ \mathbf{d}_n^T\mathbf{U}_n & \mathbf{d}_n^T\mathbf{e}_n + u_{n+1,n+1} \end{bmatrix}. \tag{1.51}$$

We can now equate corresponding blocks on the right sides of equations (1.49) and (1.51). This yields four equations. The first equation is

$$\mathbf{A}_n = \mathbf{L}_n\mathbf{U}_n, \tag{1.52}$$

which we already know, and the other three equations are

$$\mathbf{L}_n\mathbf{e}_n = \mathbf{b}_n, \tag{1.53}$$

$$\mathbf{d}_n^T \mathbf{U}_n = \mathbf{c}_n^T, \tag{1.54}$$

and

$$\mathbf{d}_n^T \mathbf{e}_n + u_{n+1,n+1} = a_{n+1,n+1}. \tag{1.55}$$

Since $\mathbf{L}_n$ is nonsingular, the solution of the linear system (1.53) determines a unique value for the vector $\mathbf{e}_n$, and since $\mathbf{U}_n$ is nonsingular, the vector $\mathbf{d}_n^T$ is uniquely determined by (1.54). Finally, we use (1.55) to give $u_{n+1,n+1}$, the final element of the upper triangular factor of $\mathbf{A}_{n+1}$. Thus, having determined $\mathbf{e}_n$, $\mathbf{d}_n^T$, and $u_{n+1,n+1}$ uniquely, we have obtained the unique factorization of $\mathbf{A}_{n+1}$, and this completes the proof. ■

The above proof is a *constructive* proof, since it shows how the factorization of a matrix $\mathbf{A}$ can be carried out by factorizing its principal submatrices in turn. Note also how easily the vectors $\mathbf{e}_n$ and $\mathbf{d}_n^T$ may be determined, since they are obtained by solving the linear systems (1.53) and (1.54), which are triangular.

Example 1.2.2 We can complete the matrix factorization

$$\begin{bmatrix} 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \\ 1 & 4 & 16 & 64 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ d_1 & d_2 & d_3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 & e_1 \\ 0 & 1 & 3 & e_2 \\ 0 & 0 & 2 & e_3 \\ 0 & 0 & 0 & u_{4,4} \end{bmatrix},$$

by using the process employed in the proof of Theorem 1.2.2. From (1.53) we obtain

$$[e_1, e_2, e_3] = [1, 7, 12],$$

and then we find from (1.54) that

$$[d_1, d_2, d_3] = [1, 3, 3].$$

Finally, we derive from (1.55) that $u_{4,4} = 6$. ■

A more efficient way of factorizing $\mathbf{A} = \mathbf{LU}$ is to find the elements of $\mathbf{L}$ and $\mathbf{U}$ in the following order: We begin by finding the first row of $\mathbf{U}$ and then the first column of $\mathbf{L}$. We continue in this way, finding next the second row of $\mathbf{U}$ and then the second column of $\mathbf{L}$, and so on. Note that before we commence the factorization of an $n \times n$ matrix, given that $\mathbf{L}$ and $\mathbf{U}$ are triangular matrices and $\mathbf{L}$ has units on the diagonal, only n^2 elements out of the $2n^2$ elements of $\mathbf{L}$ and $\mathbf{U}$ remain to be determined.

Example 1.2.3 Let us factorize the following matrix. We have put "bullets" in place of the 4^2 unknown elements of $\mathbf{L}$ and $\mathbf{U}$.

$$\mathbf{A} = \begin{bmatrix} 2 & 1 & -1 & 2 \\ -4 & -3 & 2 & -3 \\ 2 & 3 & 2 & -1 \\ -2 & -1 & 4 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ \bullet & 1 & 0 & 0 \\ \bullet & \bullet & 1 & 0 \\ \bullet & \bullet & \bullet & 1 \end{bmatrix} \begin{bmatrix} \bullet & \bullet & \bullet & \bullet \\ 0 & \bullet & \bullet & \bullet \\ 0 & 0 & \bullet & \bullet \\ 0 & 0 & 0 & \bullet \end{bmatrix}.$$

We find that the first row of $\mathbf{U}$ is the same as the first row of $\mathbf{A}$. Then, to give the correct values for the elements in the first column of $\mathbf{A}$, we find that the first column of $\mathbf{L}$ is the vector $[1, -2, 1, -1]^T$. Next, to complete the second row of $\mathbf{A}$, we find that the second row of $\mathbf{U}$ is the vector $[0, -1, 0, 1]$, and to complete the second column of $\mathbf{A}$ we find that the second column of $\mathbf{L}$ is the vector $[0, 1, -2, 0]^T$. We continue the factorization by completing the third row of $\mathbf{U}$, the third column of $\mathbf{L}$, and then obtain the final element of the matrix $\mathbf{U}$. The complete factorization is

$$\begin{bmatrix} 2 & 1 & -1 & 2 \\ -4 & -3 & 2 & -3 \\ 2 & 3 & 2 & -1 \\ -2 & -1 & 4 & -2 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ -1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 2 & 1 & -1 & 2 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix},$$

as may be easily verified. ■

Suppose we wish to solve a system of equations of the form $\mathbf{Ax} = \mathbf{b}$, where $\mathbf{A}$ is a square matrix that has been factorized to give $\mathbf{A} = \mathbf{LU}$, as described above. Then

$$\mathbf{Ax} = \mathbf{b} \;\Leftrightarrow\; \mathbf{L}(\mathbf{Ux}) = \mathbf{b}.$$

If we now write $\mathbf{Ux} = \mathbf{y}$, then the solution of the original system of equations $\mathbf{Ax} = \mathbf{b}$ is equivalent to solving the two systems

$$\mathbf{Ly} = \mathbf{b} \qquad \text{and} \qquad \mathbf{Ux} = \mathbf{y}. \tag{1.56}$$

Although, having factorized $\mathbf{A}$, we now have two systems to solve instead of one, the total number of calculations required is greatly reduced, since the matrices $\mathbf{L}$ and $\mathbf{U}$ are triangular. First we find the vector $\mathbf{y}$ in (1.56) by solving $\mathbf{Ly} = \mathbf{b}$; we find the first element of the intermediate vector $\mathbf{y}$ immediately from the first equation, substitute it into the second equation to find the second element of $\mathbf{y}$, and so on. This process of finding the elements of $\mathbf{y}$ one at a time, beginning with the first, is called *forward substitution*. Having found the vector $\mathbf{y}$, we then turn to the solution of the second triangular system of equations in (1.56). This time, in solving $\mathbf{Ux} = \mathbf{y}$ we find the *last* element of the vector $\mathbf{x}$ immediately from the last equation, substitute it into the second-to-last equation to find the second-to-last element of $\mathbf{x}$, and so on. This process of finding the elements of $\mathbf{x}$ one at a time, beginning with the last, is called *back substitution*. Note that the forward and back substitution processes, giving such a simple means of solving the above linear systems, are possible because the matrices are triangular.

Example 1.2.4 Let us solve the system of linear equations $\mathbf{Ax} = \mathbf{b}$, given by

$$\begin{bmatrix} 2 & 1 & -1 & 2 \\ -4 & -3 & 2 & -3 \\ 2 & 3 & 2 & -1 \\ -2 & -1 & 4 & -2 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 9 \\ -14 \\ -2 \\ -9 \end{bmatrix}. \tag{1.57}$$

We have already factorized the matrix $\mathbf{A} = \mathbf{LU}$ in Example 1.2.3. We now find the vector $\mathbf{y}$ by using forward substitution in the equations $\mathbf{Ly} = \mathbf{b}$, which are

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ -2 & 1 & 0 & 0 \\ 1 & -2 & 1 & 0 \\ -1 & 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 9 \\ -14 \\ -2 \\ -9 \end{bmatrix}.$$

We obtain $y_1 = 9$, and then evaluate y_2, y_3, and y_4 in turn, to obtain the vector $\mathbf{y} = [9, 4, -3, 3]^T$. Finally, we obtain the vector $\mathbf{x}$ by using back substitution in the equations $\mathbf{Ux} = \mathbf{y}$, which are

$$\begin{bmatrix} 2 & 1 & -1 & 2 \\ 0 & -1 & 0 & 1 \\ 0 & 0 & 3 & -1 \\ 0 & 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \\ x_3 \\ x_4 \end{bmatrix} = \begin{bmatrix} 9 \\ 4 \\ -3 \\ 3 \end{bmatrix}.$$

We find that $x_4 = 3$ and then evaluate x_3, x_2, and x_1, in turn, giving the vector $\mathbf{x} = [2, -1, 0, 3]^T$ as the solution of (1.57). ■

We now turn to the factorization of the Vandermonde matrix $\mathbf{V}$, defined by (1.7). Let us assume that the abscissas $x_0, x_1, \ldots, x_n$ are all distinct. Then, as we have already seen from (1.8), this implies that $\mathbf{V}$ is nonsingular. Since every principal submatrix of order $m > 1$ of a Vandermonde matrix is itself a Vandermonde matrix of order m, we see that all principal minors of $\mathbf{V}$ are nonzero. Hence, by Theorem 1.2.2, a Vandermonde matrix can be factorized uniquely as a product of a lower triangular matrix $\mathbf{L}$ with units on the diagonal and an upper triangular matrix $\mathbf{U}$.

We can obtain the factors of the Vandermonde matrices defined by (1.7) for $n = 1, 2, 3, \ldots$ in turn, using the construction employed in the proof of Theorem 1.2.2 and applied in Example 1.2.2. For $n = 1$, we obtain

$$\begin{bmatrix} 1 & x_0 \\ 1 & x_1 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & x_0 \\ 0 & x_1 - x_0 \end{bmatrix},$$

and for $n = 2$ the factors $\mathbf{L}$ and $\mathbf{U}$ are

$$\begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & \dfrac{x_2 - x_0}{x_1 - x_0} & 1 \end{bmatrix} \quad \text{and} \quad \begin{bmatrix} 1 & x_0 & x_0^2 \\ 0 & x_1 - x_0 & (x_1 - x_0)(x_0 + x_1) \\ 0 & 0 & (x_2 - x_1)(x_2 - x_0) \end{bmatrix}.$$

For $n = 3$, the matrix $\mathbf{V}$ has the lower triangular factor

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & \dfrac{x_2 - x_0}{x_1 - x_0} & 1 & 0 \\ 1 & \dfrac{x_3 - x_0}{x_1 - x_0} & \dfrac{(x_3 - x_1)(x_3 - x_0)}{(x_2 - x_1)(x_2 - x_0)} & 1 \end{bmatrix}$$

and the upper triangular factor

$$\begin{bmatrix} 1 & x_0 & x_0^2 & x_0^3 \\ 0 & x_1 - x_0 & (x_1 - x_0)(x_0 + x_1) & (x_1 - x_0)(x_0^2 + x_0 x_1 + x_1^2) \\ 0 & 0 & (x_2 - x_1)(x_2 - x_0) & (x_2 - x_1)(x_2 - x_0)(x_0 + x_1 + x_2) \\ 0 & 0 & 0 & (x_3 - x_2)(x_3 - x_1)(x_3 - x_0) \end{bmatrix}.$$

Let us now consider the $(n + 1) \times (n + 1)$ matrix $\mathbf{V}$ and its factors $\mathbf{L}$ and $\mathbf{U}$ for a general value of n. It is convenient to number the rows and columns of these $(n + 1) \times (n + 1)$ matrices from 0 to n instead of, more usually, from 1 to $n + 1$. From the above evidence for $n = 1, 2$, and 3, it is not hard to conjecture that for a general value of n, the nonzero elements of $\mathbf{L}$ are given by

$$l_{i,j} = \prod_{t=0}^{j-1} \frac{x_i - x_{j-t-1}}{x_j - x_{j-t-1}}, \qquad 0 \le j \le i \le n, \tag{1.58}$$

where an empty product (which occurs when $j = 0$) denotes 1. It is a little harder to spot the pattern in the elements of $\mathbf{U}$. We note that there appears to be a common factor in the elements of each row: For example, for $n = 3$ the elements in the second row of $\mathbf{U}$ have the common factor $x_1 - x_0$. In this case, the quantities that remain after removing the common factor are

$$0, \quad 1, \quad x_0 + x_1, \quad x_0^2 + x_0 x_1 + x_1^2,$$

which are complete symmetric functions. For the matrix $\mathbf{U}$ with $n = 3$, we see that for $0 \le i \le 3$, the elements in the ith row have the common factor $\pi_i(x_i)$, where π_i is defined by (1.11). We therefore conjecture that the nonzero elements of the general matrix $\mathbf{U}$ are given by

$$u_{i,j} = \tau_{j-i}(x_0, \ldots, x_i)\pi_i(x_i), \qquad 0 \le i \le j \le n, \tag{1.59}$$

where again an empty product (which occurs when $i = 0$) has the value 1.

Theorem 1.2.3 The $(n + 1) \times (n + 1)$ matrix $\mathbf{V}$ can be factorized as the product of the triangular matrices $\mathbf{L}$ and $\mathbf{U}$ whose elements are given by (1.58) and (1.59), respectively.

Proof. Let $\mathbf{L}$ and $\mathbf{U}$ be defined by (1.58) and (1.59), respectively. Then the (i,j)th element of $\mathbf{LU}$ is

$$\sum_{k=0}^{n} l_{i,k}u_{k,j}.$$

Note that since $\mathbf{U}$ is upper triangular, $u_{k,j} = 0$ for $k > j$, and thus we can replace n by j as the upper limit in the above sum of products $l_{i,k}u_{k,j}$. Then, using (1.58) and (1.59), we obtain

$$\sum_{k=0}^{j} l_{i,k}u_{k,j} = \sum_{k=0}^{j} \prod_{t=0}^{k-1} \frac{x_i - x_{k-t-1}}{x_k - x_{k-t-1}} \cdot \tau_{j-k}(x_0, \ldots, x_k) \prod_{t=0}^{k-1} (x_k - x_t).$$

This gives

$$\sum_{k=0}^{j} l_{i,k}u_{k,j} = \sum_{k=0}^{j} \tau_{j-k}(x_0, \ldots, x_k) \prod_{t=0}^{k-1} (x_i - x_{k-t-1}), \tag{1.60}$$

which, in view of (1.48), yields

$$\sum_{k=0}^{j} l_{i,k}u_{k,j} = \sum_{k=0}^{j} f[x_0, \ldots, x_k] \prod_{t=0}^{k-1} (x_i - x_{k-t-1}),$$

where $f(x) = x^j$. But the latter expression is just Newton's divided difference form (1.19) of the interpolating polynomial for $f(x) = x^j$, evaluated at $x = x_i$. Since the interpolating polynomial for $f(x) = x^j$ is simply x^j, it follows that

$$\sum_{k=0}^{j} f[x_0, \ldots, x_k] \prod_{t=0}^{k-1} (x_i - x_{k-t-1}) = x_i^j,$$

completing the proof that

$$\mathbf{LU} = \mathbf{V}. \qquad \blacksquare$$

A different factorization of $\mathbf{V}$ as a product of a lower and upper triangular matrix was obtained by Gohberg and Koltracht [20]. The factorization described above is due to Oruç [39]. (See also Oruç and Phillips [41].) We can scale the elements of $\mathbf{L}$ and $\mathbf{U}$, defined by (1.58) and (1.59), to give lower and upper triangular matrices $\mathbf{L}^*$ and $\mathbf{U}^*$ whose nonzero elements are

$$l_{i,j}^* = \prod_{t=0}^{j-1} (x_i - x_{j-i-1}), \qquad 0 \le j \le i \le n, \tag{1.61}$$

and

$$u_{i,j}^* = \tau_{j-i}(x_0, \ldots, x_i), \qquad 0 \le i \le j \le n. \tag{1.62}$$

It is clear from (1.58), (1.59), (1.61), and (1.62) that

$$l_{i,k}u_{k,j} = l^*_{i,k}u^*_{k,j},$$

and the argument following (1.60) shows that $\mathbf{V} = \mathbf{L}^*\mathbf{U}^*$. Note that in the factorization $\mathbf{V} = \mathbf{LU}$ the lower triangular matrix $\mathbf{L}$ has units on the diagonal, whereas in the factorization $\mathbf{V} = \mathbf{L}^*\mathbf{U}^*$ the upper triangular matrix $\mathbf{U}^*$ has units on the diagonal. For further references on the factorization of the Vandermonde matrix, see Higham [25].

Example 1.2.5 We consider again the table of values

x	0	1	2	3
$f(x)$	6	−3	−6	9

for which we found Newton's divided difference form of the interpolating polynomial in Example 1.1.1. Let us now evaluate the interpolating polynomial directly from the solution of the appropriate Vandermonde system (1.6), using the $\mathbf{LU}$ factorization. Thus we will first find an intermediate vector $\mathbf{y}$ by solving the lower triangular system, as in (1.56), and use this as the right side of an upper triangular system to obtain the vector $[a_0, a_1, a_2, a_3]^T$, the solution of the Vandermonde system (1.6), whose elements are the coefficients of the interpolating polynomial.

On substituting $x_i = i$ for $0 \le i \le 3$ in the 4×4 triangular factors $\mathbf{L}$ and $\mathbf{U}$ defined by (1.58) and (1.59), we obtain the factorization of the Vandermonde matrix $\mathbf{V}$,

$$\mathbf{V} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 1 & 1 \\ 1 & 2 & 4 & 8 \\ 1 & 3 & 9 & 27 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 6 \end{bmatrix} = \mathbf{LU}.$$

Then, as in (1.56), we next solve the lower triangular equations

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 1 & 1 & 0 & 0 \\ 1 & 2 & 1 & 0 \\ 1 & 3 & 3 & 1 \end{bmatrix} \begin{bmatrix} y_1 \\ y_2 \\ y_3 \\ y_4 \end{bmatrix} = \begin{bmatrix} 6 \\ -3 \\ -6 \\ 9 \end{bmatrix}$$

by forward substitution to obtain $\mathbf{y} = [6, -9, 6, 12]^T$. This intermediate vector $\mathbf{y}$ becomes the right side in the upper triangular system

$$\begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 1 & 1 \\ 0 & 0 & 2 & 6 \\ 0 & 0 & 0 & 6 \end{bmatrix} \begin{bmatrix} a_0 \\ a_1 \\ a_2 \\ a_3 \end{bmatrix} = \begin{bmatrix} 6 \\ -9 \\ 6 \\ 12 \end{bmatrix},$$

which we solve by back substitution to give $a_3 = 2$, $a_2 = -3$, $a_1 = -8$, and $a_0 = 6$. Thus the required interpolating polynomial is

$$p_3(x) = 2x^3 - 3x^2 - 8x + 6,$$

which agrees with the result obtained by using Newton's divided difference formula in Example 1.1.1. ■

We conclude this section by obtaining a matrix that transforms the Vandermonde matrix into the Newton matrix. If we replace x by $-1/x$ and n by $m-1$ in (1.44), and then multiply throughout by x^m, we obtain

$$\pi_m(x) = \sum_{r=0}^{m} (-1)^r \sigma_r(x_0, \ldots, x_{m-1}) x^{m-r},$$

where $\pi_m(x)$ is defined by (1.11). On reversing the order of the summation, we have

$$\pi_m(x) = \sum_{r=0}^{m} (-1)^{m-r} \sigma_{m-r}(x_0, \ldots, x_{m-1}) x^r. \tag{1.63}$$

Thus we may write

$$\begin{bmatrix} \pi_0(x) \\ \pi_1(x) \\ \vdots \\ \pi_n(x) \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1 \\ x \\ \vdots \\ x^n \end{bmatrix}, \tag{1.64}$$

where $\mathbf{A}$ is the lower triangular matrix whose elements are given by

$$a_{i,j} = \begin{cases} (-1)^{i-j} \sigma_{i-j}(x_0, \ldots, x_{i-1}), & i \geq j, \\ 0, & i < j, \end{cases}$$

for $0 \leq i, j \leq n$. As we have already remarked, the monomials $1, x, \ldots, x^n$ are a basis for the set of all polynomials of degree at most n, and the set of polynomials $\pi_0, \ldots, \pi_n$ is another basis. The matrix $\mathbf{A}$ is a *transformation* matrix, which transforms the first of these bases into the second one. If we now substitute $x = x_i$ in (1.64), we see that $\mathbf{A}$ transforms a *row* of the Vandermonde matrix $\mathbf{V}$, defined by (1.7), into the corresponding row of the Newton matrix $\mathbf{M}$, defined by (1.15). Thus we have $\mathbf{M}^T = \mathbf{A}\mathbf{V}^T$, so that

$$\mathbf{M} = \mathbf{V}\mathbf{A}. \tag{1.65}$$

Problem 1.2.1 Deduce from (1.44) the recurrence relation

$$\sigma_r(x_0, \ldots, x_n) = \sigma_r(x_0, \ldots, x_{n-1}) + x_n \, \sigma_{r-1}(x_0, \ldots, x_{n-1}),$$

where $r \geq 1$ and $n \geq 1$.

Problem 1.2.2 Show that $\sigma_r(x_0, \ldots, x_n)$ is a sum of $\binom{n+1}{r}$ terms.

Problem 1.2.3 Show by induction on r, using the recurrence relation (1.45), that $\tau_r(x_0, \ldots, x_n)$ is a sum of $\binom{n+r}{r}$ terms.

Problem 1.2.4 Verify directly from a divided difference table that

$$\tau_3(x_0, x_1, x_2) = f[x_0, x_1, x_2],$$

where $f(x) = x^5$.

Problem 1.2.5 Deduce directly from Definition 1.2.2 that

$$\tau_1(1, 2, \ldots, n) = \frac{1}{2}n(n+1).$$

Problem 1.2.6 Using the recurrence relation (1.45) and the result of Problem 1.2.5, show by induction on n that

$$\tau_2(1, 2, \ldots, n) = \frac{1}{24}n(n+1)(n+2)(3n+1).$$

Problem 1.2.7 Use the recurrence relation (1.45) and the result of Problem 1.2.6 to show by induction on n that

$$\tau_3(1, 2, \ldots, n) = \frac{1}{48}n^2(n+1)^2(n+2)(n+3).$$

Problem 1.2.8 Deduce from (1.45) that

$$\tau_r(1, \ldots, n) = \sum_{s=1}^{n} s\,\tau_{r-1}(1, \ldots, s),$$

and so show by induction on r that $\tau_r(1, \ldots, n)$ is a polynomial in n of degree $2r$.

Problem 1.2.9 With the choice of abscissas $x_j = j$ for $0 \le j \le n$, verify that the nonzero elements of the $(n+1) \times (n+1)$ triangular factors $\mathbf{L}$ and $\mathbf{U}$ (see (1.58) and (1.59)) of the Vandermonde matrix $\mathbf{V}$ satisfy

$$l_{i,j} = \begin{pmatrix} i \\ j \end{pmatrix}, \quad 0 \le j \le i \le n,$$

and

$$u_{i,j} = i!\,\tau_{j-i}(0, 1, \ldots, n), \quad 0 \le i \le j \le n.$$

Problem 1.2.10 Given the table

x	1	2	3	4
$f(x)$	-16	-13	-4	17

derive the interpolating polynomial $p_3 \in P_3$ for f by factorizing the Vandermonde matrix, and hence solve the Vandermonde equations.

Problem 1.2.11 Recalling that

$$\det(\mathbf{BC}) = \det \mathbf{B} \det \mathbf{C} \qquad \text{and} \qquad \det \mathbf{B}^T = \det \mathbf{B},$$

where $\mathbf{B}$ and $\mathbf{C}$ are any square matrices of the same order and $\mathbf{B}^T$ denotes the transpose of the matrix $\mathbf{B}$, deduce from (1.65) that $\det \mathbf{M} = \det \mathbf{V}$, where $\mathbf{M}$ is the Newton matrix defined by (1.15), and $\mathbf{V}$ is the Vandermonde matrix defined by (1.7).

Problem 1.2.12 Let $\mathbf{A} = \mathbf{LU}$, where $\mathbf{A}$ is a nonsingular square matrix, $\mathbf{L}$ is upper triangular, and $\mathbf{U}$ is lower triangular. Express $\mathbf{A}^T$ as the product of a lower and an upper triangular matrix.

1.3 Forward Differences

When we compute divided differences, as in Table 1.1, we repeatedly calculate quotients of the form

$$\frac{f[x_{j+1}, \ldots, x_{j+k+1}] - f[x_j, \ldots, x_{j+k}]}{x_{j+k+1} - x_j}, \tag{1.66}$$

where k has the same value throughout any one column of the divided difference table. We note that $k = 0$ for first-order divided differences, in column 3 of Table 1.1, $k = 1$ for the second-order divided differences in the next column, and so on. Now let the abscissas x_j be equally spaced, so that $x_j = x_0 + jh$, where $h \neq 0$ is a constant. Then, since

$$x_{j+k+1} - x_j = (k+1)h,$$

we see from (1.66) that the denominators of the divided differences are constant in any one column. In this case, it is natural to concentrate on the numerators of the divided differences, which are simply *differences*. We write

$$f(x_{j+1}) - f(x_j) = f(x_j + h) - f(x_j) = \Delta f(x_j), \tag{1.67}$$

which is called a first difference. The symbol Δ is the Greek capital delta, denoting *difference*. It follows that with equally spaced x_j, we can express a first-order divided difference in terms of a first difference, as

$$f[x_j, x_{j+1}] = \frac{\Delta f(x_j)}{h},$$

since $x_{j+1} - x_j = h$. If we make the linear change of variable

$$s = \frac{1}{h}(x - x_0), \tag{1.68}$$

then the abscissa $x_j = x_0 + jh$ is mapped to j. Without any loss of generality, we will write $x_j = j$. Then

$$f[x_j, x_{j+1}] = \Delta f(x_j), \tag{1.69}$$

and

$$f[x_j, x_{j+1}, x_{j+2}] = \frac{f[x_{j+1}, x_{j+2}] - f[x_j, x_{j+1}]}{x_{j+2} - x_j} = \frac{1}{2}\left(\Delta f(x_{j+1}) - \Delta f(x_j)\right).$$

It is convenient to define

$$\Delta^2 f(x_j) = \Delta f(x_{j+1}) - \Delta f(x_j),$$

and consequently, the second-order divided difference may be expressed in the form

$$f[x_j, x_{j+1}, x_{j+2}] = \frac{1}{2}\,\Delta^2 f(x_j). \tag{1.70}$$

We call $\Delta^2 f(x_j)$ a second-order forward difference. The reader may wish to explore what happens when we express a third-order divided difference in terms of second-order forward differences. Based on such evidence, it then seems natural to define higher-order forward differences recursively, as

$$\Delta^{k+1} f(x_j) = \Delta\left(\Delta^k f(x_j)\right) = \Delta^k f(x_{j+1}) - \Delta^k f(x_j), \tag{1.71}$$

for $k \geq 1$, where $\Delta^1 f(x_j) = \Delta f(x_j)$. It is helpful to extend the definition of (1.71) to include $k = 0$, defining

$$\Delta^0 f(x_j) = f(x_j). \tag{1.72}$$

On pursuing the evaluation of a third-order divided difference, we obtain

$$f[x_j, x_{j+1}, x_{j+2}, x_{j+3}] = \frac{1}{6}\,\Delta^3 f(x_j),$$

and it is not difficult to guess and justify the following general relation that connects divided differences and forward differences.

Theorem 1.3.1 For all $j, k \geq 0$, we have

$$f[x_j, x_{j+1}, \ldots, x_{j+k}] = \frac{1}{k!}\,\Delta^k f(x_j), \tag{1.73}$$

where $x_j = j$.

Proof. The proof is by induction on k. The result clearly holds for $k = 0$ and all $j \geq 0$. Let us assume that it holds for some $k \geq 0$ and all $j \geq 0$. Then

$$\begin{aligned} f[x_j, \ldots, x_{j+k+1}] &= \frac{f[x_{j+1}, \ldots, x_{j+k+1}] - f[x_j, \ldots, x_{j+k}]}{x_{j+k+1} - x_j} \\ &= \frac{1}{k+1} \left(\frac{\Delta^k f(x_{j+1})}{k!} - \frac{\Delta^k f(x_j)}{k!} \right) \\ &= \frac{1}{(k+1)!} \Delta^{k+1} f(x_j), \end{aligned}$$

on using (1.71). This shows that (1.73) holds when k is replaced by $k+1$, which completes the proof. ∎

In mathematical tables, functions are usually tabulated at equal intervals, as in the first tables of logarithms, which appeared in the early seventeenth century. This gave impetus to the study of interpolation at equally spaced abscissas. Let us therefore see how Newton's divided difference formula (1.19) simplifies when we interpolate at the abscissas $0, 1, \ldots, n$. First we have

$$\pi_k(x) = x(x-1)(x-2)\cdots(x-k+1),$$

for $k > 0$, with $\pi_0(x) = 1$, and (1.73) gives

$$f[0, 1, \ldots, k] = \frac{1}{k!} \Delta^k f(0).$$

It then follows from (1.19) that the interpolating polynomial for f constructed at the abscissas $0, 1, \ldots, n$ may be written in the form

$$p_n(x) = f(0) + \frac{\Delta f(0)}{1!}\pi_1(x) + \frac{\Delta^2 f(0)}{2!}\pi_2(x) + \cdots + \frac{\Delta^n f(0)}{n!}\pi_n(x).$$

We can express p_n in terms of binomial coefficients, since we have

$$\frac{1}{k!}\,\pi_k(x) = \frac{x(x-1)(x-2)\cdots(x-k+1)}{k!} = \binom{x}{k},$$

thus giving the simpler form

$$p_n(x) = f(0) + \Delta f(0) \binom{x}{1} + \cdots + \Delta^n f(0) \binom{x}{n}. \tag{1.74}$$

This is called the *forward difference formula* for the interpolating polynomial, which we can write more succinctly as

$$p_n(x) = \sum_{k=0}^{n} \Delta^k f(0) \binom{x}{k}.$$

$$
\begin{array}{llll}
f(x_0) & & & \\
 & \Delta f(x_0) & & \\
f(x_1) & & \Delta^2 f(x_0) & \\
 & \Delta f(x_1) & & \Delta^3 f(x_0) \\
f(x_2) & & \Delta^2 f(x_1) & \\
 & \Delta f(x_2) & & \Delta^3 f(x_1) \\
f(x_3) & & \Delta^2 f(x_2) & \\
 & \Delta f(x_3) & & \\
f(x_4) & & &
\end{array}
$$

TABLE 1.3. A systematic scheme for calculating forward differences.

Although the forward difference form of the interpolating polynomial is often attributed to Isaac Newton and his contemporary James Gregory (1638–1675), it was used before their time by Henry Briggs (1556–1630) and Thomas Harriot (1560–1621). For further information on the history of interpolation see, for example, Goldstine [21], Edwards [15], Phillips [44].

To evaluate the forward difference formula (1.74), we first compute a table of forward differences (see Table 1.3), which is laid out in a manner similar to that of the divided difference Table 1.1. The only entries in Table 1.3 that are required for the evaluation of the interpolating polynomial $p_n(x)$, defined by (1.74), are the first numbers in each column, namely, $f(x_0)$, $\Delta f(x_0)$, and so on. From the uniqueness of the interpolating polynomial, if $f(x)$ is itself a polynomial of degree k, then its interpolating polynomial $p_n(x)$ will be equal to $f(x)$ for $n \geq k$. It follows from (1.74) that kth-order differences of a polynomial of degree k must be constant, and differences of order greater than k must be zero.

Example 1.3.1 In Example 1.1.1 we found Newton's divided difference form of the interpolating polynomial for the function f, given in a table that we repeat here:

x	0	1	2	3
$f(x)$	6	−3	−6	9

This time we will use the forward difference formula (1.74). First we compute the following table of differences, whose entries are defined as in Table 1.3 above:

$$
\begin{array}{llll}
6 & & & \\
 & -9 & & \\
-3 & & 6 & \\
 & -3 & & 12 \\
-6 & & 18 & \\
 & 15 & & \\
9 & & &
\end{array}
$$

Only the first number from each column of this table is needed in constructing the interpolating polynomial, which, from (1.74), is

$$p_3(x) = 6\binom{x}{0} - 9\binom{x}{1} + 6\binom{x}{2} + 12\binom{x}{3},$$

and this can be rewritten in the form

$$p_3(x) = 6 - 9x + 3x(x-1) + 2x(x-1)(x-2), \tag{1.75}$$

as we found in Example 1.1.1. ■

Let us suppose that we are given a function f in some implicit form, and know its value at a sufficiently large number of abscissas. Then if we happen to know that f is a polynomial, we can obviously determine it explicitly by evaluating the interpolating polynomial. This is illustrated in the following example.

Example 1.3.2 We apply this technique to evaluate the complete symmetric function $\tau_2(1,\ldots,n)$, given that $\tau_1(1,\ldots,n) = \frac{1}{2}n(n+1)$, so that the recurrence relation (1.45) gives

$$\tau_2(1,\ldots,n) - \tau_2(1,\ldots,n-1) = \frac{1}{2}n^2(n+1). \tag{1.76}$$

Now, $\tau_2(1,\ldots,n)$ is a function of n, which we know from Problem 1.2.8 to be a polynomial of degree 4, and we know the first differences of τ_2 from (1.76). We can therefore construct a difference table for τ_2. In the column for τ_2 we insert its value of 1 for $n = 1$ and put "bullets" in place of the other values of τ_2:

n	$\tau_2(1,\ldots,n)$					
1	1					
		6				
2	•		12			
		18		10		
3	•		22		3	
		40		13		
4	•		35		3	
		75		16		
5	•		51			
		126				
6	•					

The third column in the above table contains the values of $\frac{1}{2}n^2(n+1)$ for $2 \le n \le 6$. We have continued the table as far as $n = 6$, one value of n more than is necessary, so as to obtain two values in the sixth column of the table,

corresponding to fourth differences of τ_2. These two values, both equal to 3, give a reassuring check on the fact that fourth differences of a polynomial of degree four are constant. We can now write down $\tau_2(1,\ldots,n)$ in its interpolating polynomial form, using the forward difference form (1.74). We will need to replace x in (1.74) by $n-1$, since $x=0$ corresponds to $n=1$, which is $n-1=0$, and insert the appropriate forward differences. We can thus write $\tau_2(1,\ldots,n)$ in the form

$$\binom{n-1}{0}+6\binom{n-1}{1}+12\binom{n-1}{2}+10\binom{n-1}{3}+3\binom{n-1}{4},$$

and a small calculation yields

$$\tau_2(1,\ldots,n)=\frac{1}{24}n(n+1)(n+2)(3n+1), \tag{1.77}$$

as obtained by other means in Problem 1.2.6. Note how, even before obtaining the above explicit expression for τ_2, we could have determined its values for $2\le n\le 6$ by using the column of first differences in the above table. Thus we obtain, in turn,

$$\tau_2(1,2)=1+6=7,\quad \tau_2(1,2,3)=7+18=25,$$

and so on, finally obtaining $\tau_2(1,\ldots,6)=266$, which agrees with (1.77). We can also add to our above table by inserting numbers, one at a time, beginning with the sixth column by inserting a 3, since the fourth differences are known to be constant. Then, in turn we can insert

$$3+16=19,\quad 19+51=70,\quad 70+126=196,\quad 196+266=462$$

in columns 5, 4, 3, and 2, respectively. This last number in column 2 then gives $\tau_2(1,\ldots,7)=462$, in agreement with (1.77). ■

In our above study of interpolation at equally spaced abscissas we found it convenient to make a linear change of variable so that $x_j=j$. We will now write down explicit forms, for *any* set of equally spaced abscissas, for the main results derived above for the special case where $x_j=j$. Thus we readily find (see Problem 1.3.1) that when $x_j=x_0+jh$, for all $j\ge 0$, (1.73) becomes

$$f[x_j,x_{j+1},\ldots,x_{j+k}]=\frac{1}{k!\,h^k}\Delta^k f(x_j), \tag{1.78}$$

and (see Problem 1.3.2) the forward difference form of the interpolating polynomial (1.74) becomes

$$p_n(x_0+sh)=f(x_0)+\Delta f(x_0)\binom{s}{1}+\cdots+\Delta^n f(x_0)\binom{s}{n}, \tag{1.79}$$

where the new variable s is defined by $x = x_0 + sh$. It also follows immediately from (1.33) and (1.78) that if the conditions of Theorem 1.1.2 apply, then

$$\frac{\Delta^n f(x_0)}{h^n} = f^{(n)}(\xi), \tag{1.80}$$

where $\xi \in (x_0, x_n)$. On letting $h \to 0$, we obtain

$$\lim_{h\to 0} \frac{\Delta^n f(x_0)}{h^n} = f^{(n)}(x_0), \tag{1.81}$$

provided that $f^{(n)}$ is continuous.

It is easy to see, by using induction on k, that $\Delta^k f(x_j)$ may be expressed as a sum of multiples of $f(x_j)$, $f(x_{j+1})$, $\ldots$, $f(x_{j+k})$. In fact, we may prove by induction or deduce from (1.21), the symmetric form for a divided difference, that

$$\Delta^k f(x_j) = \sum_{r=0}^{k} (-1)^r \binom{k}{r} f(x_{j+k-r}). \tag{1.82}$$

Let us recall the Leibniz rule for the differentiation of a product of two differentiable functions: If the kth derivatives of f and g both exist, then

$$\frac{d^k}{dx^k}\left(f(x)g(x)\right) = \sum_{r=0}^{k} \binom{k}{r} \frac{d^r}{dx^r} f(x) \, \frac{d^{k-r}}{dx^{k-r}} g(x). \tag{1.83}$$

This rule is named after Gottfried Leibniz (1646–1716). We now state an analogous result involving the kth *difference* of a product. We omit the proof, since (see Problem 1.3.5) it is easily deduced from Theorem 1.3.3.

Theorem 1.3.2 For any integer $k \geq 0$, we have

$$\Delta^k \left(f(x_j)g(x_j)\right) = \sum_{r=0}^{k} \binom{k}{r} \Delta^r f(x_j) \, \Delta^{k-r} g(x_{j+r}). \qquad \blacksquare \tag{1.84}$$

The Leibniz rule for differentiation (1.83) can be proved directly by induction. However, if we divide both sides of (1.84) by h^k and take limits as $h \to 0$, then the Leibniz rule for differentiation (1.83) follows immediately, on using (1.81).

In the proof of the Leibniz rule for the divided difference of a product, it is convenient to use the "operator" notation for divided differences, which we introduced in (1.20). It is easily verified that

$$[x_0, x_1]fg = [x_0]f \cdot [x_0, x_1]g + [x_0, x_1]f \cdot [x_1]g, \tag{1.85}$$

and, with a little more work, we can show that

$$[x_0, x_1, x_2]fg = [x_0]f \cdot [x_0, x_1, x_2]g + [x_0, x_1]f \cdot [x_1, x_2]g + [x_0, x_1, x_2]f \cdot [x_2]g.$$

This beautiful relation suggests the form of the general result, which we now state:

Theorem 1.3.3 The kth divided difference of a product of two functions satisfies the relation

$$[x_j, x_{j+1}, \ldots, x_{j+k}]fg = \sum_{r=0}^{k} [x_j, \ldots, x_{j+r}]f \cdot [x_{j+r}, \ldots, x_{j+k}]g \qquad (1.86)$$

for $k \geq 0$.

Proof. We can set $j = 0$, without loss of generality, to simplify the presentation of the proof. It is clear that this result holds for $k = 0$, and we saw above in (1.85) that it also holds for $k = 1$. We complete the proof by induction. Let us assume that (1.86) holds for some $k \geq 1$. Then let us write

$$[x_0, x_1, \ldots, x_{k+1}]fg = \frac{[x_1, \ldots, x_{k+1}]fg - [x_0, \ldots, x_k]fg}{x_{k+1} - x_0}. \qquad (1.87)$$

On setting $j = 1$ in (1.86), we can express the first term in the numerator on the right of (1.87) in the form

$$[x_1, \ldots, x_{k+1}]fg = \sum_{r=0}^{k} [x_1, \ldots, x_{r+1}]f \cdot [x_{r+1}, \ldots, x_{k+1}]g. \qquad (1.88)$$

From (1.86), with $j = 0$, the second term in the numerator on the right of (1.87) is

$$[x_0, \ldots, x_k]fg = \sum_{r=0}^{k} [x_0, \ldots, x_r]f \cdot [x_r, \ldots, x_k]g. \qquad (1.89)$$

We now replace the rth term on the right of (1.88) by

$$\Big\{[x_0, \ldots, x_r]f + (x_{r+1} - x_0)[x_0, \ldots, x_{r+1}]f\Big\} \cdot [x_{r+1}, \ldots, x_{k+1}]g \qquad (1.90)$$

and the rth term on the right of (1.89) by

$$[x_0, \ldots, x_r]f \cdot \Big\{[x_{r+1}, \ldots, x_{k+1}]g - (x_{k+1} - x_r)[x_r, \ldots, x_{k+1}]g\Big\}. \qquad (1.91)$$

Note that (1.90) and (1.91) have the term $[x_0, \ldots, x_r]f \cdot [x_{r+1}, \ldots, x_{k+1}]g$ in common. This pair of terms cancel because $[x_0, \ldots, x_k]fg$ has a negative sign in the numerator on the right of (1.87). This numerator may therefore be expressed in the form $S_1 + S_2$, say, where

$$S_1 = \sum_{r=0}^{k} (x_{r+1} - x_0)[x_0, \ldots, x_{r+1}]f \cdot [x_{r+1}, \ldots, x_{k+1}]g, \qquad (1.92)$$

obtained from summing the uncancelled term in (1.90), and

$$S_2 = \sum_{r=0}^{k} (x_{k+1} - x_r)[x_0, \ldots, x_r]f \cdot [x_r, \ldots, x_{k+1}]g, \qquad (1.93)$$

obtained similarly from (1.91). It remains only to replace r by $s-1$ in (1.92), and then write r in place of s, to give

$$S_1 = \sum_{r=1}^{k+1}(x_r - x_0)[x_0, \ldots, x_r]f \cdot [x_r, \ldots, x_{k+1}]g. \tag{1.94}$$

We then obtain from (1.94) and (1.93) that

$$S_1 + S_2 = (x_{k+1} - x_0)\sum_{r=0}^{k+1}[x_0, \ldots, x_r]f \cdot [x_r, \ldots, x_{k+1}]g$$

and, on dividing both sides of the latter equation by $x_{k+1} - x_0$, we deduce from (1.87) that

$$[x_0, x_1, \ldots, x_{k+1}]fg = \sum_{r=0}^{k+1}[x_0, x_1, \ldots, x_r]f \cdot [x_r, \ldots, x_{k+1}]g.$$

This completes the proof. ■

There is a variant of the forward difference formula in which the interpolating polynomial is written in terms of *backward* differences. We define the first-order backward difference,

$$\nabla f(x_j) = f(x_j) - f(x_{j-1}), \tag{1.95}$$

where the abscissas $x_j = x_0 + jh$ are equally spaced. Then we may write

$$f[x_{j-1}, x_j] = \frac{\nabla f(x_j)}{h}.$$

When we explore how a second-order divided difference might be written in terms of ∇, we find it natural to define

$$\nabla^2 f(x_j) = \nabla f(x_j) - \nabla f(x_{j-1}),$$

for then

$$f[x_{j-2}, x_{j-1}, x_j] = \frac{1}{2h}\left(\frac{\nabla f(x_j)}{h} - \frac{\nabla f(x_{j-1})}{h}\right) = \frac{\nabla^2 f(x_j)}{2h^2}.$$

Recall that the order of the arguments in a divided difference does not matter, and so we could alternatively write

$$f[x_j, x_{j-1}, x_{j-2}] = \frac{\nabla^2 f(x_j)}{2h^2}.$$

We now proceed in the same way as we did above in our study of forward differences and, following the behaviour of divided differences, we find it expedient to define higher-order backward differences recursively, writing

$$\nabla^{k+1} f(x_j) = \nabla\left(\nabla^k f(x_j)\right) = \nabla^k f(x_j) - \nabla^k f(x_{j-1}), \tag{1.96}$$

$$
\begin{array}{llll}
f(x_0) & & & \\
 & \nabla f(x_1) & & \\
f(x_1) & & \nabla^2 f(x_2) & \\
 & \nabla f(x_2) & & \nabla^3 f(x_3) \\
f(x_2) & & \nabla^2 f(x_3) & \\
 & \nabla f(x_3) & & \nabla^3 f(x_4) \\
f(x_3) & & \nabla^2 f(x_4) & \\
 & \nabla f(x_4) & & \\
f(x_4) & & &
\end{array}
$$

TABLE 1.4. A systematic scheme for calculating backward differences.

for $k \geq 0$, where $\nabla^0 f(x_j) = f(x_j)$ and $\nabla^1 f(x_j) = \nabla f(x_j)$. It is then easily verified by induction on k that

$$f[x_{n-j}, x_{n-j+1}, \ldots, x_n] = f[x_n, x_{n-1}, \ldots, x_{n-j}] = \frac{\nabla^j f(x_n)}{j!\, h^j}. \tag{1.97}$$

We now recast the divided difference formula (1.19) by taking the distinct abscissas $x_0, x_1, \ldots, x_n$ in reverse order, to give

$$p_n(x) = \sum_{j=0}^{n} f[x_n, \ldots, x_{n-j}] \prod_{r=0}^{j-1} (x - x_{n-r}), \tag{1.98}$$

where the empty product (corresponding to $j = 0$) denotes 1. On making the change of variable $x = x_n + sh$, we obtain

$$\prod_{r=0}^{j-1} (x - x_{n-r}) = h^j \prod_{r=0}^{j-1} (s + r) = (-1)^j h^j \prod_{r=0}^{j-1} (-s - r),$$

and, on using this and (1.97), we have

$$f[x_n, \ldots, x_{n-j}] \prod_{r=0}^{j-1} (x - x_{n-r}) = (-1)^j \binom{-s}{j} \nabla^j f(x_n).$$

If we now sum this last equation over j, we see from (1.98) that

$$p_n(x_n + sh) = \sum_{j=0}^{n} (-1)^j \binom{-s}{j} \nabla^j f(x_n), \tag{1.99}$$

which is called the *backward difference formula* for the interpolating polynomial.

To evaluate the backward difference formula (1.99), we first compute a table of backward differences (see Table 1.4), which is similar to the forward difference Table 1.3. The only entries in Table 1.4 required for evaluating

$p_n(x)$, defined by (1.99), are the last numbers in each column, namely $f(x_n)$, $\nabla f(x_n)$, and so on. Note that the *numbers* in Tables 1.3 and 1.4 are the same. Thus, for example, $\nabla f(x_1) = \Delta f(x_0)$, $\nabla^2 f(x_3) = \Delta^2 f(x_1)$, and in general,

$$\nabla^k f(x_j) = \Delta^k f(x_{j-k}), \tag{1.100}$$

for $j \geq k \geq 0$.

Example 1.3.3 Let us evaluate the backward difference formula for the data that we used in Example 1.3.1 to evaluate the forward difference formula. Since the abscissas are 0, 1, 2, and 3, we have $n = 3$, $h = 1$, and $x_3 = 3$ in (1.99). To evaluate (1.99) we require the values

$$f(x_3) = 9, \quad \nabla f(x_3) = 15, \quad \nabla^2 f(x_3) = 18, \quad \nabla^3 f(x_3) = 12,$$

obtained from the difference table in Example 1.3.1. We can thus write

$$p_3(s+3) = 9 - 15(-s) + 18\frac{(-s)(-s-1)}{2} - 12\frac{(-s)(-s-1)(-s-2)}{6},$$

which simplifies to give

$$p_3(s+3) = 9 + 28s + 15s^2 + 2s^3. \tag{1.101}$$

As a check, let us substitute $s = x-3$ in (1.101), and simplify the resulting polynomial in x to give

$$p_3(x) = 6 - 8x - 3x^2 + 2x^3,$$

which agrees with the expression for $p_3(x)$ given in (1.75) at the end of Example 1.3.1. ■

Problem 1.3.1 If $x_j = x_0 + jh$, for $j \geq 0$, show by induction on k that

$$f[x_j, x_{j+1}, \ldots, x_{j+k}] = \frac{1}{k!\, h^k}\Delta^k f(x_j).$$

Problem 1.3.2 Given that $x_j = x_0 + jh$, for $j \geq 0$, make the change of variable $x = x_0 + sh$ and hence, using the result of Problem 1.3.1, show that a typical term of Newton's divided difference formula (1.19) may be expressed as

$$f[x_0, x_1, \ldots, x_j] \cdot \pi_j(x) = \frac{\Delta^j f(x_0)}{j!\, h^j} \cdot h^j s(s-1)\cdots(s-j+1),$$

and thus derive the forward difference formula,

$$p_n(x_0 + sh) = f(x_0) + \Delta f(x_0)\begin{pmatrix} s \\ 1 \end{pmatrix} + \cdots + \Delta^n f(x_0)\begin{pmatrix} s \\ n \end{pmatrix}.$$

Problem 1.3.3 Consider the function 2^x evaluated at $x = 0, 1, 2$, and so on. Show that

$$\Delta\, 2^x = 2^{x+1} - 2^x = 2^x,$$

and thus show by induction on k, using (1.71), that

$$\Delta^k\, 2^x = 2^x$$

for all nonnegative integers k. Hence verify that the interpolating polynomial for 2^x constructed at $x = 0, 1, \ldots, n$ is

$$p_n(x) = \sum_{r=0}^{n} \binom{x}{r}.$$

Problem 1.3.4 Let us define

$$S(n) = 0^4 + 1^4 + 2^4 + \cdots + n^4,$$

for $n \geq 0$. Evaluate $S(n)$ for $0 \leq n \leq 5$ and compute a table of forward differences. Hence, on the assumption that S is a polynomial in n of degree 5, show that

$$S(n) = \binom{n}{1} + 15\binom{n}{2} + 50\binom{n}{3} + 60\binom{n}{4} + 24\binom{n}{5},$$

and verify that this simplifies to give

$$\sum_{r=1}^{n} r^4 = \frac{1}{30} n(n+1)(2n+1)(3n^2+3n-1), \quad n \geq 1.$$

Problem 1.3.5 Beginning with (1.86), the expression for the divided difference of a product of two functions, use (1.78) to deduce the expression (1.84) for a forward difference of a product.

Problem 1.3.6 Verify (1.97), the relation between divided differences and backward differences.

Problem 1.3.7 Beginning with the symmetric form (1.21), show that for equally spaced x_j,

$$\Delta^k f(x_j) = \sum_{r=0}^{k} (-1)^r \binom{k}{r} f(x_{j+k-r}),$$

and deduce that

$$\nabla^k f(x_j) = \sum_{r=0}^{k} (-1)^r \binom{k}{r} f(x_{j-r}).$$

1.4 Central Differences

When we use the forward difference formula (1.74), we begin with $f(x_0)$, and then add terms involving $\Delta f(x_0)$, $\Delta^2 f(x_0)$, and so on. Thus, obviously, the first term involves only $f(x_0)$, the first two terms involve only $f(x_0)$ and $f(x_1)$, and so on. As we add more terms in building up the forward difference formula, we bring in values $f(x_j)$ that are increasingly farther from x_0. Likewise, as we add successive terms in the backward difference formula (1.99), we bring in values $f(x_j)$ that are increasingly farther from x_n. These observations motivated the study of *central* differences. Let $x_j = x_0 + jh$ be defined for *all* integers j, and not just for $j \geq 0$. Then we define

$$\delta f(x) = f(x + \tfrac{1}{2}h) - f(x - \tfrac{1}{2}h), \tag{1.102}$$

which we call a first-order central difference. In a similar fashion to the definition of higher-order forward and backward differences, we define higher-order central differences recursively from

$$\delta^{k+1} f(x) = \delta\left(\delta^k f(x)\right), \tag{1.103}$$

for $k \geq 0$, where $\delta^0 f(x) = f(x)$ and $\delta^1 f(x) = \delta f(x)$. It readily follows from (1.102) and (1.103) that

$$\delta^2 f(x_j) = f(x_{j+1}) - 2f(x_j) + f(x_{j-1}),$$

so that

$$\delta^2 f(x_j) = \Delta^2 f(x_{j-1}) = \nabla^2 f(x_{j+1}).$$

Now let us begin with the forward difference formula and, in place of the original abscissas

$$x_0, x_1, x_2, x_3, x_4, \ldots,$$

where these are distinct, but are otherwise arbitrary, we will use abscissas taken in the order

$$x_0, x_1, x_{-1}, x_2, x_{-2}, x_3, x_{-3}, \ldots,$$

where each x_j is equal to $x_0 + jh$. On putting $x = x_0 + sh$, we may verify that the divided difference formula (1.19) begins

$$f(x_0) + \binom{s}{1} \delta f(x_0 + \tfrac{1}{2}h) + \binom{s}{2} \delta^2 f(x_0) + \binom{s+1}{3} \delta^3 f(x_0 + \tfrac{1}{2}h),$$

where the general even-order and odd-order terms are

$$\binom{s+k-1}{2k} \delta^{2k} f(x_0) \quad \text{and} \quad \binom{s+k}{2k+1} \delta^{2k+1} f(x_0 + \tfrac{1}{2}h), \tag{1.104}$$

$$
\begin{array}{llll}
f(x_{-2}) & & & \\
 & \delta f(x_{-1}-\frac{1}{2}h) & & \\
f(x_{-1}) & & \delta^2 f(x_{-1}) & \\
 & \delta f(x_0-\frac{1}{2}h) & & \delta^3 f(x_0-\frac{1}{2}h) \\
f(x_0) & & \delta^2 f(x_0) & \\
 & \delta f(x_0+\frac{1}{2}h) & & \delta^3 f(x_0+\frac{1}{2}h) \\
f(x_1) & & \delta^2 f(x_1) & \\
 & \delta f(x_1+\frac{1}{2}h) & & \\
f(x_2) & & &
\end{array}
$$

TABLE 1.5. A systematic scheme for calculating central differences.

respectively. The above form of the interpolating polynomial is known as Gauss's forward formula, named after C. F. Gauss (1777–1855). If, on the other hand, we introduce the abscissas in the order

$$x_0, x_{-1}, x_1, x_{-2}, x_2, x_{-3}, x_3, \ldots,$$

we obtain a form of the interpolating polynomial that begins

$$f(x_0)+\binom{s}{1}\delta f(x_0-\tfrac{1}{2}h)+\binom{s+1}{2}\delta^2 f(x_0)+\binom{s+1}{3}\delta^3 f(x_0-\tfrac{1}{2}h),$$

where the general even-order and odd-order terms are

$$\binom{s+k}{2k}\delta^{2k} f(x_0) \quad \text{and} \quad \binom{s+k}{2k+1}\delta^{2k+1} f(x_0-\tfrac{1}{2}h), \tag{1.105}$$

respectively. This second form of the interpolating polynomial in terms of central differences is known as Gauss's backward formula. Table 1.5 shows a difference table involving central differences. We note how each of the two formulas of Gauss chooses differences that lie on a zigzag path through the difference table, beginning with $f(x_0)$.

To achieve symmetry about x_0, it is tempting to take the *mean* of these two interpolating formulas of Gauss. Another item of notation is helpful in pursuing this idea: Let us write

$$\mu f(x_0) = \frac{1}{2}\left(f(x_0+\tfrac{1}{2}h)+f(x_0-\tfrac{1}{2}h)\right), \tag{1.106}$$

where μ is called the *averaging operator*. Then the means of the odd-order differences occurring in the two interpolating formulas of Gauss may be written as $\mu\delta f(x_0)$, $\mu\delta^3 f(x_0)$, and so on, and the mean of Gauss's two formulas begins

$$f(x_0)+\binom{s}{1}\mu\delta f(x_0)+\frac{s}{2}\binom{s}{1}\delta^2 f(x_0)+\binom{s+1}{3}\mu\delta^3 f(x_0),$$

where the general even-order (for $k \geq 1$) and odd-order terms are

$$\frac{s}{2k}\begin{pmatrix} s+k-1 \\ 2k-1 \end{pmatrix}\delta^{2k}f(x_0) \quad \text{and} \quad \begin{pmatrix} s+k \\ 2k+1 \end{pmatrix}\mu\delta^{2k+1}f(x_0), \quad (1.107)$$

respectively. This is called Stirling's interpolation formula, named after James Stirling (1692–1770). We may verify that

$$\mu\delta f(x_0) = \frac{1}{2}\left(f(x_1) - f(x_{-1})\right),$$

and if we truncate Stirling's interpolation formula after the first three terms, we obtain

$$p_2(x_0 + sh) = f(x_0) + \frac{s}{2}(f(x_1) - f(x_{-1})) + \frac{s^2}{2}(f(x_1) - 2f(x_0) + f(x_{-1})).$$

If we truncate Stirling's interpolation formula after the term involving $\delta^{2k}f(x_0)$, we obtain the interpolating polynomial $p_{2k}(x_0 + sh)$, which interpolates f at the abscissas x_{-k}, $x_{-k+1}, \ldots, x_{k-1}$, x_k.

Problem 1.4.1 Verify that

$$\mu\delta^3 f(x_0) = \delta^3(\mu f(x_0))$$

and that

$$\mu\delta^3 f(x_0) = \frac{1}{2}(f(x_2) - 2f(x_1) + 2f(x_{-1}) - f(x_{-2})).$$

Problem 1.4.2 Derive an expansion of $\delta^{2k+1}f(x_0)$ by adapting the corresponding relation (see (1.82)) for forward differences. Hence verify that

$$\mu\delta^{2k+1}f(x_0) = \frac{1}{2}\sum_{r=0}^{k}(-1)^r a_r(f(x_{k+1-r}) - f(x_{-k-1+r})),$$

where $a_0 = 1$ and

$$a_r = \begin{pmatrix} 2k+1 \\ r \end{pmatrix} - \begin{pmatrix} 2k+1 \\ r-1 \end{pmatrix}$$

for $1 \leq r \leq k$.

Problem 1.4.3 Show that

$$\delta^{2k}f(x_0) = (-1)^k\begin{pmatrix} 2k \\ k \end{pmatrix}f(x_0) + \sum_{r=0}^{k-1}(-1)^r\begin{pmatrix} 2k \\ r \end{pmatrix}(f(x_{k-r}) + f(x_{-k+r})).$$

1.5 q-Differences

As we saw in Section 1.1, the divided difference formula (1.19) allows us to interpolate on *any* set of distinct abscissas. Nevertheless, we derived special forms of the interpolating polynomial when the intervals between consecutive abscissas are equal, namely the forward difference formula (1.74), the backward difference formula (1.99), and the interpolating formulas of Gauss and Stirling based on central differences. In this section we will consider another special form of the interpolating polynomial, where the intervals between consecutive abscissas are not equal, but are in geometric progression. Let us denote the common ratio of the geometric progression by q. Without loss of generality, we can make a linear change of variable, shifting the origin so that $x_0 = 0$, and scaling the axis so that $x_1 - x_0 = 1$. It then follows that x_j is equal to the q-integer $[j]$, defined by

$$[j] = \begin{cases} (1-q^j)/(1-q), & q \neq 1, \\ j, & q = 1. \end{cases}$$

(More material on q-integers is presented in Chapter 8.) As we observed in Section 1.3, when we compute divided differences in Table 1.1 we repeatedly calculate quotients of the form

$$\frac{f[x_{j+1}, \ldots, x_{j+k+1}] - f[x_j, \ldots, x_{j+k}]}{x_{j+k+1} - x_j},$$

where k has the same value throughout any one column of the divided difference table, and we noted that the above denominator is independent of j when the abscissas x_j are equally spaced. Let us explore what happens to this denominator when $x_j = [j]$. We have

$$x_{j+k+1} - x_j = \frac{1-q^{j+k+1}}{1-q} - \frac{1-q^j}{1-q} = q^j[k+1], \tag{1.108}$$

which is not independent of j, although it does have the common factor $[k+1]$. Now when $k = 0$ we have

$$f[x_j, x_{j+1}] = \frac{f(x_{j+1}) - f(x_j)}{x_{j+1} - x_j} = \frac{f(x_{j+1}) - f(x_j)}{q^j}.$$

It is convenient to define

$$f(x_{j+1}) - f(x_j) = \Delta_q f(x_j), \tag{1.109}$$

so that

$$f[x_j, x_{j+1}] = \frac{\Delta_q f(x_j)}{q^j}. \tag{1.110}$$

Thus the q-difference $\Delta_q f(x_j)$ is exactly the same as the forward difference $\Delta f(x_j)$. However, we will see that the kth-order q-difference $\Delta_q^k f(x_j)$,

which we define below, and the kth order forward difference $\Delta^k f(x_j)$ are not the same for $k \geq 2$, unless $q = 1$. From (1.22), the recurrence relation for divided differences, and (1.110) we next obtain

$$f[x_j, x_{j+1}, x_{j+2}] = \left(\frac{\Delta_q f(x_{j+1})}{q^{j+1}} - \frac{\Delta_q f(x_j)}{q^j} \right) \Big/ \, (x_{j+2} - x_j) \, .$$

It follows from (1.108) that

$$x_{j+2} - x_j = q^j [2],$$

and thus the second-order divided difference may be written as

$$f[x_j, x_{j+1}, x_{j+2}] = \frac{\Delta_q f(x_{j+1}) - q \, \Delta_q f(x_j)}{q^{2j+1}[2]}. \tag{1.111}$$

In view of (1.111) we *define*

$$\Delta_q^2 f(x_j) = \Delta_q f(x_{j+1}) - q \, \Delta_q f(x_j),$$

so that we may write

$$f[x_j, x_{j+1}, x_{j+2}] = \frac{\Delta_q^2 f(x_j)}{q^{2j+1}[2]}.$$

It is helpful to gather more evidence, by evaluating the third-order divided difference $f[x_j, x_{j+1}, x_{j+2}, x_{j+3}]$ when each abscissa x_j is equal to $[j]$. We find that

$$f[x_j, x_{j+1}, x_{j+2}, x_{j+3}] = \frac{\Delta_q^2 f(x_{j+1}) - q^2 \, \Delta_q f(x_j)}{q^{3j+3}[3]!},$$

where $[3]! = [3]\,[2]\,[1]$. It thus seems natural to define higher-order q-differences recursively as follows. We write

$$\Delta_q^{k+1} f(x_j) = \Delta_q^k f(x_{j+1}) - q^k \, \Delta_q^k f(x_j) \tag{1.112}$$

for all integers $k \geq 0$, where $\Delta_q^0 f(x_j) = f(x_j)$ and $\Delta_q^1 f(x_j) = \Delta_q f(x_j)$. Note that (1.112) reduces to (1.71), the corresponding relation for forward differences, when $q = 1$. We now state and prove the relation between divided differences and q-differences.

Theorem 1.5.1 For all $j, k \geq 0$, we have

$$f[x_j, x_{j+1}, \ldots, x_{j+k}] = \frac{\Delta_q^k f(x_j)}{q^{k(2j+k-1)/2}\,[k]!}, \tag{1.113}$$

where each x_j equals $[j]$, and $[k]! = [k]\,[k-1] \cdots [1]$.

Proof. The proof is by induction on k. The result clearly holds for $k = 0$ and all $j \geq 0$. Let us assume that it holds for some $k \geq 0$ and all $j \geq 0$. Then

$$\begin{aligned} f[x_j, \ldots, x_{j+k+1}] &= \frac{f[x_{j+1}, \ldots, x_{j+k+1}] - f[x_j, \ldots, x_{j+k}]}{x_{j+k+1} - x_j} \\ &= \frac{1}{q^j[k+1]} \left(\frac{\Delta_q^k f(x_{j+1})}{q^{k(2j+k+1)/2}\,[k]!} - \frac{\Delta_q^k f(x_j)}{q^{k(2j+k-1)/2}\,[k]!} \right) \\ &= \frac{\Delta_q^k f(x_{j+1}) - q^k\, \Delta_q^k f(x_j)}{q^{(k+1)(2j+k)/2}\,[k+1]!} \\ &= \frac{\Delta_q^{k+1} f(x_j)}{q^{(k+1)(2j+k)/2}\,[k+1]!}, \end{aligned}$$

on using (1.112). This shows that (1.113) holds when k is replaced by $k+1$, and this completes the proof. ■

Now let us see how Newton's divided difference formula (1.19) simplifies when we interpolate at the abscissas $x_j = [j]$ for $j = 0, \ldots, n$. First we have

$$\pi_k(x) = x(x - [1])(x - [2]) \cdots (x - [k-1]),$$

for $k > 0$, with $\pi_0(x) = 1$. On writing $x = [t]$, we have

$$x - [j] = [t] - [j] = \frac{1 - q^t}{1 - q} - \frac{1 - q^j}{1 - q} = q^j[t - j],$$

and so

$$\pi_k([t]) = q^{k(k-1)/2}\,[t][t-1] \cdots [t-k+1]. \tag{1.114}$$

Thus, using (1.113) and (1.114), we have

$$\pi_k([t])\, f[x_0, \ldots, x_k] = q^{k(k-1)/2}\,[t][t-1] \cdots [t-k+1] \frac{\Delta_q^k f(x_0)}{q^{k(k-1)/2}\,[k]!}.$$

This may be expressed more simply in the form

$$\pi_k([t])\, f[x_0, \ldots, x_k] = \begin{bmatrix} t \\ k \end{bmatrix} \Delta_q^k f(x_0), \tag{1.115}$$

where

$$\begin{bmatrix} t \\ k \end{bmatrix} = \frac{[t]\,[t-1] \cdots [t-k+1]}{[k]!} \tag{1.116}$$

is a q-binomial coefficient, whose properties are discussed in Chapter 8. Thus, when each x_j equals $[j]$, the divided difference formula may be written as

$$p_n(x) = p_n([t]) = \sum_{k=0}^{n} \begin{bmatrix} t \\ k \end{bmatrix} \Delta_q^k f(x_0), \tag{1.117}$$

which we will call the q-difference form of the interpolating polynomial. We value (1.117) because it is a nice generalization of the forward difference formula (1.74), which we recover when we choose $q = 1$. However, if we wish to *evaluate* the interpolating polynomial when the abscissas are at the q-integers, we use the divided difference formula (1.19) rather than (1.117).

We will now state and verify an explicit expression for a kth q-difference $\Delta_q^k f(x_j)$ as a sum of multiples of values of f.

Theorem 1.5.2 For $q > 0$ and all $j, k \geq 0$,

$$\Delta_q^k f(x_j) = \sum_{r=0}^{k} (-1)^r q^{r(r-1)/2} \left[\begin{array}{c} k \\ r \end{array} \right] f(x_{j+k-r}), \tag{1.118}$$

where each x_j equals $[j]$.

Proof. This obviously holds for $k = 0$ and all j, when both sides of (1.118) reduce to $f(x_j)$. Let us assume that (1.118) holds for some integer $k \geq 0$ and all integers $j \geq 0$. We now begin with the recurrence relation (1.112), which we repeat here for convenience:

$$\Delta_q^{k+1} f(x_j) = \Delta_q^k f(x_{j+1}) - q^k \Delta_q^k f(x_j). \tag{1.119}$$

On replacing j by $j + 1$ in (1.118), the first term on the right of (1.119) may be written as

$$\Delta_q^k f(x_{j+1}) = \sum_{r=0}^{k} (-1)^r q^{r(r-1)/2} \left[\begin{array}{c} k \\ r \end{array} \right] f(x_{j+k+1-r}), \tag{1.120}$$

and on replacing r by $s - 1$ in (1.118) and then writing r in place of s, the term $\Delta_q^k f(x_j)$ on the right of (1.119) becomes

$$\Delta_q^k f(x_j) = \sum_{r=1}^{k+1} (-1)^{r-1} q^{(r-1)(r-2)/2} \left[\begin{array}{c} k \\ r-1 \end{array} \right] f(x_{j+k+1-r}). \tag{1.121}$$

It remains only to check from (1.119), (1.120), and (1.121) that the expansion of $\Delta_q^{k+1} f(x_j)$ as a sum of multiples of values of f agrees with (1.118) with k replaced by $k+1$. It is easy to check that the terms involving $f(x_j)$ and $f(x_{j+k+1})$ are correct, on putting $r = k + 1$ in (1.121) and $r = 0$ in (1.120), respectively. For $1 \leq r \leq k$ the expansion of $\Delta_q^{k+1} f(x_j)$ contains two contributions involving $f(x_{j+k+1-r})$. We combine these to give

$$(-1)^r q^{r(r-1)/2} \left(\left[\begin{array}{c} k \\ r \end{array} \right] + q^{k+1-r} \left[\begin{array}{c} k \\ r-1 \end{array} \right] \right) = (-1)^r q^{r(r-1)/2} \left[\begin{array}{c} k+1 \\ r \end{array} \right],$$

for $1 \leq r \leq k$, on using the Pascal-type relation (8.8). This verifies that (1.118) holds when k is replaced by $k + 1$, and completes the proof by induction. ■

As a corollary of the last theorem, on putting $q = 1$ in (1.118) we obtain a corresponding expansion for a kth-order forward difference, valid for $x_j = j$. The latter expansion is the special case of (1.82) where each x_j equals j.

The following expression for the kth q-difference of a product was given by Koçak and Phillips [30]:

$$\Delta_q^k \left(f(x_j)g(x_j)\right) = \sum_{r=0}^{k} \left[\begin{array}{c} k \\ r \end{array} \right] \Delta_q^r f(x_j)\, \Delta_q^{k-r} g(x_{j+r}), \tag{1.122}$$

where $x_j = [j]$. Observe that if we set $q = 1$, (1.122) reduces to (1.84), the summation we gave in Section 1.3 for the forward difference of a product of two functions, for the case where $x_j = j$. We now show how (1.122) may be deduced from (1.86), the expression given above for the kth divided difference of a product of two functions, using the relation (1.113), which connects q differences and divided differences. For it follows immediately from (1.86) and (1.113) that

$$\frac{\Delta_q^k (f(x_j)g(x_j))}{q^{k(2j+k-1)/2}\,[k]!} = \sum_{r=0}^{k} \frac{\Delta_q^r f(x_j)}{q^{r(2j+r-1)/2}\,[r]!}\, \frac{\Delta_q^{k-r} g(x_{j+r})}{q^{(k-r)(2j+k+r-1)/2}\,[k-r]!}.$$

Then, since

$$k(2j+k-1) = r(2j+r-1) + (k-r)(2j+k+r-1),$$

the powers of q in the denominators of the above equation cancel, and (1.122) follows on multiplying throughout by $[k]!$, since

$$\frac{[k]!}{[r]![k-r]!} = \left[\begin{array}{c} k \\ r \end{array} \right].$$

Problem 1.5.1 If $x = [t]$, show that

$$t = \log_q(1 - (1-q)x).$$

If $0 < q < 1$, verify that the above relation between t and x holds for $-\infty < x < 1/(1-q)$.

Problem 1.5.2 Show that for any fixed integer r such that $0 \le r \le k$,

$$\prod_{j \ne r} ([r] - [j]) = (-1)^{k-r} q^{r(2k-r-1)/2} [r]![k-r]!,$$

where the product is taken over all j from 0 to k, but excluding r. Hint: Split the product into two factors, one corresponding to the values of j such that $0 \le j < r$, and the other to the values of j such that $r < j \le k$.

Problem 1.5.3 Use (1.113) and (1.21) to show that

$$\Delta_q^k f(x_0) = q^{k(k-1)/2}\,[k]! \sum_{r=0}^{k} \frac{f(x_r)}{\prod_{j\neq r}([r]-[j])},$$

and use the result in Problem 1.5.2 to deduce that

$$\Delta_q^k f(x_0) = \sum_{r=0}^{k} (-1)^{k-r} q^{(k-r)(k-r-1)/2} \left[\begin{array}{c} k \\ r \end{array} \right] f(x_r).$$

By reversing the order of the latter summation, replacing r by $k-r$, show that

$$\Delta_q^k f(x_0) = \sum_{r=0}^{k} (-1)^{r} q^{r(r-1)/2} \left[\begin{array}{c} k \\ r \end{array} \right] f(x_{k-r}),$$

which is (1.118) with $j = 0$.

Problem 1.5.4 Evaluate the fundamental polynomial $L_i(x)$, defined by (1.9), when the interpolating points are given by $x_j = [j]$, for $0 \leq j \leq n$, and show that

$$L_i([t]) = (-1)^{n-i}\, \frac{q^{(n-i)(n-i+1)/2}}{[i]![n-i]!} \prod_{j\neq i} [t-j],$$

where the product is taken over all integers j from 0 to n, but excluding $j = i$.

Problem 1.5.5 Deduce from the result in Problem 1.5.4 that if we interpolate at the abscissas $0, \ldots, n$, the fundamental polynomial $L_i(x)$ is given by

$$L_i(x) = \frac{(-1)^{n-i}}{i!(n-i)!} \prod_{j\neq i} (x-j).$$

Problem 1.5.6 Verify the identity

$$\sum_{i=0}^{n} (-1)^{n-i} \left(\begin{array}{c} n \\ i \end{array} \right) \prod_{j\neq i} (x-j) = n!.$$

Problem 1.5.7 Derive a q-analogue of the identity in Problem 1.5.6.

2
Best Approximation

2.1 The Legendre Polynomials

Given a function f defined on $[-1, 1]$, let us write

$$\|f\| = \left(\int_{-1}^{1} [f(x)]^2 dx \right)^{1/2}. \tag{2.1}$$

We call $\|f\|$ the *square norm* of f. It can be thought of as a measure of the "size" of f. The reason for taking the square root in (2.1) is so that the norm satisfies the condition

$$\|\lambda f\| = |\lambda| \cdot \|f\|, \tag{2.2}$$

for all real λ. The square norm, which is analogous to the notion of length in n-dimensional Euclidean space, obviously satisfies the *positivity* condition

$$\|f\| > 0 \quad \text{unless } f(x) \equiv 0, \text{ the zero function, when } \|f\| = 0, \tag{2.3}$$

and, not so obviously, satisfies the *triangle inequality*

$$\|f + g\| \leq \|f\| + \|g\|, \tag{2.4}$$

for all f and g. It is very easy to check that properties (2.2) and (2.3) hold for the square norm (2.1). The third property (2.4) is a little more difficult to justify. It may be verified by expressing the integrals as limits of sums and then applying the result in Problem 2.1.1. The square norm is a special case of a general norm, which we now define.

Definition 2.1.1 A *norm* $\|\cdot\|$ on a given *linear space* S is a mapping from S to the real numbers that satisfies the three properties given by (2.2), (2.3), and (2.4). ■

Note that a linear space contains a zero element, and the norm of the zero element is zero. Two examples of linear spaces are the linear space of n-dimensional vectors, and the linear space of continuous functions defined on a finite interval, say $[-1, 1]$, which we denote by $C[-1, 1]$. In the latter case the three best known norms are the square norm, defined by (2.1), the *maximum norm*, defined by

$$\|f\| = \max_{-1 \le x \le 1} |f(x)|, \tag{2.5}$$

and the norm defined by

$$\|f\| = \int_{-1}^{1} |f(x)| dx. \tag{2.6}$$

The three properties (2.2), (2.3), and (2.4), called the *norm axioms*, are all easily verified for the norms defined by (2.5) and (2.6). The norms given by (2.1) and (2.6) are special cases of the p-norm, defined by

$$\|f\| = \left(\int_{-1}^{1} |f(x)|^p dx \right)^{1/p}, \tag{2.7}$$

for any $p \ge 1$, and the maximum norm (2.5) is obtained by letting $p \to \infty$ in (2.7). The restriction $p \ge 1$ is necessary so that the p-norm satisfies the triangle inequality, which follows by expressing the integrals as limits of sums and applying Minkowski's inequality,

$$\left(\sum_{j=1}^{n} |x_j + y_j|^p \right)^{1/p} \le \left(\sum_{j=1}^{n} |x_j|^p \right)^{1/p} + \left(\sum_{j=1}^{n} |y_j|^p \right)^{1/p}, \tag{2.8}$$

for $p \ge 1$. (See Davis [10] for a proof of (2.8).)

Example 2.1.1 Consider the linear space whose elements are the row vectors

$$\mathbf{x} = (x_1, x_2, \ldots, x_n),$$

where the x_j are all real, with the usual addition of vectors

$$\mathbf{x} + \mathbf{y} = (x_1 + y_1, x_2 + y_2, \ldots, x_n + y_n).$$

We also have multiplication by a scalar, defined by

$$\lambda \mathbf{x} = (\lambda x_1, \lambda x_2, \ldots, \lambda x_n),$$

where λ is any real number. Then

$$\|\mathbf{x}\| = \max_{1\le j\le n} |x_j|, \tag{2.9}$$

$$\|\mathbf{x}\| = |x_1| + \cdots + |x_n|, \tag{2.10}$$

$$\|\mathbf{x}\| = \left(x_1^2 + \cdots + x_n^2\right)^{1/2} \tag{2.11}$$

are all norms for this linear space. Except for the verification of the triangle inequality for the last norm (see Problem 2.1.1), it is easy to check that (2.9), (2.10), and (2.11) are indeed norms. (Corresponding to the zero function in (2.3) we have the zero vector, whose elements are all zero.) These vector norms are analogous to the norms (2.5), (2.6), and (2.1), respectively, given above for the linear space of functions defined on $[-1, 1]$. There is also the p-norm,

$$\|\mathbf{x}\| = \left(|x_1|^p + |x_2|^p + \cdots + |x_n|^p\right)^{1/p}, \tag{2.12}$$

for any $p \ge 1$, which is analogous to the norm defined by (2.7). It is easy to verify that the norm defined by (2.12) satisfies the properties (2.2) and (2.3), and we can apply Minkowski's inequality (2.8) to justify that it also satisfies the triangle inequality (2.4). If we put $p = 1$ and 2 in (2.12), we recover the norms (2.10) and (2.11), and if we let $p \to \infty$ in (2.12), we recover the norm (2.9). ■

Now, given f defined on $[-1, 1]$, let us seek the minimum value of $\|f - p\|$, for all $p \in P_n$, where $\|\cdot\|$ denotes the square norm. For *any* given norm, a $p \in P_n$ that minimizes $\|f - p\|$ is called a *best approximation* for f with respect to that norm. It can be shown (see Davis [10]) that for any norm and any $n \ge 0$, a best approximation always exists. Let us write

$$p(x) = \sum_{r=0}^{n} a_r q_r(x), \tag{2.13}$$

where $\{q_0, q_1, \ldots, q_n\}$ is some basis for P_n, so that any polynomial in P_n can be written as a sum of multiples of the q_r, as in (2.13). To find a best approximation, we can dispense with the square root in (2.1), since the problem of minimizing $\|f - p\|$ is equivalent to finding the minimum value of

$$\|f - p\|^2 = \int_{-1}^{1} [f(x) - p(x)]^2 dx = E(a_0, \ldots, a_n),$$

say, where p is given by (2.13). Thus we need to equate to zero the partial derivatives of E with respect to each a_s, and we obtain

$$0 = \frac{\partial}{\partial a_s} E(a_0, \ldots, a_n) = \int_{-1}^{1} 2[f(x) - p(x)] \cdot [-q_s(x)] dx,$$

for $0 \le s \le n$. This gives a system of linear equations to determine the coefficients a_s, and we will write these as

$$\sum_{r=0}^{n} c_{r,s} a_r = b_s, \quad 0 \le s \le n,$$

where

$$b_s = \int_{-1}^{1} f(x) q_s(x) dx, \quad 0 \le s \le n,$$

and

$$c_{r,s} = \int_{-1}^{1} q_r(x) q_s(x) dx, \quad 0 \le r, s \le n.$$

If we can now choose the basis $\{q_0, q_1, \ldots, q_n\}$ so that

$$\int_{-1}^{1} q_r(x) q_s(x) dx = 0, \quad r \ne s, \quad 0 \le r, s \le n, \tag{2.14}$$

then the above linear system will be immediately solved, giving

$$a_s = \int_{-1}^{1} f(x) q_s(x) dx \Big/ \int_{-1}^{1} [q_s(x)]^2 dx, \quad 0 \le s \le n. \tag{2.15}$$

Definition 2.1.2 The set of functions $\{q_0, q_1, \ldots, q_n\}$ is called an *orthogonal* basis if (2.14) holds, and if, in addition,

$$\int_{-1}^{1} [q_r(x)]^2 dx = 1, \quad 0 \le r \le n, \tag{2.16}$$

it is called an *orthonormal* basis. An orthogonal basis can obviously be made orthonormal by scaling each polynomial q_r appropriately. ■

We can construct the elements of an orthogonal basis, beginning with $q_0(x) = 1$ and choosing each q_k so that it satisfies the k conditions

$$\int_{-1}^{1} x^r q_k(x) dx = 0, \quad 0 \le r < k. \tag{2.17}$$

We then say that q_k is orthogonal on $[-1, 1]$ to all polynomials in P_{k-1}, and thus (2.14) holds. To determine each $q_k \in P_k$ uniquely, we will scale q_k so that its coefficient of x^k is unity. We say that q_r and q_s are mutually orthogonal if $r \ne s$. These orthogonal polynomials are named after A. M. Legendre (1752–1833). Let us now write

$$q_k(x) = x^k + d_{k-1} x^{k-1} + \cdots + d_1 x + d_0,$$

and solve a system of k linear equations, derived from (2.17), to obtain the coefficients d_0, $d_1, \ldots, d_{k-1}$. Beginning with $q_0(x) = 1$, we find that

$q_1(x) = x$, and $q_2(x) = x^2 - \frac{1}{3}$. At this stage, we will refer to any multiples of the polynomials q_k as Legendre polynomials, although we will restrict the use of this name later to the particular *multiples* of these polynomials that assume the value 1 at $x = 1$. The following theorem shows that the Legendre polynomials satisfy a simple recurrence relation, and in Theorem 2.1.2 we will show that the recurrence relation can be simplified still further.

Theorem 2.1.1 The Legendre polynomials, scaled so that their leading coefficients are unity, satisfy the recurrence relation

$$q_{n+1}(x) = (x - \alpha_n)q_n(x) - \beta_n q_{n-1}(x), \tag{2.18}$$

where

$$\alpha_n = \int_{-1}^{1} x[q_n(x)]^2 dx \Big/ \int_{-1}^{1} [q_n(x)]^2 dx, \tag{2.19}$$

and

$$\beta_n = \int_{-1}^{1} [q_n(x)]^2 dx \Big/ \int_{-1}^{1} [q_{n-1}(x)]^2 dx, \tag{2.20}$$

for all $n \geq 1$, where $q_0(x) = 1$ and $q_1(x) = x$.

Proof. It is clear that q_0 and q_1 are mutually orthogonal. To complete the proof it will suffice to show that for $n \geq 1$, if $q_0, q_1, \ldots, q_n$ denote the Legendre polynomials of degree up to n, each with leading coefficient unity, then the polynomial q_{n+1} defined by (2.18), with α_n and β_n defined by (2.19) and (2.20), respectively, is orthogonal to all polynomials in P_n. Now, since q_{n-1} and q_n are orthogonal to all polynomials in P_{n-2} and P_{n-1}, respectively, it follows from the recurrence relation (2.18) that q_{n+1} is orthogonal to all polynomials in P_{n-2}. For if we multiply (2.18) throughout by $q_m(x)$ and integrate over $[-1, 1]$, it is clear from the orthogonality property that

$$\int_{-1}^{1} q_{n+1}(x)q_m(x)dx = \int_{-1}^{1} xq_n(x)q_m(x)dx = 0,$$

for $0 \leq m \leq n - 2$, since we can write

$$xq_m(x) = q_{m+1}(x) + r_m(x),$$

where $r_m \in P_m$. The proof will be completed if we can show that q_{n+1} is orthogonal to q_{n-1} and q_n. If we multiply (2.18) throughout by $q_{n-1}(x)$, integrate over $[-1, 1]$, and use the fact that q_{n-1} and q_n are orthogonal, we obtain

$$\int_{-1}^{1} q_{n+1}(x)q_{n-1}(x)dx = \int_{-1}^{1} xq_n(x)q_{n-1}(x)dx - \beta_n \int_{-1}^{1} [q_{n-1}(x)]^2 dx.$$

Since $xq_{n-1}(x) = q_n(x) + r_{n-1}(x)$, where $r_{n-1} \in P_{n-1}$, it follows from the orthogonality property that

$$\int_{-1}^{1} xq_n(x)q_{n-1}(x)dx = \int_{-1}^{1} [q_n(x)]^2 dx.$$

From this and the above definition of β_n, it follows that q_{n+1} is orthogonal to q_{n-1}. If we now multiply (2.18) throughout by $q_n(x)$ and integrate over $[-1,1]$, it then follows from the orthogonality of q_{n-1} and q_n, and the definition of α_n, that q_{n+1} is orthogonal to q_n. ■

The following two theorems tell us more about the Legendre polynomials. First we show that the recurrence relation in (2.18) can be simplified.

Theorem 2.1.2 The Legendre polynomials, scaled so that their leading coefficients are unity, satisfy the recurrence relation

$$q_{n+1}(x) = xq_n(x) - \beta_n q_{n-1}(x), \tag{2.21}$$

for $n \geq 1$, with $q_0(x) = 1$ and $q_1(x) = x$, where β_n is defined by (2.20). Further, q_n is an even function when n is even, and is an odd function when n is odd.

Proof. We begin by noting that $q_0(x) = 1$ is an even function, and $q_1(x) = x$ is odd. Let us assume that for some $k \geq 0$, the polynomials $q_0, q_1, \ldots, q_{2k+1}$ are alternately even and odd, and that $\alpha_n = 0$ in (2.19) for all $n \leq 2k$. It then follows from (2.19) that $\alpha_{2k+1} = 0$, and then from the recurrence relation (2.18) that q_{2k+2} is an even function. Another inspection of (2.19) shows that $\alpha_{2k+2} = 0$, and then (2.18) shows that q_{2k+3} is an odd function. The proof is completed by induction. Later in this section we will determine the value of β_n. ■

Theorem 2.1.3 The Legendre polynomial of degree n has n distinct zeros in the interior of the interval $[-1,1]$.

Proof. Since q_n is orthogonal to 1 for $n > 0$,

$$\int_{-1}^{1} q_n(x)dx = 0, \quad n > 0,$$

and thus q_n must have at least one zero in the interior of $[-1,1]$. If $x = x_1$ were a multiple zero of q_n, for $n \geq 2$, then $q_n(x)/(x-x_1)^2$ would be a polynomial in P_{n-2} and so be orthogonal to q_n, giving

$$0 = \int_{-1}^{1} \frac{q_n(x)}{(x-x_1)^2}\, q_n(x)dx = \int_{-1}^{1} \left(\frac{q_n(x)}{x-x_1}\right)^2 dx,$$

which is impossible. Thus the zeros of q_n are all distinct. Now suppose that q_n has exactly $k \geq 1$ zeros in the interior of $[-1,1]$, and that

$$q_n(x) = (x-x_1)\cdots(x-x_k)\, r(x) = \pi_k(x)r(x),$$

say, where $r(x)$ does not change sign in $(-1,1)$. Then if $k < n$, it follows from the orthogonality property that

$$0 = \int_{-1}^{1} \pi_k(x)q_n(x)dx = \int_{-1}^{1} [\pi_k(x)]^2 r(x)dx,$$

which is impossible, since $[\pi_k(x)]^2 r(x)$ does not change sign. Thus we must have $k = n$, and consequently $r(x) = 1$, which completes the proof. ■

We will now obtain an explicit form for the recurrence relation (2.21) for the Legendre polynomials. Consider derivatives of the function

$$(x^2 - 1)^n = (x - 1)^n (x + 1)^n.$$

It is easily verified, using the Leibniz rule for differentiation (1.83), that

$$\frac{d^j}{dx^j}(x^2 - 1)^n = 0, \quad 0 \le j \le n - 1, \quad \text{for} \quad x = \pm 1, \tag{2.22}$$

and

$$\frac{d^n}{dx^n}(x^2 - 1)^n = 2^n n! \quad \text{for} \quad x = 1. \tag{2.23}$$

If we define

$$Q_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n}(x^2 - 1)^n, \tag{2.24}$$

it is clear that Q_n is a polynomial of degree n, and that $Q_n(1) = 1$. The relation (2.24) is called a Rodrigues formula, after O. Rodrigues. We now state and prove the following lemma.

Lemma 2.1.1 If u and v are both n times differentiable on $[-1, 1]$, and v and its first $n - 1$ derivatives are zero at both endpoints $x = \pm 1$, then

$$\int_{-1}^{1} u(x) v^{(n)}(x) dx = (-1)^n \int_{-1}^{1} u^{(n)}(x) v(x) dx, \quad n \ge 1. \tag{2.25}$$

Proof. Using integration by parts, we have

$$\int_{-1}^{1} u(x) v^{(n)}(x) dx = -\int_{-1}^{1} u'(x) v^{(n-1)}(x) dx,$$

and the proof is completed by using induction on n. ■

In particular, if g is any function that is n times differentiable, we obtain

$$\int_{-1}^{1} g(x) \frac{d^n}{dx^n}(x^2 - 1)^n dx = \int_{-1}^{1} g^{(n)}(x)(1 - x^2)^n dx, \tag{2.26}$$

and the latter integral is zero if $g \in P_{n-1}$. We deduce that the polynomial $Q_n \in P_n$ is orthogonal to all polynomials in P_{n-1}, and thus must be a multiple of the Legendre polynomial q_n. It can be shown that

$$|Q_n(x)| \le 1, \quad \text{for} \quad |x| \le 1, \tag{2.27}$$

the maximum modulus of 1 being attained at the endpoints $x = \pm 1$. (This inequality is derived in Rivlin [48], via a cleverly arranged sequence of

results involving Q_n and its first derivative.) To derive an explicit form of the recurrence relation (2.21), we write

$$(x^2-1)^n = x^{2n} - nx^{2n-2} + \cdots,$$

and so obtain from (2.24) that

$$Q_n(x) = \frac{1}{2^n n!}\left(\frac{(2n)!}{n!}x^n - \frac{n(2n-2)!}{(n-2)!}x^{n-2} + \cdots\right), \tag{2.28}$$

for $n \geq 2$. Since the Legendre polynomial q_n has leading term x^n, with coefficient unity, it follows from (2.28) that

$$Q_n(x) = \mu_n q_n(x), \quad \text{where} \quad \mu_n = \frac{1}{2^n}\begin{pmatrix} 2n \\ n \end{pmatrix}, \tag{2.29}$$

and

$$q_n(x) = x^n - \frac{n(n-1)}{2(2n-1)}x^{n-2} + \cdots. \tag{2.30}$$

If we now use (2.30) in the recurrence relation (2.21), we obtain

$$x^{n+1} - \frac{(n+1)n}{2(2n+1)}x^{n-1} + \cdots = x^{n+1} - \frac{n(n-1)}{2(2n-1)}x^{n-1} - \beta_n x^{n-1} + \cdots,$$

and on equating coefficients of x^{n-1}, we find that

$$\beta_n = \frac{(n+1)n}{2(2n+1)} - \frac{n(n-1)}{2(2n-1)} = \frac{n^2}{4n^2-1}. \tag{2.31}$$

It was convenient to work with q_n, which is scaled so that its coefficient of x^n is unity, in the early part of our discussion on the Legendre polynomials. This enabled us to simplify the recurrence relation given in Theorems 2.1.1 and 2.1.2. In order to refine the recurrence relation, we then introduced Q_n, the multiple of q_n scaled so that $Q_n(1) = 1$. In the mathematical literature it is Q_n that is called *the* Legendre polynomial. If we use (2.29) to replace each q_n in the recurrence relation (2.21) by the appropriate multiple of Q_n, and use (2.31), we obtain the following remarkably simple recurrence relation, which deserves to be expressed as a theorem.

Theorem 2.1.4 The Legendre polynomials Q_n satisfy the recurrence relation

$$(n+1)Q_{n+1}(x) = (2n+1)xQ_n(x) - nQ_{n-1}(x), \tag{2.32}$$

for $n \geq 1$, where $Q_0(x) = 1$ and $Q_1(x) = x$. ■

We find from the recurrence relation (2.32) and its initial conditions that the first few Legendre polynomials are

$$1, \quad x, \quad \frac{1}{2}(3x^2-1), \quad \frac{1}{2}(5x^3-3x), \quad \frac{1}{8}(35x^4-30x^2+3). \tag{2.33}$$

The Legendre polynomial $Q_n(x)$ (see Problem 2.1.7) satisfies the second-order differential equation

$$(1-x^2)Q_n''(x) - 2xQ_n'(x) + n(n+1)Q_n(x) = 0. \tag{2.34}$$

Let us return to a problem that we posed near the beginning of this section: Given a function f defined on $[-1,1]$, let us seek the polynomial $p_n \in P_n$ that minimizes $\|f - p_n\|$, where $\|\cdot\|$ is the square norm. Our solution to this problem was given in terms of the polynomials q_r, and if we recast this in terms of the Legendre polynomials proper, we obtain the solution

$$p_n(x) = \sum_{r=0}^{n} a_r Q_r(x), \tag{2.35}$$

where

$$a_r = \int_{-1}^{1} f(x)Q_r(x)dx \Big/ \int_{-1}^{1} [Q_r(x)]^2 dx, \quad 0 \le r \le n. \tag{2.36}$$

We call the polynomial p_n the best square norm approximation or, more commonly, the *least squares* approximation for f on $[-1,1]$. We remark, in passing, that the partial sum of a Fourier series has this same property of being a least squares approximation. If we let $n \to \infty$ in (2.35), the resulting infinite series, if it exists, is called the *Legendre series* for f. From (2.36) and Problem 2.1.6, the Legendre coefficients may be expressed in the form

$$a_r = \frac{1}{2}(2r+1)\int_{-1}^{1} f(x)Q_r(x)dx, \quad r \ge 0, \tag{2.37}$$

and we have the following result.

Theorem 2.1.5 The partial sum of the Legendre series for f is even or odd if f is even or odd, respectively.

Proof. As we saw in Theorem 2.1.2, the polynomial q_n is an even function when n is even, and is an odd function when n is odd, and (2.29) shows that this holds also for its multiple, the Legendre polynomial Q_n. It follows from (2.37) that the Legendre coefficient a_r is zero if r is odd and f is even, and is also zero if r is even and f is odd. Thus the Legendre series for f contains only even- or odd-order Legendre polynomials when f is even or odd, respectively. This completes the proof. ■

If f is sufficiently differentiable, we can derive another expression for the Legendre coefficients by expressing $Q_s(x)$ in (2.37) in its Rodrigues form, given in (2.24), and then use (2.26) to give

$$a_r = \frac{2r+1}{2^{r+1}r!}\int_{-1}^{1} f^{(r)}(x)(1-x^2)^r dx. \tag{2.38}$$

We now give an estimate of the size of the Legendre coefficient a_r for f in terms of $f^{(r)}$.

Theorem 2.1.6 If $f^{(r)}$ is continuous on $[-1,1]$, the Legendre coefficient a_r is given by

$$a_r = \frac{2^r\, r!}{(2r)!}\, f^{(r)}(\xi_r), \tag{2.39}$$

where $\xi_r \in (-1,1)$.

Proof. Since $f^{(r)}$ is continuous on $[-1,1]$, it follows from the mean value theorem for integrals, Theorem 3.1.2, that

$$a_r = \frac{2r+1}{2^{r+1}r!} f^{(r)}(\xi_r) I_r,$$

where

$$I_r = \int_{-1}^{1} (1-x^2)^r dx.$$

Then, using integration by parts, we find that

$$I_r = 2r \int_{-1}^{1} x^2 (1-x^2)^{r-1} dx = 2r(I_{r-1} - I_r).$$

Hence

$$I_r = \frac{2r}{2r+1} I_{r-1} = \frac{2r}{2r+1}\frac{2r-2}{2r-1}\cdots\frac{2}{3} I_0,$$

where $I_0 = 2$, and (2.39) follows easily. ■

Example 2.1.2 Let us use (2.38) to compute the Legendre coefficients for the function e^x. First we derive from (2.38)

$$a_0 = \frac{1}{2}(e - e^{-1}) \approx 1.175201, \qquad a_1 = 3e^{-1} \approx 1.103638.$$

If we write

$$J_r = \frac{1}{2^r\, r!} \int_{-1}^{1} e^x (1-x^2)^r dx, \tag{2.40}$$

then (2.38) with $f(x) = e^x$ yields $a_r = \frac{1}{2}(2r+1)J_r$. On using integration by parts twice on the above integral for J_r, we obtain

$$J_r = -(2r-1)J_{r-1} + J_{r-2},$$

and hence obtain the recurrence relation

$$\frac{a_r}{2r+1} = -a_{r-1} + \frac{a_{r-2}}{2r-3}, \tag{2.41}$$

with a_0 and a_1 as given above. The next few Legendre coefficients for e^x, rounded to six decimal places, are as follows:

n	2	3	4	5	6	7
a_n	0.357814	0.070456	0.009965	0.001100	0.000099	0.000008

All the Legendre coefficients for e^x are positive. (See Problem 2.1.10.) Before leaving this example, we remark that the recurrence relation (2.41) is numerically unstable, since the error in the computed value of a_r is approximately $2r+1$ times the error in a_{r-1}. Thus, if we need to compute a_r for a large value of r, we should estimate it directly from the expression (2.40) for J_r, using a numerical integration method. ■

It is convenient to write

$$(f,g) = \int_{-1}^{1} f(x)g(x)dx, \tag{2.42}$$

and we call (f,g) an *inner product* of f and g. It follows from (2.35), the orthogonality of the Legendre polynomials Q_s, and (2.36) that

$$(p_n, Q_s) = a_s(Q_s, Q_s) = (f, Q_s). \tag{2.43}$$

Since

$$(f - p_n, Q_s) = (f, Q_s) - (p_n, Q_s),$$

we see from (2.43) that

$$(f - p_n, Q_s) = 0, \quad 0 \le s \le n. \tag{2.44}$$

For further material on inner products, see Davis [10], Deutsch [14]. The following theorem describes another property concerning the difference between f and p_n.

Theorem 2.1.7 The error term $f - p_n$ changes sign on at least $n+1$ points in the interior of $[-1,1]$, where p_n is the partial sum of the Legendre series for f.

Proof. First we have from (2.44) that $(f - p_n, Q_0) = 0$. Since $Q_0(x) = 1$, it follows that there is at least one point in $(-1,1)$ where $f - p_n$ changes sign. Suppose that $f(x) - p_n(x)$ changes sign at k points,

$$-1 < x_1 < x_2 < \cdots < x_k < 1,$$

and at no other points in $(-1,1)$, with $1 \le k < n+1$. Then $f(x) - p_n(x)$ and the function $\pi_k(x) = (x - x_1)\cdots(x - x_k)$ change sign at the x_j, and at no other points in $(-1,1)$. Thus we must have $(f - p_n, \pi_k) \neq 0$. On the other hand, since π_k may be written as a sum of multiples of $Q_0, Q_1, \ldots, Q_k$, it follows from (2.44) that $(f - p_n, \pi_k) = 0$, which gives a contradiction. We deduce that $k \ge n+1$, which completes the proof. ■

Since, as we have just established, there are at least $n+1$ points in $(-1,1)$ where $p_n(x)$ and $f(x)$ are equal, the best approximant p_n must be an interpolating polynomial for f. This observation leads us to the following interesting expression for $\|f - p_n\|$, of a similar form to the error term for the Taylor polynomial or the error term for the interpolating polynomial, which are given in (1.35) and (1.25), respectively.

Theorem 2.1.8 Let p_n denote the least squares approximation for f on $[-1,1]$. Then if $f^{(n+1)}$ is continuous on $[-1,1]$ there exists a number ζ in $(-1,1)$ such that

$$\|f-p_n\| = \frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}}\,\frac{|f^{(n+1)}(\zeta)|}{(n+1)!} \sim \frac{\sqrt{\pi}}{2^{n+1}}\,\frac{|f^{(n+1)}(\zeta)|}{(n+1)!}, \tag{2.45}$$

where μ_{n+1} is defined in (2.29).

Proof. We begin with a comment on the notation used in (2.45). We write $u_n \sim v_n$ to mean that

$$\lim_{n\to\infty}\frac{u_n}{v_n} = 1,$$

and we say that u_n and v_n are asymptotically equal, as $n \to \infty$.

Since the best approximant p_n is an interpolating polynomial for f, we have from (1.25) that

$$f(x) - p_n(x) = (x-x_0)(x-x_1)\cdots(x-x_n)\frac{f^{n+1}(\xi_x)}{(n+1)!},$$

where $\xi_x \in (-1,1)$, and the x_j are distinct points in $(-1,1)$. Taking the square norm, and applying the mean value theorem for integrals, Theorem 3.1.2, we find that

$$\|f-p_n\| = \frac{|f^{(n+1)}(\xi)|}{(n+1)!}\,\|(x-x_0)\cdots(x-x_n)\|, \tag{2.46}$$

for some $\xi \in (-1,1)$. If p_n^* denotes the interpolating polynomial for f on the zeros of the Legendre polynomial Q_{n+1}, we similarly have

$$\|f-p_n^*\| = \frac{|f^{(n+1)}(\eta)|}{(n+1)!}\,\|(x-x_0^*)\cdots(x-x_n^*)\|, \tag{2.47}$$

where $\eta \in (-1,1)$, and the x_j^* are the zeros of Q_{n+1}. We then see from (2.47) and Problem 2.1.9 that

$$\|f-p_n^*\| = \frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}}\,\frac{|f^{(n+1)}(\eta)|}{(n+1)!},$$

where μ_{n+1} is defined in (2.29). Since p_n is the least squares approximation for f, we have

$$\|f-p_n\| \le \|f-p_n^*\| = \frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}}\,\frac{|f^{(n+1)}(\eta)|}{(n+1)!}, \tag{2.48}$$

where $\eta \in (-1,1)$. Also, we may deduce from (2.46) and Problem 2.1.9 that

$$\|f-p_n\| \ge \frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}}\,\frac{|f^{(n+1)}(\xi)|}{(n+1)!}, \tag{2.49}$$

where $\xi \in (-1,1)$. We can now combine (2.48) and (2.49), and use the continuity of $f^{(n+1)}$ to give

$$\|f - p_n\| = \frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}}\,\frac{|f^{(n+1)}(\zeta)|}{(n+1)!}, \quad \zeta \in (-1,1),$$

and we complete the proof by applying Stirling's formula (see Problem 2.1.12) to give

$$\frac{1}{\mu_{n+1}}\sqrt{\frac{2}{2n+3}} \sim \frac{\sqrt{\pi}}{2^n}. \qquad \blacksquare$$

Problem 2.1.1 Let

$$\|x\| = \left(x_1^2 + \cdots + x_n^2\right)^{1/2}.$$

Verify the inequality

$$(x_1y_1 + \cdots + x_ny_n)^2 \le (x_1^2 + \cdots + x_n^2)(y_1^2 + \cdots + y_n^2)$$

by showing that it is equivalent to

$$\sum_{i\neq j}(x_iy_j - x_jy_i)^2 \ge 0,$$

where the latter sum of $\frac{1}{2}n(n-1)$ terms is taken over all distinct pairs of numbers i and j chosen from the set $\{1,\ldots,n\}$. Deduce the inequality

$$|x_1y_1 + \cdots + x_ny_n| \le \|x\| \cdot \|y\|,$$

and hence justify the triangle inequality for this norm.

Problem 2.1.2 Define

$$f(x) = \begin{cases} 0, & -1 \le x < 0, \\ x, & 0 \le x \le 1, \end{cases}$$

and define $g(x) = f(-x),\ -1 \le x \le 1$. Show that

$$\|f\| = \|g\| = \frac{1}{(p+1)^{1/p}} \quad \text{and} \quad \|f+g\| = \frac{2^{1/p}}{(p+1)^{1/p}},$$

where $\|\cdot\|$ denotes the p-norm, defined by (2.7). Deduce that $p \ge 1$ is a necessary condition for the triangle inequality to hold.

Problem 2.1.3 Scale the Legendre polynomials 1, x, $\frac{1}{2}(3x^2-1)$, and $\frac{1}{2}(5x^3-3x)$ so that they form an orthonormal basis for P_3.

Problem 2.1.4 Verify (2.22) and (2.23).

Problem 2.1.5 By writing down the binomial expansion of $(x^2-1)^n$ and differentiating it n times, deduce from (2.24) that

$$Q_n(x) = \frac{1}{2^n}\sum_{j=0}^{[n/2]}(-1)^j \begin{pmatrix} n \\ j \end{pmatrix} \begin{pmatrix} 2n-2j \\ n \end{pmatrix} x^{n-2j},$$

where $[n/2]$ denotes the integer part of $n/2$.

Problem 2.1.6 We have, from (2.20) and (2.31),

$$\beta_n = \int_{-1}^{1} [q_n(x)]^2\,dx \Big/ \int_{-1}^{1} [q_{n-1}(x)]^2\,dx = \frac{n^2}{4n^2-1},$$

for $n \geq 1$. Use this and (2.29) to deduce that

$$\int_{-1}^{1} [Q_n(x)]^2\,dx \Big/ \int_{-1}^{1} [Q_{n-1}(x)]^2\,dx = \frac{2n-1}{2n+1},$$

for $n \geq 1$, and hence show that

$$\int_{-1}^{1} [Q_n(x)]^2\,dx = \frac{2}{2n+1}$$

for $n \geq 0$.

Problem 2.1.7 If $y(x)$ and $p(x)$ are functions that are twice differentiable, use integration by parts twice to show that

$$\begin{aligned}\int_{-1}^{1} \frac{d}{dx}\{(1-x^2)y'(x)\}\,p(x)dx &= -\int_{-1}^{1}(1-x^2)\,y'(x)p'(x)dx \\ &= \int_{-1}^{1}\frac{d}{dx}\{(1-x^2)p'(x)\}\,y(x)dx.\end{aligned}$$

Now write $y(x) = Q_n(x)$ and let $p \in P_{n-1}$, and deduce that

$$\frac{d}{dx}\{(1-x^2)Q_n'(x)\} = (1-x^2)Q_n''(x) - 2xQ_n'(x)$$

is orthogonal to all polynomials in P_{n-1}, and thus must be a multiple of the Legendre polynomial Q_n. Using (2.28), equate coefficients of x^n in

$$(1-x^2)Q_n''(x) - 2xQ_n'(x) = \mu Q_n(x),$$

to give

$$-n(n-1) - 2n = \mu,$$

and thus show that Q_n satisfies the second-order differential equation given in (2.34).

Problem 2.1.8 Show that the polynomial $p \in P_n$ that minimizes

$$\int_{-1}^{1} [x^{n+1} - p(x)]^2 dx$$

is $p(x) = x^{n+1} - q_{n+1}(x)$, where q_{n+1} is the Legendre polynomial Q_{n+1} scaled so that it has leading coefficient unity. Hint: Write

$$x^{n+1} - p(x) = q_{n+1}(x) + \sum_{j=0}^{n} \gamma_j q_j(x),$$

and use the orthogonality properties of the q_j.

Problem 2.1.9 Deduce from the result in the previous problem that for the square norm on $[-1, 1]$, the minimum value of $\|(x - x_0) \cdots (x - x_n)\|$ is attained when the x_j are the zeros of the Legendre polynomial Q_{n+1}, and use the result in Problem 2.1.6 to show that

$$\min_{x_j} \|(x - x_0) \cdots (x - x_n)\| = \frac{1}{\mu_{n+1}} \sqrt{\frac{2}{2n+3}},$$

where μ_{n+1} is defined in (2.29).

Problem 2.1.10 Deduce from (2.38) that if f and all its derivatives are nonnegative on $[-1, 1]$, then all coefficients of the Legendre series for f are nonnegative.

Problem 2.1.11 Verify that

$$\frac{d}{dx}(x^2 - 1)^n = 2nx(x^2 - 1)^{n-1},$$

differentiate n times, and use the Leibniz rule (1.83) to show that

$$\frac{d^{n+1}}{dx^{n+1}}(x^2 - 1)^n = 2n \left(x \frac{d^n}{dx^n}(x^2 - 1)^{n-1} + n \frac{d^{n-1}}{dx^{n-1}}(x^2 - 1)^{n-1} \right),$$

for $n \geq 1$. Finally, divide by $2^n n!$ and use the Rodrigues formula (2.24) to obtain the relation

$$Q_n'(x) = xQ_{n-1}'(x) + nQ_{n-1}(x), \quad n \geq 1.$$

Problem 2.1.12 Use Stirling's formula,

$$n! \sim \sqrt{2\pi n} \left(\frac{n}{e} \right)^n,$$

to show that (see (2.29))

$$\mu_n = \frac{1}{2^n} \begin{pmatrix} 2n \\ n \end{pmatrix} \sim \frac{2^n}{\sqrt{\pi n}}.$$

2.2 The Chebyshev Polynomials

Many of the ideas and formulas presented in the last section concerning the Legendre polynomials can be generalized by introducing a *weight function*. This generalization leads to an infinite number of systems of orthogonal polynomials. We will pay particular attention to one such system, the Chebyshev polynomials, named after P. L. Chebyshev (1821–1894).

Given any integrable function ω that is nonnegative and not identically zero on $[-1, 1]$, we can construct a sequence of polynomials (q_n^ω), where q_n^ω is of degree n, has leading coefficient unity, and satisfies

$$\int_{-1}^{1} \omega(x) x^r q_n^\omega(x) dx = 0, \qquad 0 \le r < n. \tag{2.50}$$

The polynomials q_n^ω are said to be orthogonal on $[-1, 1]$ with respect to the weight function ω, and the scaled Legendre polynomials q_n are recovered by putting $\omega(x) = 1$. The generalized orthogonal polynomials q_n^ω, like the Legendre polynomials, satisfy a recurrence relation of the form (2.18), where the coefficients α_n and β_n are given by

$$\alpha_n = \int_{-1}^{1} \omega(x) x [q_n^\omega(x)]^2 dx \Big/ \int_{-1}^{1} \omega(x) [q_n^\omega(x)]^2 dx \tag{2.51}$$

and

$$\beta_n = \int_{-1}^{1} \omega(x) [q_n^\omega(x)]^2 dx \Big/ \int_{-1}^{1} \omega(x) [q_{n-1}^\omega(x)]^2 dx. \tag{2.52}$$

Further, if the weight function ω is even, then $\alpha_n = 0$, the even-order orthogonal polynomials are even functions, and the odd-order polynomials are odd, as given by Theorem 2.1.2 for the Legendre polynomials. The above statements about the generalized orthogonal polynomials are easily verified by inserting the weight function ω appropriately and repeating the arguments used above in the special case where $\omega(x) = 1$. See, for example, Davis and Rabinowitz [11].

Let us now define

$$\|f\| = \left(\int_{-1}^{1} \omega(x) [f(x)]^2 dx \right)^{1/2}. \tag{2.53}$$

It can be shown that $\|\cdot\|$ in (2.53) satisfies the three properties (2.2), (2.3), and (2.4), and so indeed defines a norm, which we call a weighted square norm. When we choose $\omega(x) = 1$, we recover the square norm (2.1). We then find that

$$\|f - p\| = \left(\int_{-1}^{1} \omega(x) [f(x) - p(x)]^2 dx \right)^{1/2}$$

is minimized over all $p \in P_n$ by choosing $p = p_n$, where

$$p_n(x) = \sum_{r=0}^{n} a_r q_r^{\omega}(x), \tag{2.54}$$

and each coefficient a_r is given by

$$a_r = \int_{-1}^{1} \omega(x) f(x) q_r^{\omega}(x) dx \Big/ \int_{-1}^{1} \omega(x) [q_r^{\omega}(x)]^2 dx, \quad 0 \le r \le n. \tag{2.55}$$

If we let $n \to \infty$, we obtain a generalized orthogonal expansion for f whose coefficients are given by (2.55). In particular, if $\omega(x) = 1$, we obtain the Legendre series. We can easily adapt Theorem 2.1.5 to show that if the weight function ω is even, then the partial sum of the generalized orthogonal series for f is even or odd if f is even or odd, respectively.

The choice of weight function

$$\omega(x) = (1-x)^{\alpha}(1+x)^{\beta}, \quad \alpha, \beta > -1, \tag{2.56}$$

leads to a two-parameter system of orthogonal polynomials that are called the Jacobi polynomials. These include the following, as special cases:

$\alpha = \beta = 0$	Legendre polynomials,
$\alpha = \beta = -\frac{1}{2}$	Chebyshev polynomials,
$\alpha = \beta = \frac{1}{2}$	Chebyshev polynomials of the second kind,
$\alpha = \beta$	ultraspherical polynomials.

Note that the first three systems of orthogonal polynomials listed above are all special cases of the fourth system, the *ultraspherical* polynomials, whose weight function is the even function $(1-x^2)^{\alpha}$. Thus the ultraspherical polynomials satisfy a recurrence relation of the form given in Theorem 2.1.2 for the Legendre polynomials. Further, the ultraspherical polynomial of degree n is even or odd, when n is even or odd, respectively. If we choose ω as the Jacobi weight function (2.56), then (2.54) and (2.55) define a partial Jacobi series, and it is clear that when $\alpha = \beta$, the resulting partial ultraspherical series is even or odd if f is even or odd, respectively.

To investigate the Jacobi polynomials, we begin with the function

$$(1-x)^{-\alpha}(1+x)^{-\beta} \frac{d^n}{dx^n}(1-x)^{n+\alpha}(1+x)^{n+\beta}. \tag{2.57}$$

First, using the Leibniz rule (1.83) for differentiating a product, we can show that this function is a polynomial of degree n. Then, following the method we used for the Legendre polynomials, beginning with Lemma

2.1.1, we can show that this polynomial is orthogonal on $[-1,1]$, with respect to the Jacobi weight function given in (2.56), to all polynomials in P_{n-1}. We then define the Jacobi polynomial of degree n as

$$Q_n^{(\alpha,\beta)}(x) = \frac{(-1)^n}{2^n n!}(1-x)^{-\alpha}(1+x)^{-\beta}\frac{d^n}{dx^n}(1-x)^{n+\alpha}(1+x)^{n+\beta}. \quad (2.58)$$

It is easy to adapt the proof of Theorem 2.1.3 to show that the Jacobi polynomial of degree n has n distinct zeros in the interior of $[-1,1]$.

The following theorem generalizes Theorem 2.1.4.

Theorem 2.2.1 The Jacobi polynomials satisfy the recurrence relation

$$Q_{n+1}^{(\alpha,\beta)}(x) = (a_n x + b_n)Q_n^{(\alpha,\beta)}(x) - c_n Q_{n-1}^{(\alpha,\beta)}(x), \quad (2.59)$$

where

$$a_n = \frac{(2n+\alpha+\beta+1)(2n+\alpha+\beta+2)}{2(n+1)(n+\alpha+\beta+1)},$$
$$b_n = \frac{(\alpha^2-\beta^2)(2n+\alpha+\beta+1)}{2(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)},$$
$$c_n = \frac{(n+\alpha)(n+\beta)(2n+\alpha+\beta+2)}{(n+1)(n+\alpha+\beta+1)(2n+\alpha+\beta)}.$$

Proof. This can be justified by following the same method as we used to prove Theorem 2.1.4. See Davis [10]. ■

We can use the Leibniz rule (1.83) again to show that

$$Q_n^{(\alpha,\beta)}(1) = \frac{1}{n!}(n+\alpha)(n-1+\alpha)\cdots(1+\alpha) = \begin{pmatrix} n+\alpha \\ n \end{pmatrix}, \quad (2.60)$$

which is independent of β. Likewise (see Problem 2.2.2), we can show that $Q_n^{(\alpha,\beta)}(-1)$ is independent of α. We can express (2.60) in the form

$$Q_n^{(\alpha,\beta)}(1) = \frac{\Gamma(n+\alpha+1)}{\Gamma(n+1)\Gamma(\alpha+1)}, \quad (2.61)$$

where Γ is the gamma function, defined by

$$\Gamma(x) = \int_0^\infty t^{x-1}e^{-t}dt. \quad (2.62)$$

It is not hard (see Problem 2.2.1) to deduce from (2.62) that the gamma function satisfies the difference equation

$$\Gamma(x+1) = x\Gamma(x), \quad (2.63)$$

with $\Gamma(1) = 1$, and hence justify (2.61).

From (2.55) the Jacobi series has coefficients that satisfy

$$a_r = \int_{-1}^{1} \omega(x) f(x) Q_r^{(\alpha,\beta)}(x) dx \Big/ \int_{-1}^{1} \omega(x) \left[Q_r^{(\alpha,\beta)}(x) \right]^2 dx, \tag{2.64}$$

where $\omega(x) = (1-x)^\alpha (1+x)^\beta$. Then, if f is sufficiently differentiable, it follows from (2.58) and Lemma 2.1.1 that the numerator on the right of (2.64) may be written as

$$\int_{-1}^{1} \omega(x) f(x) Q_r^{(\alpha,\beta)}(x) dx = \frac{1}{2^r\, r!} \int_{-1}^{1} \omega(x) f^{(r)}(x) (1-x^2)^r dx, \tag{2.65}$$

where $\omega(x) = (1-x)^\alpha (1+x)^\beta$. The denominator on the right of (2.64) can be expressed (see Davis [10]) in the form

$$\int_{-1}^{1} \omega(x) [Q_r^{(\alpha,\beta)}(x)]^2 dx = \frac{2^{\alpha+\beta+1}}{(2r+\alpha+\beta+1)} \frac{\Gamma(r+\alpha+1)\Gamma(r+\beta+1)}{\Gamma(r+1)\Gamma(r+\alpha+\beta+1)},$$

and we note that this is consistent with our findings in Problem 2.1.6 for the special case of the Legendre polynomials.

One might argue that the simplest of all the Jacobi polynomials are the Legendre polynomials, since these have the simplest weight function. However, there are good reasons for saying that the simplest Jacobi polynomials are the Chebyshev polynomials, which have weight function $(1-x^2)^{-1/2}$. Note that the latter weight function is singular at the endpoints $x = \pm 1$. The Chebyshev polynomials are usually denoted by $T_n(x)$, and are uniquely defined by

$$\int_{-1}^{1} (1-x^2)^{-1/2} T_r(x) T_s(x) dx = 0, \quad r \neq s, \tag{2.66}$$

where

$$T_r \in P_r \quad \text{and} \quad T_r(1) = 1, \quad r \geq 0. \tag{2.67}$$

It follows from the above definition of the Chebyshev polynomials that T_n is a multiple of the Jacobi (also ultraspherical) polynomial $Q_n^{(-1/2,-1/2)}$. With $\alpha = \beta = -\frac{1}{2}$ in (2.60), we can show that

$$Q_n^{(-1/2,-1/2)}(1) = \frac{1}{2^{2n}} \begin{pmatrix} 2n \\ n \end{pmatrix} \sim \frac{1}{\sqrt{\pi n}}, \tag{2.68}$$

and, since $T_n(1) = 1$, it follows that

$$Q_n^{(-1/2,-1/2)}(x) = \frac{1}{2^{2n}} \begin{pmatrix} 2n \\ n \end{pmatrix} T_n(x). \tag{2.69}$$

The following theorem shows that the Chebyshev polynomials can be expressed in a very simple form in terms of the cosine function.

Theorem 2.2.2 Let us define

$$t_n(x) = \cos n\theta, \quad \text{where } x = \cos\theta, \quad -1 \le x \le 1, \quad n \ge 0. \tag{2.70}$$

Then $t_n = T_n$, the Chebyshev polynomial of degree n.

Proof. On making the substitution $x = \cos\theta$, the interval $-1 \le x \le 1$ corresponds to $0 \le \theta \le \pi$, and we obtain

$$\int_{-1}^{1} (1-x^2)^{-1/2} t_r(x) t_s(x) dx = \int_0^{\pi} \cos r\theta \cos s\theta \, d\theta, \quad r \ne s.$$

Since

$$\cos r\theta \cos s\theta = \frac{1}{2}(\cos(r+s)\theta + \cos(r-s)\theta),$$

we readily see that

$$\int_{-1}^{1} (1-x^2)^{-1/2} t_r(x) t_s(x) dx = 0, \quad r \ne s.$$

From the substitution $x = \cos\theta$, we see that $x = 1$ corresponds to $\theta = 0$, and thus $t_r(1) = 1$. Now

$$\cos(n+1)\theta + \cos(n-1)\theta = 2\cos n\theta \cos\theta,$$

which, using (2.70), yields the recurrence relation

$$t_{n+1}(x) = 2x\, t_n(x) - t_{n-1}(x), \quad n \ge 1,$$

and we see from (2.70) that $t_0(x) = 1$ and $t_1(x) = x$. Thus, by induction, each t_r belongs to P_r. It follows that $t_r = T_r$ for all $r \ge 0$, and this completes the proof. ■

We have just shown that

$$T_n(x) = \cos n\theta, \quad \text{where } x = \cos\theta, \quad -1 \le x \le 1, \quad n \ge 0, \tag{2.71}$$

and that the Chebyshev polynomials satisfy the recurrence relation

$$T_{n+1}(x) = 2xT_n(x) - T_{n-1}(x), \quad n \ge 1, \tag{2.72}$$

with $T_0(x) = 1$ and $T_1(x) = x$. We find that the first few Chebyshev polynomials are

$$1, \quad x, \quad 2x^2 - 1, \quad 4x^3 - 3x, \quad 8x^4 - 8x^2 + 1. \tag{2.73}$$

Given how simply the Chebyshev polynomial is expressed in (2.71), we can very easily find its zeros and turning values. We write

$$T_n(x) = 0 \quad \Rightarrow \quad \cos n\theta = 0 \quad \Rightarrow \quad n\theta = (2j-1)\frac{\pi}{2}, \tag{2.74}$$

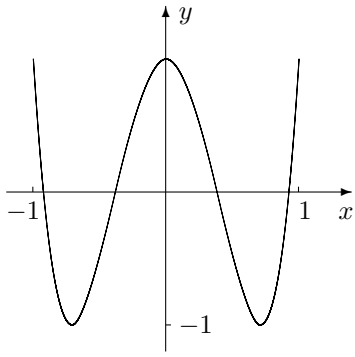

FIGURE 2.1. Graph of the Chebyshev polynomial $T_4(x)$.

where j is an integer. Thus

$$T_n(x) = 0 \quad \Rightarrow \quad x = \cos\theta_j, \qquad \text{where} \quad \theta_j = \frac{(2j-1)\pi}{2n}, \tag{2.75}$$

for some integer j. We see from the graph of $x = \cos\theta$ that as θ takes all values between 0 and π, the function $x = \cos\theta$ is monotonic decreasing, and takes all values between 1 and -1. The choice of $j = 1, 2, \ldots, n$ in (2.75) gives n distinct zeros of T_n in $(-1, 1)$, and since $T_n \in P_n$, *all* the zeros of T_n are given by the n values

$$x_j = \cos\frac{(2j-1)\pi}{2n}, \quad 1 \le j \le n. \tag{2.76}$$

Since for $x \in [-1, 1]$ we can express $T_n(x)$ in the form $\cos n\theta$, where $x = \cos\theta$, it is clear that the maximum modulus of T_n on $[-1, 1]$ is 1. This is attained for values of θ such that $|\cos n\theta| = 1$, and

$$\cos n\theta = \pm 1 \quad \Rightarrow \quad n\theta = j\pi \quad \Rightarrow \quad x = \cos(j\pi/n) = \tau_j,$$

say, where j is an integer. Thus the Chebyshev polynomial T_n attains its maximum modulus of 1 at the $n + 1$ points $\tau_j = \cos(j\pi/n)$, for $0 \le j \le n$, and T_n alternates in sign over this set of points. For we have

$$T_n(\tau_j) = \cos j\pi = (-1)^j, \quad 0 \le j \le n.$$

These $n + 1$ points of maximum modulus are called the *extreme points* of T_n. The Chebyshev polynomial T_n is the only polynomial in P_n whose maximum modulus is attained on $n + 1$ points of $[-1, 1]$. Note that although we have expressed $T_n(x)$ in terms of the cosine function for the interval $-1 \le x \le 1$ only, the Chebyshev polynomial is defined by its recurrence relation (2.72) for all real x. Outside the interval $[-1, 1]$, we can

(see Problem 2.2.6) express $T_n(x)$ in terms of the hyperbolic cosine. For further information on the Chebyshev polynomials, see Rivlin [49].

Let us consider the Chebyshev series, the orthogonal series based on the Chebyshev polynomials. From (2.55) we see that the Chebyshev coefficients are determined by the ratio of two integrals, and (see Problem 2.2.12) the integral in the denominator is

$$\int_{-1}^{1}(1-x^2)^{-1/2}[T_r(x)]^2dx = \begin{cases} \pi, & r=0, \\ \frac{1}{2}\pi, & r>0. \end{cases} \tag{2.77}$$

Thus the infinite Chebyshev series for f is

$$\frac{1}{2}a_0 + \sum_{r=1}^{\infty} a_r T_r(x), \tag{2.78}$$

where the first coefficient is halved so that the relation

$$a_r = \frac{2}{\pi}\int_{-1}^{1}(1-x^2)^{-1/2}f(x)T_r(x)dx \tag{2.79}$$

holds for all $r \geq 0$. By making the substitution $x = \cos\theta$, we can express (2.79) alternatively in the form

$$a_r = \frac{2}{\pi}\int_0^{\pi} f(\cos\theta)\cos r\theta\, d\theta, \quad r \geq 0. \tag{2.80}$$

We will write

$$f(x) \sim \frac{1}{2}a_0 + \sum_{r=1}^{\infty} a_r T_r(x) \tag{2.81}$$

to signify that the series on the right of (2.81) is the Chebyshev series for the function f. (It should cause no confusion that we used the symbol $\sim$ earlier in a different sense in the statement of Theorem 2.1.8.)

Example 2.2.1 Let us derive the Chebyshev series for $\sin^{-1}x$. On substituting $x = \cos\theta$, we have $\sin^{-1}x = \frac{\pi}{2} - \theta$, and obtain from (2.80) that

$$a_0 = \frac{2}{\pi}\int_0^{\pi}(\tfrac{\pi}{2} - \theta)\, d\theta = 0$$

and

$$a_r = \frac{2}{\pi}\int_0^{\pi}(\tfrac{\pi}{2} - \theta)\cos r\theta\, d\theta = \frac{2}{\pi r}\int_0^{\pi}\sin r\theta\, d\theta, \quad r > 0,$$

on using integration by parts. Thus

$$a_r = \frac{2}{\pi r^2}(1-(-1)^r), \quad r > 0,$$

which is zero when $r > 0$ is even, and we obtain

$$\sin^{-1}x \sim \frac{4}{\pi}\sum_{r=1}^{\infty}\frac{T_{2r-1}(x)}{(2r-1)^2}. \qquad \blacksquare$$

If f is sufficiently differentiable, we can use (2.65) and (2.69) in (2.79) to show that the Chebyshev coefficients for f can be expressed in the form

$$a_r = \frac{2^{r+1}\, r!}{\pi\,(2r)!} \int_{-1}^{1} f^{(r)}(x)(1-x^2)^{r-1/2} dx. \tag{2.82}$$

We see from (2.82) that if f and all its derivatives are nonnegative on $[-1, 1]$, then all its Chebyshev coefficients are nonnegative. It is not hard to see that the same holds for the coefficients of any Jacobi series. We also have the following estimate for the Chebyshev coefficients, which can be justified by using a similar method to that used in proving the analogous result for the Legendre coefficients in Theorem 2.1.6. See Problem 2.2.17.

Theorem 2.2.3 If $f^{(r)}$ is continuous in $(-1, 1)$, then the Chebyshev coefficient a_r is given by

$$a_r = \frac{1}{2^{r-1}\, r!}\, f^{(r)}(\xi_r), \tag{2.83}$$

where $\xi_r \in (-1, 1)$. ■

Example 2.2.2 Consider (2.82) when $f(x) = e^x$, and write

$$I_r = \int_{-1}^{1} e^x (1-x^2)^{r-1/2} dx.$$

Then, on integrating by parts twice, we find that

$$I_r = -(2r-1)(2r-2)I_{r-1} + (2r-1)(2r-3)I_{r-2}, \quad r \geq 2.$$

From (2.82) we see that the Chebyshev coefficient a_r for e^x is

$$a_r = \frac{2^{r+1}\, r!}{\pi\,(2r)!} I_r,$$

and thus we obtain

$$a_r = -(2r-2)a_{r-1} + a_{r-2}, \quad r \geq 2.$$

Like the recurrence relation that we derived in (2.41) for the Legendre coefficients for e^x, this recurrence relation is numerically unstable and is thus of limited practical use. ■

The members of any orthogonal system form a basis for the polynomials. Thus, given an infinite power series, we could transform it to give a series involving the terms of a given orthogonal system. We will illustrate this with the Chebyshev polynomials, which are particularly easy to manipulate. Let us begin by writing

$$x = \cos\theta = \frac{1}{2}\left(e^{i\theta} + e^{-i\theta}\right),$$

where $i^2 = -1$, and we can also write

$$T_n(x) = \cos n\theta = \frac{1}{2}\left(e^{in\theta} + e^{-in\theta}\right). \tag{2.84}$$

Then we have

$$x^n = \frac{1}{2^n}\left(e^{i\theta} + e^{-i\theta}\right)^n,$$

and on using the binomial expansion, we obtain

$$x^n = \frac{1}{2^n}\sum_{r=0}^{n}\binom{n}{r} e^{ir\theta}e^{-i(n-r)\theta} = \frac{1}{2^n}\sum_{r=0}^{n}\binom{n}{r} e^{i(2r-n)\theta}. \tag{2.85}$$

We can combine the rth and $(n-r)$th terms within the latter sum in (2.85), using (2.84) and the fact that the two binomial coefficients involved are equal, to give

$$\binom{n}{r} e^{i(2r-n)\theta} + \binom{n}{n-r} e^{-i(2r-n)\theta} = 2\binom{n}{r} T_{n-2r}(x). \tag{2.86}$$

If n is odd, the terms in the sum in (2.85) combine in pairs, as in (2.86), to give a sum of multiples of odd-order Chebyshev polynomials. We obtain

$$x^{2n+1} = \frac{1}{2^{2n}}\sum_{r=0}^{n}\binom{2n+1}{r} T_{2n+1-2r}(x). \tag{2.87}$$

If n is even in (2.85), we have an odd number of terms in the sum, of which all but one pair off to give the terms involving the even-order Chebyshev polynomials $T_n, T_{n-2}, \ldots, T_2$, leaving a single term, which corresponds to $r = \frac{1}{2}n$ and gives the contribution involving T_0. Thus we obtain

$$x^{2n} = \frac{1}{2^{2n}}\binom{2n}{n} T_0(x) + \frac{1}{2^{2n-1}}\sum_{r=0}^{n-1}\binom{2n}{r} T_{2n-2r}(x). \tag{2.88}$$

Now suppose we have a function f expressed as an infinite power series,

$$f(x) = \sum_{r=0}^{\infty} c_r x^r, \tag{2.89}$$

and let this series be uniformly convergent on $[-1, 1]$. We then express each monomial x^r in terms of the Chebyshev polynomials, using (2.87) when r is of the form $2n+1$, and (2.88) when r is of the form $2n$. On collecting together all terms involving T_0, T_1, and so on, we obtain a series of the form

$$f(x) = \frac{1}{2}a_0 + \sum_{r=1}^{\infty} a_r T_r(x). \tag{2.90}$$

It follows from the uniform convergence of the power series in (2.89) that the series defined by (2.90) is also uniformly convergent. Thus, for any integer $s \geq 0$, we may multiply (2.90) throughout by $(1-x^2)^{-1/2}T_s(x)$ and integrate term by term over $[-1,1]$. Due to the orthogonality property (2.66), we obtain

$$\int_{-1}^{1}(1-x^2)^{-1/2}f(x)T_s(x)dx = a_s\int_{-1}^{1}(1-x^2)^{-1/2}[T_s(x)]^2dx, \quad s>0,$$

and when $s=0$ we have

$$\int_{-1}^{1}(1-x^2)^{-1/2}f(x)dx = \frac{1}{2}a_0\int_{-1}^{1}(1-x^2)^{-1/2}dx.$$

In view of (2.77) and (2.79), we can see that (2.90) is *the* Chebyshev series for f. The relation between the Chebyshev coefficients a_r and the coefficients c_r in the power series then follows from (2.87) and (2.88). We obtain

$$a_r = \frac{1}{2^{r-1}}\sum_{s=0}^{\infty}\frac{1}{2^{2s}}\begin{pmatrix} r+2s \\ s \end{pmatrix}c_{r+2s}, \tag{2.91}$$

for all $r \geq 0$.

Example 2.2.3 With $f(x)=e^x$ in (2.89), we have $c_r = 1/r!$, and we see from (2.91) that the Chebyshev coefficients for e^x are

$$a_r = \frac{1}{2^{r-1}}\sum_{s=0}^{\infty}\frac{1}{2^{2s}}\cdot\frac{1}{s!(r+s)!}, \tag{2.92}$$

for all $r \geq 0$. The first two Chebyshev coefficients for e^x, rounded to six decimal places, are $a_0 = 2.532132$ and $a_1 = 1.130318$. The next few are given in the following table:

r	2	3	4	5	6	7
a_r	0.271495	0.044337	0.005474	0.000543	0.000045	0.000003

We note from (2.92) that for large r,

$$a_r \sim \frac{1}{2^{r-1}}\cdot\frac{1}{r!},$$

compared with $c_r = 1/r!$. ■

The inner product (2.42) can be generalized by incorporating a weight function ω, writing

$$(f,g) = \int_{-1}^{1}\omega(x)f(x)g(x)dx. \tag{2.93}$$

Then we can show that

$$(f - p_n, q_s^\omega) = 0, \quad 0 \le s \le n, \tag{2.94}$$

where p_n is the partial orthogonal series for f with respect to the weight function ω, and the polynomials q_s^ω are orthogonal on $[-1, 1]$ with respect to ω. We may justify (2.94) in the same way that we justified its special case (2.44). Then, using the same method of proof as we used for Theorem 2.1.7, we obtain the following generalization.

Theorem 2.2.4 The error term $f - p_n$ changes sign on at least $n+1$ points in the interior of $[-1, 1]$, where p_n is the partial sum of the orthogonal series for f with respect to a given weight function ω. ■

In particular, the error term of the partial Chebyshev series $f - p_n$ changes sign on at least $n+1$ points in the interior of $[-1, 1]$. We now consider three theorems that generalize results, stated earlier, related to the Legendre polynomials.

Theorem 2.2.5 Given a weight function ω, the polynomial $p \in P_n$ that minimizes

$$\int_{-1}^{1} \omega(x)[x^{n+1} - p(x)]^2 dx$$

is $p(x) = x^{n+1} - q_{n+1}^\omega(x)$, where q_{n+1}^ω is the orthogonal polynomial of degree $n + 1$, with leading coefficient unity, with respect to the weight function ω.

Proof. The proof is similar to that used in Problem 2.1.8 to verify the special case where $\omega(x) = 1$. ■

Theorem 2.2.6 Given the norm (2.53) based on the weight function ω, the minimum value of $\|(x - x_0) \cdots (x - x_n)\|$ is attained when the x_j are the zeros of the orthogonal polynomial q_{n+1}^ω.

Proof. This result follows immediately from Theorem 2.2.5. ■

Theorem 2.2.7 Given the norm (2.53) based on the weight function ω, let p_n denote the best weighted square norm approximation, with respect to ω, for a given function f. Then if $f^{(n+1)}$ is continuous on $[-1, 1]$, there exists a number ζ in $(-1, 1)$ such that

$$\|f - p_n\| = \frac{|f^{(n+1)}(\zeta)|}{(n+1)!} \|(x - x_0^*) \cdots (x - x_n^*)\|, \tag{2.95}$$

where the x_j^* denote the zeros of the orthogonal polynomial q_{n+1}^ω.

Proof. This result may be justified in the same way as Theorem 2.1.8, which is concerned with the special case where $\omega(x) = 1$. ■

A special case of the last theorem is Theorem 2.1.8 concerning the error of the truncated Legendre series. Another special case of Theorem 2.2.7, which we now give as a separate theorem, concerns the error of the Chebyshev series.

Theorem 2.2.8 Let p_n denote the truncated Chebyshev series of degree n for f on $[-1, 1]$. Then if $f^{(n+1)}$ is continuous on $[-1, 1]$, there exists a number ζ in $(-1, 1)$ such that

$$\|f - p_n\| = \frac{\sqrt{\pi}}{2^{n+1/2}} \frac{|f^{(n+1)}(\zeta)|}{(n+1)!}, \tag{2.96}$$

where the norm is given by (2.53) with $\omega(x) = (1 - x^2)^{-1/2}$.

Proof. This result follows from Theorem 2.2.7 and Problem 2.2.18. ■

We continue this section with a brief account of the system of polynomials that are orthogonal on $[-1, 1]$ with respect to the weight function $(1-x^2)^{1/2}$. As we already mentioned, these are called the Chebyshev polynomials of the second kind, a special case of the ultraspherical polynomials. Like *the* Chebyshev polynomials T_n, the Chebyshev polynomials of the second kind can be expressed simply in terms of circular functions. For $n \geq 0$, let us write

$$U_n(x) = \frac{\sin(n+1)\theta}{\sin\theta}, \quad \text{where } x = \cos\theta, \quad -1 \leq x \leq 1. \tag{2.97}$$

On making the substitution $x = \cos\theta$, the interval $-1 \leq x \leq 1$ is mapped to $0 \leq \theta \leq \pi$, and we obtain

$$\int_{-1}^{1} (1 - x^2)^{1/2} U_r(x) U_s(x) dx = \int_0^{\pi} \sin(r+1)\theta \sin(s+1)\theta \, d\theta, \quad r \neq s.$$

Since

$$\sin(r+1)\theta \sin(s+1)\theta = \frac{1}{2}(\cos(r-s)\theta - \cos(r+s+2)\theta),$$

we find that

$$\int_{-1}^{1} (1 - x^2)^{1/2} U_r(x) U_s(x) dx = 0, \quad r \neq s,$$

showing that the functions U_r are orthogonal on $[-1, 1]$ with respect to the weight function $(1 - x^2)^{1/2}$. It remains to show that $U_n \in P_n$. Now,

$$\sin(n+2)\theta + \sin n\theta = 2\sin(n+1)\theta\cos\theta,$$

and if we divide throughout by $\sin\theta$ and use (2.97), we obtain the recurrence relation

$$U_{n+1}(x) = 2xU_n(x) - U_{n-1}(x), \quad n \geq 1. \tag{2.98}$$

We observe from (2.97) that $U_0(x) = 1$ and $U_1(x) = 2x$, and it follows from (2.98) by induction that $U_n \in P_n$ for each n. Thus (2.97) defines a system of polynomials that are orthogonal on $[-1, 1]$ with respect to the weight function $(1 - x^2)^{1/2}$.

Since $x = 1$ corresponds to $\theta = 0$ under the transformation $x = \cos\theta$, for $x \in [-1, 1]$, we have

$$U_n(1) = \lim_{\theta \to 0} \frac{\sin(n+1)\theta}{\sin\theta} = n + 1, \quad n \geq 0,$$

using L'Hospital's rule. Since, as for all the ultraspherical polynomials, U_n is even or odd when n is even or odd, $U_n(-1) = (-1)^n(n + 1)$. The zeros of U_n (see Problem 2.2.21) are

$$x = \cos(j\pi/(n + 1)), \quad 1 \leq j \leq n.$$

Let us consider the orthogonal series based on the Chebyshev polynomials of the second kind, whose weight function is $(1 - x^2)^{1/2}$. As we saw in (2.55), an orthogonal coefficient is determined by the ratio of two integrals. Let us begin with the integral in the denominator on the right side of (2.55), with $\omega(x) = (1 - x^2)^{1/2}$, and make the substitution $x = \cos\theta$ to give

$$\int_{-1}^{1} (1 - x^2)^{1/2} [U_n(x)]^2 dx = \int_0^{\pi} \sin^2(n + 1)\theta \, d\theta = \frac{1}{2}\pi,$$

for all $n \geq 0$. Thus if b_n denotes the coefficient of U_n in the orthogonal series, it follows from (2.55) that

$$b_n = \frac{2}{\pi} \int_0^{\pi} \sin\theta \sin(n + 1)\theta \, f(\cos\theta) d\theta, \tag{2.99}$$

for all $n \geq 0$.

We can derive a simple relation between the coefficients b_n of a series of the second kind, defined by (2.99), and the coefficients a_n of *the* Chebyshev series, defined by (2.80). For we can write

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

for all $n \geq 0$, and thus, comparing (2.80) and (2.99), we have

$$b_n = \frac{1}{2}(a_n - a_{n+2}), \quad n \geq 0. \tag{2.100}$$

We can obtain the relation (2.100) otherwise by formally comparing the first and second kinds of Chebyshev expansions of a given function, as we

did above in deriving (2.91) by comparing a Chebyshev series with a power series. We begin by writing

$$\sin(n+1)\theta - \sin(n-1)\theta = 2\cos n\theta \sin\theta.$$

On dividing throughout by $\sin\theta$, we obtain

$$U_n(x) - U_{n-2}(x) = 2T_n(x), \quad n \geq 2, \tag{2.101}$$

and we also have

$$U_0(x) = T_0(x) \qquad \text{and} \qquad U_1(x) = 2T_1(x). \tag{2.102}$$

Then

$$\frac{1}{2}a_0T_0(x) + \sum_{r=1}^{\infty} a_rT_r(x) = \frac{1}{2}\sum_{r=0}^{1} a_rU_r(x) + \frac{1}{2}\sum_{r=2}^{\infty} a_r(U_r(x) - U_{r-2}(x)).$$

If we now express the right side of the latter equation as

$$\frac{1}{2}\sum_{r=0}^{1} a_rU_r(x) + \frac{1}{2}\sum_{r=2}^{\infty} a_r(U_r(x) - U_{r-2}(x)) = \sum_{r=0}^{\infty} b_rU_r(x),$$

it is clear that the relation between the b_n and the a_n is as given above in (2.100).

To complete this section on the Chebyshev polynomials, we will consider again the Hermite interpolating polynomial $p_{2n+1}(x)$, defined in (1.38), which interpolates a given function $f(x)$, and whose first derivative interpolates $f'(x)$, at $n+1$ arbitrary abscissas $x_0, x_1, \ldots, x_n$. Let us take the x_i to be the zeros of T_{n+1} arranged in the order $-1 < x_0 < x_1 < \cdots < x_n < 1$, so that

$$x_i = \cos\frac{(2n-2i+1)\pi}{2n+2}, \quad 0 \leq i \leq n.$$

It then follows from Problem 1.1.3 that the fundamental polynomial on the zeros of T_{n+1} is

$$L_i(x) = \frac{T_{n+1}(x)}{(x-x_i)T'_{n+1}(x_i)},$$

and on using the expression for the derivative of the Chebyshev polynomial in Problem 2.2.11, we can write this as

$$L_i(x) = (-1)^{n-i}\frac{T_{n+1}(x)\sqrt{1-x_i^2}}{(n+1)(x-x_i)}.$$

It also follows from Problem 1.1.3 and the expressions for the first two derivatives of the Chebyshev polynomial in Problem 2.2.11 that

$$L'_i(x_i) = \frac{1}{2}\frac{T''_{n+1}(x_i)}{T'_{n+1}(x_i)} = \frac{1}{2}\frac{x_i}{1-x_i^2}.$$

On substituting these results into (1.39) and (1.40), we find that the Hermite interpolating polynomial on the zeros of T_{n+1} is given by

$$p_{2n+1}(x) = \sum_{i=0}^{n} [f(x_i)u_i(x) + f'(x_i)v_i(x)],$$

where

$$u_i(x) = \left(\frac{T_{n+1}(x)}{n+1}\right)^2 \frac{1 - x_i x}{(x - x_i)^2} \tag{2.103}$$

and

$$v_i(x) = \left(\frac{T_{n+1}(x)}{n+1}\right)^2 \left(\frac{1 - x_i^2}{x - x_i}\right). \tag{2.104}$$

Problem 2.2.1 Use integration by parts on (2.62) to show that

$$\Gamma(x+1) = x\Gamma(x).$$

Show that $\Gamma(1) = 1$, and $\Gamma(n+1) = n!$, for all integers $n \geq 0$, and thus verify (2.61).

Problem 2.2.2 Use the Leibniz rule (1.83) to obtain

$$Q_n^{(\alpha,\beta)}(-1) = (-1)^n \begin{pmatrix} n+\beta \\ n \end{pmatrix} = (-1)^n \frac{\Gamma(n+\beta+1)}{\Gamma(n+1)\Gamma(\beta+1)},$$

a companion formula for (2.61).

Problem 2.2.3 Verify that Theorem 2.2.1 reduces to Theorem 2.1.4 when we put $\alpha = \beta = 0$.

Problem 2.2.4 Deduce from (2.60) that

$$Q_n^{(-1/2,-1/2)}(1) = \frac{1}{2^{2n}} \begin{pmatrix} 2n \\ n \end{pmatrix},$$

and use Stirling's formula (see Problem 2.1.12) to show that

$$Q_n^{(-1/2,-1/2)}(1) \sim \frac{1}{\sqrt{\pi n}}.$$

Problem 2.2.5 Verify from Theorem 2.2.1 that

$$Q_{n+1}^{(-1/2,-1/2)}(x) = a_n x Q_n^{(-1/2,-1/2)}(x) - c_n Q_{n-1}^{(-1/2,-1/2)}(x),$$

where

$$a_n = \frac{2n+1}{n+1} \quad \text{and} \quad c_n = \frac{4n^2 - 1}{4n(n+1)},$$

with $Q_0^{(-1/2,-1/2)}(x) = 1$ and $Q_1^{(-1/2,-1/2)}(x) = \frac{1}{2}x$.

Problem 2.2.6 For $x \geq 1$, write

$$x = \cosh\theta = \frac{1}{2}(e^{\theta} + e^{-\theta}).$$

Verify that $\cosh\theta \geq 1$ for all real θ and show also that

$$\cosh(n+1)\theta + \cosh(n-1)\theta = 2\cosh\theta\cosh n\theta.$$

Hence verify by induction, using the recurrence relation (2.72), that

$$T_n(x) = \cosh n\theta, \quad \text{where} \quad x = \cosh\theta,$$

for $x \geq 1$. Finally, show that

$$T_n(x) = (-1)^n \cosh n\theta, \quad \text{where} \quad x = -\cosh\theta,$$

for $x \leq -1$.

Problem 2.2.7 Deduce from the recurrence relation (2.72) that $T_n(1) = 1$ and $T_n(-1) = (-1)^n$ for all $n \geq 0$.

Problem 2.2.8 Using the recurrence relation (2.72), verify that

$$T_5(x) = 16x^5 - 20x^3 + 5x \quad \text{and} \quad T_6(x) = 32x^6 - 48x^4 + 18x^2 - 1.$$

Write $T_2(T_3(x)) = 2(T_3(x))^2 - 1$, and show directly that this simplifies to give $T_6(x)$. Show also that $T_3(T_2(x)) = T_6(x)$. More generally, show that

$$T_m(T_n(x)) = T_n(T_m(x)) = T_{mn}(x),$$

for all integers $m, n \geq 0$.

Problem 2.2.9 Verify that

$$T_{2n+2}(x) = 2T_2(x)T_{2n}(x) - T_{2n-2}(x),$$

for $n \geq 1$, with $T_0(x) = 1$ and $T_2(x) = 2x^2 - 1$, giving a recurrence relation that computes only the even-order Chebyshev polynomials. Find a recurrence relation that computes only the odd-order Chebyshev polynomials.

Problem 2.2.10 Verify that

$$T_{n+k}(x)T_{n-k}(x) - (T_n(x))^2 = (T_k(x))^2 - 1$$

for all $n \geq k \geq 0$.

Problem 2.2.11 Use the chain rule of differentiation to show from the definition of the Chebyshev polynomials in (2.71) that

$$T_n'(x) = \frac{n\sin n\theta}{\sin\theta} \quad \text{and} \quad T_n''(x) = \frac{-n^2\sin\theta\cos n\theta + n\sin n\theta\cos\theta}{\sin^3\theta},$$

where $x = \cos\theta$, and deduce that T_n satisfies the second-order differential equation

$$(1-x^2)T_n''(x) - xT_n'(x) + n^2T_n(x).$$

Problem 2.2.12 By making the substitution $x = \cos\theta$, show that

$$\int_{-1}^{1} (1-x^2)^{-1/2}[T_r(x)]^2 dx = \int_0^{\pi} \cos^2 r\theta \, d\theta = \frac{1}{2}\int_0^{\pi} (1+\cos 2r\theta)\, d\theta,$$

and hence show that

$$\int_{-1}^{1} (1-x^2)^{-1/2}[T_r(x)]^2 dx = \begin{cases} \pi, & r = 0, \\ \frac{1}{2}\pi, & r > 0. \end{cases}$$

Problem 2.2.13 Derive the Chebyshev series for $[(1+x)/2]^{1/2}$. Hint: Make the substitution $x = \cos\theta$ and use the identity

$$2\cos\frac{1}{2}\theta\cos r\theta = \cos\left(r+\frac{1}{2}\right)\theta + \cos\left(r-\frac{1}{2}\right)\theta.$$

Problem 2.2.14 Obtain the Chebyshev series for $\cos^{-1} x$.

Problem 2.2.15 Find the Chebyshev series for $(1-x^2)^{1/2}$.

Problem 2.2.16 Assuming that the Chebyshev series for $\sin^{-1} x$, derived in Example 2.2.1, converges uniformly to $\sin^{-1} x$ on $[-1,1]$, deduce that

$$\frac{1}{1^2} + \frac{1}{3^2} + \frac{1}{5^2} + \frac{1}{7^2} + \cdots = \frac{\pi^2}{8}.$$

Problem 2.2.17 Let

$$J_r = \int_{-1}^{1} (1-x^2)^{r-1/2} dx.$$

Verify that $J_0 = \pi$, as obtained in Problem 2.2.12, and use integration by parts to show that

$$J_r = (2r-1)\int_{-1}^{1} x^2(1-x^2)^{r-3/2} dx = (2r-1)(J_{r-1} - J_r).$$

Deduce that

$$J_r = \frac{\pi\,(2r)!}{2^{2r}\,(r!)^2},$$

and hence verify (2.83).

Problem 2.2.18 Let $x_0^*, \ldots, x_n^*$ denote the zeros of the Chebyshev polynomial T_{n+1}, and let $\|\cdot\|$ denote the weighted square norm, defined by (2.53), with weight function $(1-x^2)^{-1/2}$. Use the substitution $x = \cos\theta$ to show that

$$\|(x-x_0^*)\cdots(x-x_n^*)\| = \frac{1}{2^n}\left(\int_{-1}^{1} (1-x^2)^{-1/2}[T_{n+1}(x)]^2 dx\right)^{1/2} = \frac{\pi^{1/2}}{2^{n+1/2}}.$$

Problem 2.2.19 Begin with the Maclaurin series

$$\cos \tfrac{1}{2}\pi x = \sum_{r=0}^{\infty} (-1)^r \frac{(\pi x)^{2r}}{2^{2r}(2r)!},$$

and use the relation (2.91) to compute the first few Chebyshev coefficients for $\cos(\frac{1}{2}\pi x)$.

Problem 2.2.20 In the text it is shown that

$$U_n(1) = n + 1 \qquad \text{and} \qquad U_n(-1) = (-1)^n(n + 1).$$

Deduce these results alternatively by using an induction argument on the recurrence relation for U_n.

Problem 2.2.21 Deduce from (2.97) that the Chebyshev polynomial of the second kind U_n is zero for values of θ such that $\sin(n + 1)\theta = 0$ and $\sin\theta$ is nonzero. Show that

$$\sin(n + 1)\theta = 0 \quad \Rightarrow \quad (n + 1)\theta = j\pi \quad \Rightarrow \quad x = \cos(j\pi/(n + 1)),$$

and hence show that U_n has all its n zeros in $(-1, 1)$, at the abscissas $x = \cos(j\pi/(n + 1))$, for $1 \le j \le n$.

Problem 2.2.22 Use the relations (2.101) and (2.102), which express the Chebyshev polynomials T_n, to derive an expression for every monomial in terms of the U_n, like those given for x^{2n+1} and x^{2n} in terms of the T_n, in (2.87) and (2.88), respectively.

Problem 2.2.23 Deduce from (2.71) and (2.97) that

$$T'_{n+1}(x) = (n + 1)U_n(x),$$

and hence show that $T'_n(1) = n^2$.

Problem 2.2.24 Show that

$$U_n(x) = T_n(x) + xU_{n-1}(x), \quad n \ge 1,$$

and deduce that

$$U_n(x) = \sum_{r=0}^{n} x^r T_{n-r}(x), \quad n \ge 0.$$

Problem 2.2.25 Deduce from (2.101) that

$$U_{2n+1}(x) = 2\sum_{r=0}^{n} T_{2n+1-2r}(x)$$

and

$$U_{2n}(x) = T_0(x) + 2\sum_{r=0}^{n-1} T_{2n-2r}(x).$$

2.3 Finite Point Sets

In the last section we constructed a sequence of polynomials (q_n^ω) that are orthogonal on $[-1,1]$ with respect to a given weight function ω, and we used these as a basis for constructing best approximations for functions defined on $[-1,1]$.

We will now consider best approximations for functions that are defined on a finite set of points, say $X = \{x_0, x_1, \ldots, x_N\}$. Associated with each point x_j we will assign a positive number ω_j, called a *weight*. Then, given a function f defined on the point set X, we seek a polynomial $p \in P_n$ that minimizes

$$\sum_{j=0}^{N} \omega_j [f(x_j) - p(x_j)]^2, \tag{2.105}$$

which we will call a least squares approximation to f on X with respect to the given weights. We now define

$$\|g\| = \left(\sum_{j=0}^{N} \omega_j [g(x_j)]^2 \right)^{1/2}, \tag{2.106}$$

and we can show that this is a norm, as in Definition 2.1.1. It is analogous to the norm defined by (2.53).

Following the same method as we used in Section 2.2 for weighted square norm approximations on the interval $[-1,1]$, we construct a sequence of polynomials $(q_n^\omega)_{n=0}^N$, where q_n^ω is of degree n, has leading coefficient unity, and satisfies

$$\sum_{j=0}^{N} \omega_j x_j^r q_n^\omega(x_j) = 0, \qquad 0 \le r < n. \tag{2.107}$$

We find that the orthogonal polynomials q_n^ω satisfy the recurrence relation

$$q_{n+1}^\omega(x) = (x - \alpha_n) q_n^\omega(x) - \beta_n q_{n-1}^\omega(x), \tag{2.108}$$

where

$$\alpha_n = \sum_{j=0}^{N} \omega_j x_j [q_n^\omega(x_j)]^2 \Big/ \sum_{j=0}^{N} \omega_j [q_n^\omega(x_j)]^2 \tag{2.109}$$

and

$$\beta_n = \sum_{j=0}^{N} \omega_j [q_n^\omega(x_j)]^2 \Big/ \sum_{j=0}^{N} \omega_j [q_{n-1}^\omega(x_j)]^2. \tag{2.110}$$

Note that the last two relations are analogous to (2.51) and (2.52), respectively. We then discover that

$$\|f - p\| = \left(\sum_{j=0}^{N} \omega_j [f(x_j) - p(x_j)]^2 \right)^{1/2}$$

is minimized over all $p \in P_n$, with $n \le N$, by choosing $p = p_n$, where

$$p_n(x) = \sum_{r=0}^{n} a_r q_r^{\omega}(x) \tag{2.111}$$

and

$$a_r = \sum_{j=0}^{N} \omega_j f(x_j) q_r^{\omega}(x_j) \Big/ \sum_{j=0}^{N} \omega_j [q_r^{\omega}(x_j)]^2, \quad 0 \le r \le n. \tag{2.112}$$

We note that p_N must be the interpolating polynomial for f on the set X, since then $\|f - p_N\| = 0$, and this shows why we have restricted n to be not greater than N. We also observe that the sequence of orthogonal polynomials is unchanged if we multiply all the weights by any positive constant.

It can sometimes be more convenient to work with orthogonal polynomials that do not have leading coefficient 1. Suppose $Q_r^{\omega}(x) = c_r q_r^{\omega}(x)$, where $c_r \ne 0$ may depend on r, but is independent of x. Then it follows from (2.111) and (2.112) that if we express the best square norm approximation to f on X as

$$p_n(x) = \sum_{r=0}^{n} a_r Q_r^{\omega}(x), \tag{2.113}$$

with $n \le N$, then the coefficient a_r is given by

$$a_r = \sum_{j=0}^{N} \omega_j f(x_j) Q_r^{\omega}(x_j) \Big/ \sum_{j=0}^{N} \omega_j [Q_r^{\omega}(x_j)]^2, \quad 0 \le r \le n. \tag{2.114}$$

Note that the polynomials Q_r^{ω} will satisfy the simple recurrence relation (2.108) only if every c_r is equal to 1.

Example 2.3.1 To illustrate the the foregoing material, let us choose the set X as $\{-1, 0, 1\}$, with weights ω_j all equal, and let f be defined on X by the following table:

x	-1	0	1
$f(x)$	1	2	4

We have $q_0^{\omega}(x) = 1$ and $q_1^{\omega}(x) = x$. Then we find that $\alpha_1 = 0$ and $\beta_1 = \frac{2}{3}$, so that $q_2^{\omega}(x) = x^2 - \frac{2}{3}$. On using (2.112) we find that the orthogonal coefficients are $a_0 = \frac{7}{3}$, $a_1 = \frac{3}{2}$, and $a_2 = \frac{1}{2}$. Thus the best approximation in P_1 is $\frac{7}{3} + \frac{3}{2}x$, while that in P_2 is $\frac{7}{3} + \frac{3}{2}x + \frac{1}{2}(x^2 - \frac{2}{3})$. As expected, the last polynomial interpolates f on X. ■

We require some further notation, writing $\sum'$ to denote a sum in which the first term is halved, and $\sum''$ to denote a sum in which both the first

term and the last terms are halved. Thus

$$\sum_{j=0}^{N}{}' u_j = \frac{1}{2}u_0 + u_1 + \cdots + u_N \tag{2.115}$$

and

$$\sum_{j=0}^{N}{}'' u_j = \frac{1}{2}u_0 + u_1 + \cdots + u_{N-1} + \frac{1}{2}u_N. \tag{2.116}$$

Now let $X = \{x_0, x_1, \ldots, x_N\}$, where $x_j = \cos(\pi j/N)$, $0 \le j \le N$, and let

$$\omega_j = \begin{cases} \frac{1}{2}, & j = 0 \text{ and } N, \\ 1, & 1 \le j \le N-1. \end{cases}$$

Thus X is the set of extreme points of the Chebyshev polynomial T_N. We can verify (see Problem 2.3.1) that

$$\sum_{j=0}^{N}{}'' T_r(x_j)T_s(x_j) = 0, \qquad r \ne s, \tag{2.117}$$

and so the Chebyshev polynomials are orthogonal on the extreme points of T_N with respect to the given set of weights. It then follows from (2.113) and (2.114) that the best weighted square norm approximation with respect to this set of points and weights is

$$p_n(x) = \sum_{r=0}^{n} a_r T_r(x), \tag{2.118}$$

where

$$a_r = \sum_{j=0}^{N}{}'' f(x_j)T_r(x_j) \Big/ \sum_{j=0}^{N}{}'' [T_r(x_j)]^2, \quad 0 \le r \le n. \tag{2.119}$$

We find (again see Problem 2.3.1) that

$$\sum_{j=0}^{N}{}'' [T_r(x_j)]^2 = \begin{cases} N, & r = 0 \text{ or } N, \\ \frac{1}{2}N, & 1 \le r \le N-1. \end{cases} \tag{2.120}$$

On combining (2.118), (2.119), and (2.120) we see that the best weighted square norm approximation to f on the extreme points of T_N is

$$p_n(x) = \sum_{r=0}^{n}{}' \alpha_r T_r(x), \quad n \le N, \tag{2.121}$$

say, where

$$\alpha_r = \frac{2}{N}\sum_{j=0}^{N}{}'' f(x_j)T_r(x_j), \quad 0 \le r \le n. \tag{2.122}$$

If we choose $n = N$ in (2.121), the least squares approximation must coincide with the interpolating polynomial for f on the extreme points of T_N. We now derive a connection between the above coefficients α_r and the Chebyshev coefficients a_r, defined by (2.79). First we note that

$$T_{2kN\pm r}(x_j) = \cos\left(\frac{(2kN \pm r)\pi j}{N}\right) = \cos\left(2k\pi j \pm \frac{r\pi j}{N}\right) = \cos\left(\frac{r\pi j}{N}\right),$$

and thus

$$T_{2kN\pm r}(x_j) = T_r(x_j). \tag{2.123}$$

Let us assume that *the* Chebyshev series for f, given by (2.81), converges uniformly to f on $[-1, 1]$. Then it follows from (2.122) that

$$\begin{aligned}\alpha_r &= \frac{2}{N}\sum_{j=0}^{N}{}'' \left(\sum_{s=0}^{\infty}{}' a_s T_s(x_j)\right) T_r(x_j) \\ &= \frac{2}{N}\sum_{s=0}^{\infty}{}' a_s \sum_{j=0}^{N}{}'' T_s(x_j)T_r(x_j).\end{aligned} \tag{2.124}$$

We observe from (2.123) and the orthogonality property (2.117) that the only nonzero summations over j in the second line of (2.124) are those for which $s = r$, $2N - r$, $2N + r$, $4N - r$, $4N + r$, and so on. Thus, for $r \neq 0$ or N,

$$\alpha_r = a_r + \sum_{k=1}^{\infty}(a_{2kN-r} + a_{2kN+r}), \tag{2.125}$$

and we find that (2.125) holds also for $r = 0$ and N, on examining these cases separately.

We saw in the last section that the Chebyshev polynomials are orthogonal on the interval $[-1, 1]$ with respect to a certain weight function, and we have seen above that they are orthogonal on the extreme points of T_N with respect to certain weights. We now show that they satisfy a third orthogonality property. For the Chebyshev polynomials are also orthogonal on the set $X = \{x_1^*, \ldots, x_N^*\}$ with all weights equal, where the x_j^* are the zeros of T_N. We can verify that (see Problem 2.3.2)

$$\sum_{j=1}^{N} T_r(x_j^*)T_s(x_j^*) = \begin{cases} 0, & r \neq s \text{ or } r = s = N, \\ \frac{1}{2}N, & r = s \neq 0 \text{ or } N, \\ N, & r = s = 0. \end{cases} \tag{2.126}$$

Then, following the same method as we used in deriving (2.121) and (2.122), we find that the best square norm approximation to the function f on the zeros of T_N is

$$p_n(x) = {\sum_{r=0}^{n}}' \alpha_r^* T_r(x), \qquad n \le N-1, \tag{2.127}$$

where

$$\alpha_r^* = \frac{2}{N} \sum_{j=1}^{N} f(x_j^*) T_r(x_j^*). \tag{2.128}$$

If we choose $n = N-1$ in (2.127), the least squares approximation is just the interpolating polynomial for f on the zeros of T_N.

We can express α_r^* as a sum involving the Chebyshev coefficients, using the same method as we used above to derive the expression (2.125) for α_r. We first verify that

$$T_{2kN\pm r}(x_j^*) = (-1)^k T_r(x_j^*). \tag{2.129}$$

Assuming that the Chebyshev series for f converges uniformly to f on $[-1,1]$, we find that

$$\alpha_r^* = a_r + \sum_{k=1}^{\infty} (-1)^k (a_{2kN-r} + a_{2kN+r}) \tag{2.130}$$

for $0 \le r \le N-1$, and we see from (2.125) and (2.130) that

$$\frac{1}{2}(\alpha_r + \alpha_r^*) = a_r + \sum_{k=1}^{\infty} (a_{4kN-r} + a_{4kN+r}). \tag{2.131}$$

The three expressions (2.125), (2.130), and (2.131), connecting the a_r with the coefficients α_r and α_r^*, suggest a practical method of computing the Chebyshev coefficients a_r. We choose a value of N and then compute α_r and α_r^*, using (2.125) and (2.130). If α_r and α_r^* differ by more than an acceptable amount, we increase N and recompute α_r and α_r^*. When they agree sufficiently, we use $\frac{1}{2}(\alpha_r + \alpha_r^*)$ as an approximation to a_r.

Problem 2.3.1 Show that $T_r(x_j) = \cos(\pi r j/N)$, where the x_j are the extreme points of T_N, and hence express the left side of (2.117) as a sum of cosines, using the relation

$$\cos(\theta+\phi) + \cos(\theta-\phi) = 2\cos\theta\cos\phi.$$

Evaluate the sum by using the identity

$$\cos k\theta = \frac{\sin(k+\frac{1}{2})\theta - \sin(k-\frac{1}{2})\theta}{2\sin\frac{1}{2}\theta},$$

and hence verify the orthogonality relation (2.117). Similarly verify (2.120).

Problem 2.3.2 With $x_j^* = \cos\frac{(2j-1)\pi}{2N}$, show that

$$T_r(x_j^*) = \cos\frac{(2j-1)r\pi}{2N}$$

and follow the method of the last problem to verify (2.126). Show also that

$$T_{2kN\pm r}(x_j^*) = (-1)^k T_r(x_j^*).$$

Problem 2.3.3 Verify (2.130) and (2.131).

Problem 2.3.4 Show that α_r, defined by (2.122), is the approximation to the Chebyshev coefficient a_r that is obtained by estimating the integral

$$a_r = \frac{2}{\pi}\int_0^\pi f(\cos\theta)\cos r\theta\, d\theta$$

(see (2.80)) using the composite form of the trapezoidal rule. Show also that α_r^*, defined by (2.128), is the approximation obtained by applying the composite midpoint rule to the same integral. (These integration rules are discussed in Chapter 3.)

2.4 Minimax Approximation

A best approximation with respect to the maximum norm, defined by (2.5), is called a *minimax approximation*, since we wish to *minimize*, over all p_n in P_n, the *maximum* value of the error $|f(x) - p_n(x)|$ over $[-1, 1]$. For convenience, we will continue to work with the interval $[-1, 1]$, although what we have to say about minimax approximations may be applied to any finite interval by making a linear change of variable. Minimax approximations, which are also called *uniform approximations*, are important because of the following famous theorem due to Karl Weierstrass (1815–1897).

Theorem 2.4.1 Given any $f \in C[-1, 1]$ and any $\epsilon > 0$, there exists a polynomial p such that

$$|f(x) - p(x)| < \epsilon, \qquad -1 \le x \le 1. \qquad \blacksquare \tag{2.132}$$

There are many proofs of this key theorem and we will give two in Chapter 7, based on proofs given by two mathematicians who were born in the same year, S. N. Bernstein (1880–1968) and L. Fejér (1880–1959). As we will see in this section, the property that characterizes the error $f - p_n$ of a minimax polynomial $p_n \in P_n$ for a function $f \in C[-1, 1]$ is one that is shared with the Chebyshev polynomial T_{n+1}. As we saw in the last section, the polynomial T_{n+1} attains its maximum modulus on $[-1, 1]$ at $n + 2$ points belonging to $[-1, 1]$, and $T_{n+1}(x)$ alternates in sign on these points. It is useful to have a name for this property, which we give now.

Definition 2.4.1 A function $E(x)$ is said to *equioscillate* on $n+2$ points of $[-1,1]$ if for $-1 \le x_1 < x_2 < \cdots < x_{n+2} \le 1$,

$$|E(x_j| = \max_{-1\le x\le 1} |E(x)|, \quad 1 \le j \le n+2,$$

and

$$E(x_{j+1}) = -E(x_j), \quad 1 \le j \le n+1. \qquad \blacksquare$$

Thus T_{n+1} equioscillates on the $n+2$ points $t_j = \cos[(j-1)\pi/(n+1)]$, $1 \le j \le n+2$, since $T_{n+1}(t_j) = (-1)^{j-1}$ and the maximum modulus of T_{n+1} on $[-1,1]$ is 1.

Example 2.4.1 Consider the function

$$E(x) = |x| - \frac{1}{8} - x^2$$

on $[-1,1]$. We can easily check that

$$E(-1) = -E(-\tfrac{1}{2}) = E(0) = -E(\tfrac{1}{2}) = E(1).$$

The function E is differentiable on $[-1,1]$ except at $x=0$. For $0 < x \le 1$, we have $E(x) = x - \frac{1}{8} - x^2$ and

$$E'(x) = 1 - 2x = 0 \;\Rightarrow\quad x = \frac{1}{2}.$$

Thus E has a turning value at $x = \frac{1}{2}$, and we similarly find that E has another turning value at $x = -\frac{1}{2}$. It follows that E attains its maximum modulus on $[-1,1]$ at the five points $x = 0, \pm\frac{1}{2}, \pm 1$, and E equioscillates on these points. As we will see from Theorem 2.4.2, $\frac{1}{8} + x^2$ is the minimax approximation in P_3 for $|x|$ on $[-1,1]$. Note that although $\frac{1}{8} + x^2$ is a polynomial of degree *two*, it is the minimax approximation to $|x|$ on $[-1,1]$ out of all polynomials in P_3. $\blacksquare$

Theorem 2.4.2 Let $f \in C[-1,1]$ and suppose there exists $p \in P_n$ such that $f-p$ equioscillates on $n+2$ points belonging to $[-1,1]$. Then p is the minimax approximation in P_n for f on $[-1,1]$.

Proof. Suppose p is not the minimax approximation. We will prove the theorem by showing that this assumption cannot be true. (We then say that we have obtained a *contradiction* to the initial assumption. This style of proof is called *reductio ad absurdum*.) If p is not the minimax approximation, there must exist some $q \in P_n$ such that $p+q \in P_n$ is the minimax approximation. Now let us compare the graphs of $f-p$, which equioscillates on $n+2$ points on $[-1,1]$, and $f-p-q$. Since the latter error curve must have the smaller maximum modulus, the effect of adding q to p must be to reduce the size of the modulus of the error function $f-p$ on all $n+2$

equioscillation points. In particular, this must mean that q alternates in sign on these $n+2$ points, and thus must have at least $n+1$ zeros. Since $q \in P_n$, this is impossible, contradicting our initial assumption that p is not the minimax approximation. ■

Example 2.4.2 For each $n \geq 0$, let us define

$$p(x) = x^{n+1} - \frac{1}{2^n} T_{n+1}(x),$$

so that $p \in P_n$. (See Problem 2.4.3 for a refinement of this observation.) Then, since

$$x^{n+1} - p(x) = \frac{1}{2^n} T_{n+1}(x)$$

equioscillates on $n+2$ points belonging to $[-1,1]$, it follows from Theorem 2.4.2 that p is the minimax approximation P_n for x^{n+1} on $[-1,1]$. From Theorem 2.2.5 we see that the same p is also the best approximation for x^{n+1} on $[-1,1]$ with respect to the weighted square norm, with weight $\omega(x) = (1-x^2)^{-1/2}$. Finally, if we write $\|\cdot\|_\infty$ to denote the maximum norm, defined by (2.5), we deduce from the minimax approximation for x^{n+1} that $\|(x-x_0)(x-x_1)\cdots(x-x_n)\|_\infty$ is minimized over all choices of the abscissas x_j by choosing the x_j as the zeros of T_{n+1}. Theorem 2.2.6 shows us that this also holds when we replace $\|\cdot\|_\infty$ by the weighted square norm, defined by (2.53) with weight function $(1-x^2)^{-1/2}$. ■

In Theorem 2.4.2 we showed that the equioscillation property is a sufficient condition for a minimax approximation. The next theorem, whose proof is a little harder, shows that the equioscillation property is also *necessary*, which is why we call it the characterizing property of minimax approximation.

Theorem 2.4.3 Let $f \in C[-1,1]$, and let $p \in P_n$ denote a minimax approximation for f on $[-1,1]$. Then there exist $n+2$ points on $[-1,1]$ on which $f-p$ equioscillates.

Proof. Let us write $E(x) = f(x) - p(x)$ and assume that E equioscillates on fewer than $n+2$ points. We will show that this assumption leads to a contradiction. We can exclude the case where E is the zero function, for then there is nothing to prove. Then we argue that E must equioscillate on at least two points, for otherwise, we could add a suitable constant to p so as to reduce the maximum modulus. Thus we can assume that E equioscillates on k points, where $2 \leq k < n+2$. Now let us choose $k-1$ points,

$$-1 < x_1 < x_2 < \cdots < x_{k-1} < 1,$$

so that there is one equioscillation point of E on each of the intervals

$$[-1, x_1), (x_1, x_2), \ldots, (x_{k-2}, x_{k-1}), (x_{k-1}, 1].$$

Now we construct a polynomial q such that

$$q(x) = \pm C(x - x_1)(x - x_2) \cdots (x - x_{k-1}),$$

with $C = 1/\|(x - x_1)(x - x_2) \cdots (x - x_{k-1})\|_\infty$. It follows from the choice of C that $\|q\|_\infty = 1$. Finally we choose the plus or minus sign so that q takes the same sign as E on all points where $\|E\|_\infty$ is attained. Now let S_- denote the set of points on $[-1, 1]$ where $E(x)q(x) \leq 0$, and let S_+ denote the set of points on $[-1, 1]$ where $E(x)q(x) > 0$. Thus S_+ includes all points where $\|E\|_\infty$ is attained, and the union of S_- and S_+ is simply the interval $[-1, 1]$. We now define

$$d = \max_{x \in S_-} |E(x)| < \|E\|_\infty, \tag{2.133}$$

choose any $\theta > 0$, and define ξ as any point in $[-1, 1]$ for which

$$|E(\xi) - \theta q(\xi)| = \|E - \theta q\|_\infty. \tag{2.134}$$

Obviously, $\xi \in S_-$ or $\xi \in S_+$. If $\xi \in S_-$, it follows from (2.134) and (2.133) that

$$\|E - \theta q\|_\infty = |E(\xi)| + \theta |q(\xi)| \leq d + \theta. \tag{2.135}$$

On the other hand, if $\xi \in S_+$, it follows from (2.134) that

$$\|E - \theta q\|_\infty < \max\{|E(\xi)|, \theta |q(\xi)|\} \leq \max\{\|E\|_\infty, \theta\}. \tag{2.136}$$

In view of the definition of d in (2.133), let us now restrict the value of θ so that

$$0 < \theta < \|E\|_\infty - d.$$

Then, whether ξ is in S_- or in S_+, both (2.135) and (2.136) yield

$$\|E - \theta q\|_\infty < \|E\|_\infty.$$

This contradicts our assumption that $k < n + 2$, and completes the proof. Note that the polynomial q, defined above, is in P_{k-1}, and so we are able to force the above contradiction only if $k < n + 2$, so that $k - 1 \leq n$ and thus q is in P_n. ■

The uniqueness of a best approximation for the weighted square norm, discussed in the previous section, follows from the way we derived it as the unique solution of a system of linear equations. We now show how the uniqueness of a minimax approximation can be deduced by deploying the same ideas used above in the proof of Theorem 2.4.2.

Theorem 2.4.4 If $f \in C[-1, 1]$, there is a unique minimax approximation in P_n for f on $[-1, 1]$.

Proof. Let $p \in P_n$ denote a minimax approximation for f, and if this approximation is not unique, let $p + q \in P_n$ denote another minimax approximation. Then, using the same kind of argument as in the proof of Theorem 2.4.2, we argue that $q \in P_n$ must be alternately ≥ 0 and ≤ 0 on the $n + 2$ equioscillation points of $f - p$. We deduce that this is possible only when $q(x) \equiv 0$, which completes the proof. ■

We saw in the last section that the best approximation with respect to a weighted square norm is an interpolating polynomial for the given function. The above equioscillation theorems show that this is true also for minimax approximations, and we state this as a theorem.

Theorem 2.4.5 Let $f \in C[-1, 1]$ and let $p \in P_n$ denote the minimax polynomial for f. Then there exist $n+1$ points in $[-1, 1]$ on which p interpolates f. ■

It is convenient to introduce a new item of notation in the following definition.

Definition 2.4.2 We write

$$E_n(f) = \|f - p\|_\infty, \tag{2.137}$$

where $\| \cdot \|_\infty$ denotes the maximum norm on $[-1, 1]$, and $p \in P_n$ is the minimax approximation for $f \in C[-1, 1]$. ■

Theorem 2.4.5 leads to an estimate of the minimax error $E_n(f)$ that is similar to the estimate given in Theorem 2.2.7 for the error of a best approximation with respect to a weighted square norm.

Theorem 2.4.6 If $f \in C^{n+1}[-1, 1]$, then the error of the minimax polynomial $p \in P_n$ for f on $[-1, 1]$ satisfies

$$E_n(f) = \|f - p\|_\infty = \frac{1}{2^n} \frac{|f^{(n+1)}(\xi)|}{(n+1)!}, \tag{2.138}$$

where $\xi \in (-1, 1)$.

Proof. In the light of Theorem 2.4.5, let the minimax polynomial p interpolate f at the points $x_0, x_1, \ldots, x_n$ in $[-1, 1]$. Then, from (1.25),

$$f(x) - p(x) = (x - x_0)(x - x_1) \cdots (x - x_n) \frac{f^{(n+1)}(\xi_x)}{(n+1)!},$$

where $\xi_x \in (-1, 1)$, and thus

$$\|f - p\|_\infty \geq \|(x - x_0)(x - x_1) \cdots (x - x_n)\|_\infty \frac{\min |f^{(n+1)}(x)|}{(n+1)!}, \tag{2.139}$$

the minimum being over $[-1,1]$. It follows from our discussion in Example 2.4.2 that

$$\|(x-x_0)\cdots(x-x_n)\|_\infty \geq \|(x-x_0^*)\cdots(x-x_n^*)\|_\infty = \frac{1}{2^n},$$

where the x_j^* denote the zeros of T_{n+1}, and thus (2.139) yields

$$\|f-p\|_\infty \geq \frac{1}{2^n}\frac{\min|f^{(n+1)}(x)|}{(n+1)!}. \tag{2.140}$$

If p^* denotes the interpolating polynomial for f on the zeros of T_{n+1}, then, again using the error term for interpolation (1.25), we obtain the inequality

$$\|f-p^*\|_\infty \leq \|(x-x_0^*)(x-x_1^*)\cdots(x-x_n^*)\|_\infty \frac{\max|f^{(n+1)}(x)|}{(n+1)!},$$

and so

$$\|f-p\|_\infty \leq \|f-p^*\|_\infty \leq \frac{1}{2^n}\frac{\max|f^{(n+1)}(x)|}{(n+1)!}. \tag{2.141}$$

Finally, the theorem follows from (2.140), (2.141), and the continuity of $f^{(n+1)}$. ■

Example 2.4.3 If we apply Theorem 2.4.6 to the function e^x, we obtain

$$\frac{1}{2^n}\frac{e^{-1}}{(n+1)!} \leq \|e^x - p_n(x)\|_\infty \leq \frac{1}{2^n}\frac{e}{(n+1)!},$$

where p_n denotes the minimax polynomial in P_n for e^x on $[-1,1]$. For example, with $n=6$ and 7, we have the bounds

$$0.11\times 10^{-5} < \|e^x - p_6(x)\|_\infty < 0.85\times 10^{-5}$$

and

$$0.71\times 10^{-7} < \|e^x - p_7(x)\|_\infty < 0.53\times 10^{-6}. \quad ■$$

The next theorem shows that the minimax polynomial is not the only approximant to f that has an error term of the form (2.138).

Theorem 2.4.7 If p_n^* denotes the interpolating polynomial on the zeros of the Chebyshev polynomial T_{n+1} for $f \in C^{n+1}[-1,1]$, then

$$\|f-p_n^*\|_\infty = \frac{1}{2^n}\frac{|f^{(n+1)}(\eta)|}{(n+1)!}, \tag{2.142}$$

where $\eta \in (-1,1)$.

Proof. This is easily verified by adapting the proof of Theorem 2.4.6, and the details are left to the reader. ■

Example 2.4.4 Let p_n^* denote the interpolating polynomial for e^x on the zeros of the Chebyshev polynomial T_{n+1}. Then it follows from Theorem 2.4.7 that

$$\max_{-1\le x\le 1} |e^x - p_n^*(x)| = \frac{e^{\eta_n}}{2^n(n+1)!}, \tag{2.143}$$

where $-1 < \eta_n < 1$. It is clear that the sequence of polynomials (p_n^*) converges uniformly to e^x on $[-1,1]$. ■

We saw in Section 2.2 that when the weight function is even, the best weighted square norm approximation for a function f is even or odd, according as f is even or odd, respectively. We can deduce from the equioscillation property (see Problems 2.4.9 and 2.4.10) that a minimax approximation for a function f is even or odd, according as f is even or odd, respectively.

It is clear from Example 2.4.1 that $\frac{1}{8} + x^2$ is the minimax approximation in P_2 for $|x|$ on $[-1,1]$. However, we found that the error function equioscillates on five points, and so, by Theorem 2.4.2, $\frac{1}{8} + x^2$ is also the minimax approximation in P_3 for $|x|$, as stated in Example 2.4.1. This shows that if we seek a minimax polynomial $p_n \in P_n$ for a given function f, we could find that $f - p_n$ has more than $n+2$ equioscillation points. For example, the minimax polynomial in P_0 (a constant) for the function T_k on $[-1,1]$, with $k > 0$, is simply the zero function. In this case, the error function has $k+1$ equioscillation points. The following theorem gives a condition that ensures that $f - p_n$ has exactly $n+2$ equioscillation points.

Theorem 2.4.8 If $f \in C^{n+1}[-1,1]$ and $f^{(n+1)}$ has no zero in $(-1,1)$, then the error of the minimax polynomial $p \in P_n$ for f on $[-1,1]$ equioscillates on $n+2$ points, and on no greater number of points.

Proof. Suppose that $f - p$ has k equioscillation points in the interior of $[-1,1]$. Since there can be at most two equioscillation points at the endpoints $x = \pm 1$, Theorem 2.4.3 shows that $k \ge n$. This means that $f' - p'$ has at least $k \ge n$ zeros in $(-1,1)$. On applying Rolle's theorem, we deduce that $f'' - p''$ has at least $k-1$ zeros in $(-1,1)$. Since the nth derivative of p' is zero, the repeated application of Rolle's theorem n times to the function $f' - p'$ shows that $f^{(n+1)}$ has at least $k-n$ zeros in $(-1,1)$. Since, by our assumption, $f^{(n+1)}$ has no zero in $(-1,1)$, we deduce that $k = n$, and this completes the proof. ■

Apart from the approximation of x^{n+1} by a polynomial in P_n, and some simple cases involving low-order polynomial approximations for a few functions, we have said nothing about how to *compute* minimax approximations. We now discuss a class of algorithms based on the work of E. Ya. Remez (1896–1975) in the 1930s.

Algorithm 2.4.1 The following algorithm computes a sequence of polynomials that converges uniformly to the minimax approximation $p \in P_n$

for a given function $f \in C[-1, 1]$. Corresponding to each polynomial in the sequence there is a set X of $n+2$ points. This set of points converges to a set of points in $[-1, 1]$ on which the error $f - p$ equioscillates.

Step 1 Choose an initial set $X = \{x_1, x_2, \ldots, x_{n+2}\} \subset [-1, 1]$.

Step 2 Solve the system of linear equations

$$f(x_j) - q(x_j) = (-1)^j e, \quad 1 \le j \le n+2,$$

to obtain a real number e and a polynomial $q \in P_n$.

Step 3 Change the set X (as described below) and go to Step 2 unless a "stopping criterion" has been met. ■

Let us denote a polynomial $q \in P_n$ that occurs in Step 2 of the above Remez algorithm by

$$q(x) = a_0 + a_1 x + \cdots + a_n x^n.$$

We now verify that, provided that the x_j are distinct, the system of linear equations in Step 2 has a nonsingular matrix, and so the linear system has a unique solution. The matrix associated with these equations is

$$\mathbf{A} = \begin{bmatrix} 1 & x_1 & x_1^2 & \cdots & x_1^n & -e \\ 1 & x_2 & x_2^2 & \cdots & x_2^n & +e \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n+2} & x_{n+2}^2 & \cdots & x_{n+2}^n & (-1)^{n+2} e \end{bmatrix}.$$

Now, a necessary and sufficient condition for a matrix to be singular is that its columns be linearly dependent. If the columns of the above matrix are denoted by $\mathbf{c}_1$, $\mathbf{c}_2, \ldots, \mathbf{c}_{n+2}$, and they are linearly dependent, then there exist real numbers λ_1, $\lambda_2, \ldots, \lambda_{n+2}$, not all zero, such that

$$\lambda_1 \mathbf{c}_1 + \lambda_2 \mathbf{c}_2 + \cdots + \lambda_{n+2} \mathbf{c}_{n+2} = \mathbf{0}, \tag{2.144}$$

where $\mathbf{0}$ is the zero column vector with $n+2$ elements. If we write

$$q(x) = \lambda_1 + \lambda_2 x + \cdots + \lambda_{n+1} x^n,$$

then the vector equation (2.144) is equivalent to the scalar system of linear equations

$$q(x_j) + \lambda_{n+2}(-1)^j e = 0, \quad 1 \le j \le n+2.$$

Since this would imply that $q \in P_n$ is alternately ≥ 0 and ≤ 0 on the $n+2$ distinct x_j, this is impossible, and we conclude that the above matrix $\mathbf{A}$ is nonsingular. Thus, provided that the x_j are distinct, the system of equations in Step 2 of the Remez algorithm always has a unique solution.

We usually choose the set X in Step 1 of the Remez algorithm to be the set of $n+2$ extreme points of T_{n+1}. Then we see from Example 2.4.2 that for the special case of $f(x) = x^{n+1}$, this choice of X immediately gives us the minimax polynomial when we solve the linear system in Step 2.

The solution of the linear system in Step 2 yields a polynomial $q \in P_n$ and some number e. It follows from Theorem 2.4.2 that if $\|f - q\|_\infty$ is sufficiently close to $|e|$, then q must be close to the minimax polynomial, and we would then terminate the algorithm. Otherwise, if ξ is a point such that $|f(\xi) - q(\xi)| = \|f - q\|_\infty$, we change the point set X by including ξ and deleting one of the existing points in X so that $f - q$ still alternates in sign on the new point set. In the scheme that we have just proposed, we change only one point each time we carry out Step 3. There is a second version of the algorithm, which converges more rapidly, in which we amend the point set X by including $n+2$ local extreme values of $f - q$ and delete existing points of X so that $f - q$ still alternates in sign on the new point set.

Suppose f is an even function and we seek its minimax approximation in P_{2n+1}. Let us choose as the initial set X the $2n+3$ extreme points of T_{2n+2}. Thus X consists of the abscissa $x = 0$ and $n+1$ pairs of abscissas of the form $\pm x_j$. The minimax polynomial p must be an even polynomial (see Problem 2.4.9), and is therefore of degree $2n$ or less. We can then see from the symmetry in the linear equations in Step 2 that the same is true of the polynomial q. The linear equations in Step 2 thus contain only $n+2$ coefficients, namely, e and the $n+1$ coefficients to determine p. Because of the symmetry in X, we need write down only the $n+2$ equations corresponding to the nonnegative abscissas in X. We can see that this simplification persists as we work through the algorithm. At each stage the set X is symmetric with respect to the origin, and the polynomial q is even. A corresponding simplification also occurs when we apply the Remez algorithm to an odd function.

Example 2.4.5 Let us use the Remez algorithm to compute the minimax polynomial in P_3 for $1/(1+x^2)$ on $[-1, 1]$. We require five equioscillation points, and initially in Step 1 we choose the set of extreme points of T_4,

$$X = \left\{-1, -\tfrac{1}{2}\sqrt{2}, 0, \tfrac{1}{2}\sqrt{2}, 1\right\}.$$

In Step 2 let us write $q(x) = a_0 + a_1 x + a_2 x^2 + a_3 x^3$. We see that the solution of the system of five linear equations is obtained by choosing $a_1 = a_3 = 0$ and solving the three equations corresponding to $x = 0$, $\frac{1}{2}\sqrt{2}$, and 1. This is in agreement with our remark above concerning the application of the Remez algorithm to an even function. We thus have the equations

$$\begin{aligned} 1 - a_0 &= e, \\ \frac{2}{3} - a_0 - \frac{1}{2}a_2 &= -e, \\ \frac{1}{2} - a_0 - a_2 &= e, \end{aligned}$$

whose solution is $a_0 = \frac{23}{24}$, $a_2 = -\frac{1}{2}$, and $e = \frac{1}{24} \approx 0.041667$. In this case, the error function has just one turning value in $(0, 1)$. Although we could now differentiate the error function to determine the turning value, this would not be in the spirit of the algorithm. Instead, we simply evaluate the error function on a suitably dense set of points on $[0, 1]$, say at intervals of 0.01. We find that the extreme values on this finite point set are as follows:

x	0	0.64	1
$f(x) - p(x)$	0.041667	−0.044112	0.041667

We now change the point set to $X = \{-1, -0.64, 0, 0.64, 1\}$ and go to Step 2. This time the solution of the linear system is $a_0 = 0.957111$, $a_2 = -\frac{1}{2}$, and $e = 0.042889$, to six decimal places. We evaluate the new error function on the same finite set of points and find that the extreme values are again at $x = 0$, 0.64, and 1, where $f - q$ assumes the values 0.042889, −0.042890, and 0.042889, respectively. Since the set X cannot be changed further, we accept the current polynomial q as being sufficiently close to the true minimax polynomial p, and terminate the algorithm. If we wish greater accuracy we can repeat the calculation, but evaluate the error at each stage on a finer grid of points. If we return to the original error function $f - q$ and evaluate it at intervals of 0.001 in the vicinity of 0.64, we find that this extreme point is more precisely given by 0.644, where the value of $f - q$ is −0.044120. Therefore, we now change the point set to $X = \{-1, -0.644, 0, 0.644, 1\}$ and go to Step 2. This time the solution of the linear system is $a_0 = 0.957107$, $a_2 = -\frac{1}{2}$, and $e = 0.042893$, to six decimal places. We evaluate the new error function at intervals of 0.01, and again refine the extreme point near 0.64. We find that the extreme values are

x	0	0.644	1
$f(x) - q(x)$	0.042893	−0.042893	0.042893

Thus we have the polynomial $0.957107 - 0.5x^2$ as a refined estimate of p, with minimax error 0.042893. In fact, for the function in this example, we can find p exactly (see Problem 2.4.7). This serves as a check on our calculations here. We can see that our last estimate of p above is correct to six decimal places. ■

We now state and prove a theorem, due to C. J. de la Vallée Poussin (1866–1962), which we can apply to give lower and upper bounds for the minimax error after each iteration of the Remez algorithm.

Theorem 2.4.9 Let $f \in C[-1,1]$ and $q \in P_n$. Then if $f-q$ alternates in sign on $n+2$ points

$$-1 \le x_1 < x_2 < \cdots < x_{n+2} \le 1,$$

we have

$$\min_j |f(x_j) - q(x_j)| \le \|f-p\|_\infty \le \|f-q\|_\infty, \tag{2.145}$$

where $p \in P_n$ denotes the minimax approximation for f on $[-1,1]$.

Proof. We need concern ourselves only with the left-hand inequality in (2.145), since the right-hand inequality follows immediately from the definition of p as the minimax approximation. Now let us write

$$(f(x_j) - q(x_j)) - (f(x_j) - p(x_j)) = p(x_j) - q(x_j). \tag{2.146}$$

If the left-hand inequality in (2.145) does *not* hold, then the right-hand inequality in the following line must hold:

$$|f(x_j) - p(x_j)| \le \|f-p\|_\infty < |f(x_j) - q(x_j))|, \quad 1 \le j \le n+2.$$

The left hand inequality above is a consequence of the definition of the norm. It follows that the sign of the quantity on the left side of (2.146) is that of $f(x_j) - q(x_j)$, which alternates over the x_j. Thus $p - q \in P_n$ alternates in sign on $n+2$ points, which is impossible. This completes the proof. ■

The above theorem has the following obvious application to the Remez algorithm.

Theorem 2.4.10 At each stage in the Remez algorithm, we have

$$|e| \le E_n(f) \le \|f-q\|_\infty, \tag{2.147}$$

where $E_n(f)$ is the minimax error, and e and q are obtained from the solution of the linear equations in Step 2 of the algorithm.

Proof. From the way e and q are constructed in Step 2 of the algorithm, $f-q$ alternates in sign on the $n+2$ points x_j belonging to the set X, and

$$|e| = |f(x_j) - q(x_j)|, \quad 1 \le j \le n+2.$$

Hence (2.147) follows immediately from (2.145). ■

Let $e^{(i)}$ and $q^{(i)}$ denote the number e and the polynomial q that occur in the ith stage of the Remez algorithm. The sequence $\left(\left|e^{(i)}\right|\right)$ increases, and the sequence $\left(\left\|f-q^{(i)}\right\|_\infty\right)$ decreases, and in principle both sequences converge to the common limit $E_n(f)$. In practice, these limits are usually not attained, because we evaluate the error function at each stage at only a finite number of points. However, the inequalities (2.147) provide a reliable indicator of how close we are to the minimax polynomial at each stage.

Example 2.4.6 To illustrate the second version of the Remez algorithm, mentioned above, in which we amend the point set X by including $n+2$ local extreme values of $f-q$, let us find the minimax $p \in P_3$ for e^x on $[-1,1]$. We will take the initial set X as the set of extreme points of T_4, as we did in Example 2.4.5. The solution of the linear equations in Step 2 then yields, to six decimal places, $e = 0.005474$, and

$$q(x) = 0.994526 + 0.995682x + 0.543081x^2 + 0.179519x^3.$$

On evaluating $f-q$ at intervals of 0.01, we find that it has extreme values at $x = -1, -0.68, 0.05, 0.73$, and 1. The error $f-q$ has the value $e = 0.005474$ at $x = \pm 1$, by construction, and has the following values at the three interior extreme points:

x	-0.68	0.05	0.73
$f(x)-q(x)$	-0.005519	0.005581	-0.005537

If we now evaluate $f-q$ at intervals of 0.001 in the vicinity of the interior extreme points, we can refine these to give -0.683, 0.049, and 0.732. We therefore amend the set X, making three changes to give

$$X = \{-1, -0.683, 0.049, 0.732, 1\},$$

and repeat Step 2. This time we obtain $e = 0.005528$, and

$$q(x) = 0.994580 + 0.995668x + 0.542973x^2 + 0.179534x^3.$$

When we evaluate $f-p$ at intervals of 0.01, we find that it has extreme points at $x = -1$, -0.68, 0.05, 0.73, and 1, where it assumes the values ± 0.005528. We therefore accept this polynomial q as being sufficiently close to the minimax approximation p. The interior equioscillation points are more precisely -0.682, 0.050, and 0.732. ■

If a function is defined on a set $X = \{x_0, x_1, \ldots, x_N\}$, we can define a norm

$$\|f\| = \max_{0 \le j \le N} |f(x_j)|,$$

and seek a polynomial $p \in P_n$, with $n \le N$, that minimizes $\|f-p\|$. This is called a minimax approximation on the set X. Such minimax approximations behave much like those on a finite interval, as we have discussed in this section. For example, with $n < N$, the error function $f-p$ equioscillates on $n+2$ points in X. Also, when using a Remez algorithm, we can expect to locate a minimax polynomial on a finite interval only approximately, as we saw above, whereas we can find a minimax polynomial on a finite point set X precisely.

Problem 2.4.1 Show that the minimax polynomial $p \in P_1$ for $2/(x+3)$ on $[-1,1]$ is $p(x) = \frac{1}{2}\sqrt{2} - \frac{1}{4}x$, and find the minimax error.

Problem 2.4.2 Let $ax+b$ denote the minimax polynomial in P_1 for the function $1/x$ on the interval $[\alpha,\beta]$, where $0 < \alpha < \beta$. Show that

$$a = -\frac{1}{\alpha\beta} \quad \text{and} \quad b = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}} + \frac{1}{\sqrt{\beta}}\right)^2,$$

and that

$$\max_{\alpha\le x\le\beta}\left|ax+b-\frac{1}{x}\right| = \frac{1}{2}\left(\frac{1}{\sqrt{\alpha}} - \frac{1}{\sqrt{\beta}}\right)^2.$$

Verify that the minimax error is attained at α, β, and at the intermediate point $\sqrt{\alpha\beta}$.

Problem 2.4.3 Use the property that the Chebyshev polynomial T_n is an odd or even function when n is odd or even, respectively, to show that the polynomial p defined by

$$p(x) = x^{n+1} - \frac{1}{2^n}T_{n+1}(x)$$

is of degree $n-1$.

Problem 2.4.4 Let $p_n \in P_n$ denote the minimax polynomial for $\sin\frac{1}{2}\pi x$ on $[-1,1]$. Show that

$$0 < \|\sin\tfrac{1}{2}\pi x - p_n(x)\|_\infty \le \left(\frac{\pi}{4}\right)^{n+1}\frac{2}{(n+1)!}.$$

Problem 2.4.5 Let $f \in C[-1,1]$ and let

$$m = \min_{-1\le x\le 1} f(x), \quad M = \max_{-1\le x\le 1} f(x).$$

Show that $\frac{1}{2}(m+M) \in P_0$ is a minimax approximation for f on $[-1,1]$.

Problem 2.4.6 By first showing that the error function has turning values at $x = -\frac{2}{3}$ and $x = \frac{1}{3}$, show that the function $1/(3x+5)$ on $[-1,1]$ has the minimax approximation $\frac{1}{48}(6x^2-8x+9)$ in P_2, and determine the minimax error.

Problem 2.4.7 Verify that the minimax approximation in P_3 for the function $1/(1+x^2)$ on $[-1,1]$ is $\frac{1}{4}+\frac{1}{2}\sqrt{2}-\frac{1}{2}x^2$, by showing that the error function equioscillates on the five points $0, \pm\xi$, and ± 1, where $\xi^2 = \sqrt{2}-1$. Find the minimax error.

Problem 2.4.8 If $p \in P_n$ denotes the minimax polynomial for f on $[-1,1]$, show that $\lambda p + q$ is the minimax approximation for $\lambda f + q$, for any real λ and any $q \in P_n$. Thus show, using the result of Problem 2.4.6, that the minimax polynomial in P_2 for $(x+3)/(3x+5)$ is $(6x^2-8x+21)/36$.

Problem 2.4.9 Let f be an even function on $[-1,1]$, and let $p \in P_n$ denote the minimax approximation for f. By considering the equioscillation points, deduce that $p(-x)$ is the minimax approximation for $f(-x)$. Since $f(-x) = f(x)$ and the minimax approximation is unique, deduce that p is also an even function.

Problem 2.4.10 By adapting the argument used in the previous problem, show that a minimax approximation for an odd function is itself odd.

2.5 The Lebesgue Function

Recall equation (1.10),

$$p_n(x) = \sum_{i=0}^{n} f(x_i) L_i(x),$$

where p_n is the Lagrange form for the polynomial that interpolates f on the abscissas x_0, $x_1, \ldots, x_n$, and the fundamental polynomials L_i are defined by (1.9). If the values $f(x_i)$ all have errors of modulus not greater than $\epsilon > 0$, what can we say about the resulting size of the error in evaluating $p_n(x)$ at any point on the interval $[a,b]$? Suppose that instead of evaluating p_n, we evaluate p_n^*, where

$$p_n^*(x) = \sum_{i=0}^{n} f^*(x_i) L_i(x)$$

and

$$|f(x_i) - f^*(x_i)| \le \epsilon, \quad 0 \le i \le n.$$

It follows that

$$|p_n(x) - p_n^*(x)| \le \epsilon \, \lambda_n(x),$$

where

$$\lambda_n(x) = \sum_{i=0}^{n} |L_i(x)|, \quad a \le x \le b. \tag{2.148}$$

Hence we have

$$\max_{a \le x \le b} |p_n(x) - p_n^*(x)| \le \epsilon \, \Lambda_n, \tag{2.149}$$

where

$$\Lambda_n = \max_{a \le x \le b} \lambda_n(x). \tag{2.150}$$

Thus errors in the function values $f(x_i)$ of modulus not greater than ϵ result in an error in the interpolating polynomial whose modulus is not greater than $\epsilon \Lambda_n$. We call Λ_n the *Lebesgue* constant and λ_n the Lebesgue function

associated with the point set $\{x_0, x_1, \ldots, x_n\} \subset [a, b]$. These are named after Henri Lebesgue (1875–1941). Both the textbook by T. J. Rivlin [48] and the paper by Lev Brutman [6] contain valuable surveys on the Lebesgue functions.

Now let us consider an infinite triangular array of abscissas,

$$X: \quad \begin{array}{llll} x_0^{(0)} & & & \\ x_0^{(1)} & x_1^{(1)} & & \\ x_0^{(2)} & x_1^{(2)} & x_2^{(2)} & \\ \vdots & & & \\ x_0^{(n)} & x_1^{(n)} & \cdots & x_n^{(n)} \\ \vdots & & & \end{array} \tag{2.151}$$

where for each $n \geq 0$,

$$a \leq x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \leq b. \tag{2.152}$$

For any function f defined on $[a, b]$, we can construct a sequence of polynomials p_0, p_1, $p_2, \ldots$, where p_n interpolates f on the set of abscissas $\left\{x_0^{(n)}, x_1^{(n)}, \ldots, x_n^{(n)}\right\}$, which forms the $(n+1)$th row of the above triangular array. When we wish to emphasize their dependence on the array X, we will write $\Lambda_n(X)$, $n \geq 0$, to denote the Lebesgue constants on $[a, b]$ associated with this array of abscissas, and write the Lebesgue functions in the form $\lambda_n(X; x)$.

Example 2.5.1 In the above triangular array X let us choose $x_0^{(0)} = 0$, and for each $n \geq 1$, let

$$x_i^{(n)} = -1 + \frac{2i}{n}, \qquad 0 \leq i \leq n.$$

Thus the abscissas in each row of X are equally spaced on the interval $[-1, 1]$. Let us take $[a, b]$ in (2.152) to be $[-1, 1]$.

If we interpolate on the abscissas in the second row of X, namely $\{-1, 1\}$, the fundamental polynomials are $L_0(x) = \frac{1}{2}(1 - x)$ and $L_1(x) = \frac{1}{2}(1 + x)$. Thus the Lebesgue function on $[-1, 1]$ is

$$\lambda_1(x) = \frac{1}{2}(1 - x) + \frac{1}{2}(1 + x) = 1, \quad -1 \leq x \leq 1,$$

and the Lebesgue constant on $[-1, 1]$ is $\Lambda_1(X) = 1$.

Let us also obtain an explicit form for the Lebesgue function $\lambda_2(x)$ on $[-1, 1]$ for the set of interpolating abscissas in the third row of X, which is

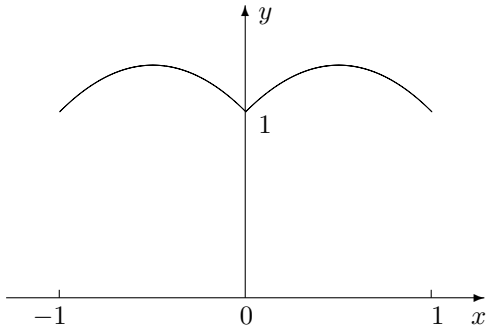

FIGURE 2.2. The Lebesgue function $\lambda_2(x)$ defined in Example 2.5.1.

$\{-1, 0, 1\}$. In this case the fundamental polynomials are

$$L_0(x) = \frac{1}{2}x(x-1), \quad L_1(x) = 1 - x^2, \quad L_1(x) = \frac{1}{2}x(x+1).$$

The Lebesgue function $\lambda_2(x)$ may be expressed as a quadratic polynomial in each of the intervals $[-1, 0]$ and $[0, 1]$. We find that

$$\lambda_2(x) = \frac{1}{2}x(x-1) + (1-x^2) - \frac{1}{2}x(1+x) = 1 - x - x^2, \quad -1 \le x \le 0,$$

and

$$\lambda_2(x) = -\frac{1}{2}x(x-1) + (1-x^2) + \frac{1}{2}x(1+x) = 1 + x - x^2, \quad 0 \le x \le 1.$$

It is easily verified that $\lambda_2(x) = 1$ for $x = 0, \pm 1$, and that $\lambda_2(x) \ge 1$ on $[-1, 1]$. We find that $\Lambda_2(X)$, the maximum value of $\lambda_2(x)$ on $[-1, 1]$ is $\frac{5}{4}$, and this value is attained at $x = \pm\frac{1}{2}$. See Figure 2.2. ■

Theorem 2.5.1 The Lebesgue function λ_n, defined by (2.148), is continuous on $[a, b]$ and is a polynomial of degree at most n on each of the subintervals $[a, x_0]$, $[x_0, x_1]$, $[x_1, x_2], \ldots, [x_{n-1}, x_n]$, and $[x_n, b]$. We also have

$$\lambda_n(x_i) = 1, \quad 0 \le i \le n, \tag{2.153}$$

and

$$\lambda_n(x) \ge 1, \quad a \le x \le b. \tag{2.154}$$

Proof. Equation (2.153) follows immediately from (2.148). To verify the inequality (2.154) we deduce from (1.42) that

$$1 = |L_0(x) + L_1(x) + \cdots + L_n(x)| \le |L_0(x)| + |L_1(x)| + \cdots + |L_n(x)| = \lambda_n(x)$$

for $a \le x \le b$. Each function $|L_i(x)|$, and therefore the Lebesgue function λ_n, is obviously a polynomial of degree at most n on each of the intervals

$[a, x_0]$, $[x_0, x_1]$, $[x_1, x_2], \ldots, [x_{n-1}, x_n]$, and $[x_n, b]$. Also, each $|L_j(x)|$ is continuous on $[a, b]$, and thus the Lebesgue function λ_n is continuous on $[a, b]$. This completes the proof. ■

We now come to the main result of this section.

Theorem 2.5.2 Let $p_n \in P_n$ denote the minimax approximation for a given function $f \in C[a, b]$, and let $p_n^* \in P_n$ denote the interpolating polynomial for f on the abscissas $x_i^{(n)}$ in the $(n + 1)$th row of the triangular array X defined in (2.151). Then

$$\|f - p_n^*\|_\infty \leq (1 + \Lambda_n(X))E_n(f), \tag{2.155}$$

where $\|\cdot\|_\infty$ denotes the maximum norm on $[a, b]$ and $E_n(f)$ is the minimax error, defined in (2.137).

Proof. We have from (1.10) that

$$p_n^*(x) = \sum_{i=0}^{n} f(x_i)L_i^{(n)}(x),$$

where $L_i^{(n)}$ is the fundamental polynomial that takes the value 1 at the abscissa $x_i^{(n)}$ and is zero on all abscissas $x_j^{(n)}$ with $j \neq i$. It then follows from the uniqueness of the interpolating polynomial that

$$p_n(x) = \sum_{i=0}^{n} p_n(x_i)L_i^{(n)}(x),$$

since this relation holds for any polynomial in P_n. On subtracting the last two equations, we immediately obtain

$$p_n^*(x) - p_n(x) = \sum_{i=0}^{n} (f(x_i) - p_n(x_i))L_i^{(n)}(x),$$

and thus

$$|p_n^*(x) - p_n(x)| \leq \lambda_n(x) \max_{0 \leq i \leq n} |f(x_i) - p_n(x_i)|,$$

where λ_n is the Lebesgue function. We deduce that

$$\|p_n - p_n^*\| = \|p_n^* - p_n\| \leq \Lambda_n(X)E_n(f). \tag{2.156}$$

Let us now write

$$f(x) - p_n^*(x) = (f(x) - p_n(x)) + (p_n(x) - p_n^*(x)),$$

from which we can derive the inequality

$$\|f - p_n^*\| \leq \|f - p_n\| + \|p_n - p_n^*\|,$$

and we immediately have

$$\|f - p_n^*\| \leq (1 + \Lambda_n(X))E_n(f),$$

which completes the proof. ■

Theorem 2.5.3 The Lebesgue constants

$$\Lambda_n(X) = \max_{a \leq x \leq b} \sum_{i=0}^{n} \left| L_i^{(n)}(x) \right|, \quad n \geq 1,$$

are unchanged if we carry out a linear transformation $x = \alpha t + \beta$, with $\alpha \neq 0$, and the triangular array of interpolating abscissas $X = \left(x_i^{(n)} \right)$ is mapped to the triangular array $T = \left(t_i^{(n)} \right)$, where $x_i^{(n)} = \alpha t_i^{(n)} + \beta$.

Proof. Let the interval $a \leq x \leq b$ be mapped to $c \leq t \leq d$ under the linear transformation $x = \alpha t + \beta$, and let the fundamental polynomial $L_i^{(n)}(x)$ be mapped to

$$M_i^{(n)}(t) = \prod_{j \neq i} \left(\frac{t - t_j^{(n)}}{t_i^{(n)} - t_j^{(n)}} \right),$$

so that the Lebesgue function $\lambda_n(x)$ is mapped to

$$\lambda_n^*(t) = \sum_{i=0}^{n} \left| M_i^{(n)}(t) \right|.$$

Finally, we define

$$\Lambda_n^*(T) = \max_{c \leq t \leq d} \lambda_n^*(t).$$

We can verify that

$$\frac{x - x_j^{(n)}}{x_i^{(n)} - x_j^{(n)}} = \frac{t - t_j^{(n)}}{t_i^{(n)} - t_j^{(n)}},$$

and hence $M_i^n(t) = L_i^{(n)}(x)$. Consequently,

$$\Lambda_n^*(T) = \max_{c \leq t \leq d} \lambda_n^*(t) = \max_{a \leq t \leq b} \lambda_n(x) = \Lambda_n(X),$$

and this completes the proof. ■

The following result shows that the Lebesgue constants are not increased if we include the endpoints a and b in the set of interpolating abscissas.

Theorem 2.5.4 Consider an infinite triangular array X, as defined in (2.151), where

$$a \leq x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} \leq b, \quad n \geq 0.$$

Now define an infinite triangular array T, where $t_0^{(0)}$ is defined arbitrarily, and for each $n > 0$, the elements in the $(n+1)$th row of T satisfy

$$x_i^{(n)} = \alpha_n t_i^{(n)} + \beta_n, \quad 0 \le i \le n.$$

In keeping with this transformation, we define

$$M_i^{(n)}(t) = L_i^{(n)}(x), \quad \text{where} \quad x = \alpha_n t + \beta_n, \quad 0 \le i \le n.$$

Then, if α_n and β_n are chosen so that $t_0^{(n)} = a$ and $t_n^{(n)} = b$, we have

$$\Lambda_n(T) = \max_{a \le t \le b} |\lambda_n^*(t)| = \max_{x_0^{(n)} \le x \le x_n^{(n)}} |\lambda_n(x)| \le \max_{a \le x \le b} |\lambda_n(x)| = \Lambda_n(X).$$

Proof. The proof of this theorem is on the same lines as that of Theorem 2.5.3. Note that since

$$x_0^{(n)} = \alpha_n t_0^{(n)} + \beta_n \quad \text{and} \quad x_n^{(n)} = \alpha_n t_n^{(n)} + \beta_n,$$

we need to choose

$$\alpha_n = \frac{x_n^{(n)} - x_0^{(n)}}{b - a} \quad \text{and} \quad \beta_n = \frac{b x_0^{(n)} - a x_n^{(n)}}{b - a}. \qquad \blacksquare$$

Example 2.5.2 Let us consider the Lebesgue constants $\Lambda_n(X)$ for interpolation at equally spaced abscissas. For convenience, we will work on the interval $[0, 1]$, rather than on $[-1, 1]$, as we did in Example 2.5.1. Thus the abscissas in the triangular array X are $x_0^{(0)} = \frac{1}{2}$ and

$$x_i^{(n)} = \frac{i}{n}, \qquad 0 \le i \le n.$$

Let us evaluate the Lebesgue function $\lambda_n(x)$ at

$$x = \xi_n = \frac{2n-1}{2n},$$

the midpoint of the subinterval $[\frac{n-1}{n}, 1]$. We already know that $\lambda_n(x)$ has the value 1 at the endpoints of this subinterval. A little calculation shows that

$$|L_i(\xi_n)| = \frac{1}{|2n - 2i - 1|} \cdot \frac{(2n-1)(2n-3)\cdots 3 \cdot 1}{2^n i!(n-i)!}.$$

We can simplify this last equation to give

$$|L_i(\xi_n)| = \frac{1}{|2n - 2i - 1|} \cdot \frac{1}{2^{2n}} \binom{2n}{n} \binom{n}{i}, \tag{2.157}$$

from which we derive the inequality

$$|L_i(\xi_n)| \geq \frac{1}{2n-1} \cdot \frac{1}{2^{2n}} \left(\begin{array}{c} 2n \\ n \end{array} \right) \left(\begin{array}{c} n \\ i \end{array} \right). \tag{2.158}$$

If we now sum the above inequality over i, and evaluate the binomial expansion of $(1+x)^n$ at $x=1$ to give

$$\sum_{i=0}^{n} \left(\begin{array}{c} n \\ i \end{array} \right) = 2^n,$$

we see from (2.148) that

$$\lambda_n(\xi_n) > \frac{1}{2n-1} \cdot \frac{1}{2^n} \left(\begin{array}{c} 2n \\ n \end{array} \right), \quad \text{where} \quad \xi_n = \frac{2n-1}{2n}. \tag{2.159}$$

Note that the above inequality holds strictly because we have equality in (2.158) only for $i=0$. Thus

$$\lambda_n(\xi_n) > \frac{\mu_n}{2n-1}, \quad \text{where} \quad \xi_n = \frac{2n-1}{2n},$$

and μ_n is defined in (2.29). If we now apply Stirling's formula, as we did to estimate μ_n in Problem 2.1.12, we find that

$$\Lambda_n(X) \geq \frac{1}{2n-1} \cdot \frac{1}{2^n} \left(\begin{array}{c} 2n \\ n \end{array} \right) \sim \frac{2^{n-1}}{\sqrt{\pi}\, n^{3/2}}. \tag{2.160}$$

In view of the factor 2^{n-1} on the right of (2.160), we have proved that the Lebesgue constants for equally spaced abscissas tend to infinity at least exponentially with n. ■

It is natural to seek an infinite triangular array of interpolating abscissas X that gives the smallest values of $\Lambda_n(X)$ for every value of n. It would be nice if for such an *optimal* array X, the sequence $(\Lambda_n(X))_{n=0}^{\infty}$ were bounded. However, this is not so, as the following result of Paul Erdős (1913–1996) shows.

Theorem 2.5.5 There exists a positive constant c such that

$$\Lambda_n(X) > \frac{2}{\pi} \log n - c, \tag{2.161}$$

for *all* infinite triangular arrays X. ■

For a proof, see [18]. A simple proof of a slightly weaker version of this theorem, that

$$\Lambda_n(X) > \frac{2}{\pi^2} \log n - 1$$

for every triangular array X, is proved in Rivlin [48].

Thus, for every choice of the array X, the sequence $(\Lambda_n(X))_{n=0}^{\infty}$ grows at least as fast as $\log n$. Moreover, as we will see below, interpolation at the zeros of the Chebyshev polynomials yields a sequence of Lebesgue constants that grows only logarithmically with n and so is close to the optimal choice of abscissas. It is therefore clear that as measured by the rate of growth of the Lebesgue constants, interpolation at the zeros of the Chebyshev polynomials is substantially superior to interpolation at equally spaced abscissas.

Let T denote the infinite triangular array, as depicted in (2.151), whose $(n+1)$th row consists of the zeros of T_{n+1}, which we will write as

$$x_i^{(n)} = \cos\theta_i, \quad \text{where} \quad \theta_i = \frac{(2n+1-2i)\pi}{2n+2}, \quad 0 \le i \le n, \tag{2.162}$$

so that

$$-1 < x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} < 1.$$

We can deduce from the result in Problem 1.1.3 that the fundamental polynomial $L_i^{(n)}$ can be expressed in the form

$$L_i^{(n)}(x) = \frac{T_{n+1}(x)}{\left(x - x_i^{(n)}\right) T'_{n+1}\left(x_i^{(n)}\right)}. \tag{2.163}$$

From the expression for the derivative of a Chebyshev polynomial given in Problem 2.2.11, we can see that

$$T'_{n+1}(x_i^{(n)}) = \frac{(n+1)\sin(n+1)\theta_i}{\sin\theta_i} = (-1)^{n-i}\left(\frac{n+1}{\sin\theta_i}\right),$$

and hence, with $x = \cos\theta$,

$$L_i^{(n)}(x) = (-1)^{n-i}\frac{\cos(n+1)\theta}{n+1} \cdot \frac{\sin\theta_i}{\cos\theta - \cos\theta_i}, \tag{2.164}$$

so that

$$\lambda_n(T;x) = \frac{|\cos(n+1)\theta|}{n+1} \sum_{i=0}^{n} \frac{\sin\theta_i}{|\cos\theta - \cos\theta_i|}. \tag{2.165}$$

S. N. Bernstein (1880–1968) obtained the asymptotic estimate (see [4])

$$\Lambda_n(T) \sim \frac{2}{\pi}\log(n+1), \tag{2.166}$$

and D. L. Berman [2] obtained the upper bound

$$\Lambda_n(T) < 4\sqrt{2} + \frac{2}{\pi}\log(n+1), \tag{2.167}$$

which, together with Bernstein's asymptotic estimate, tells us more about $\Lambda_n(T)$. Luttmann and Rivlin [36] conjectured that $\Lambda_n(T) = \lambda_n(T;1)$, and showed that

$$\lim_{n\to\infty}\left[\lambda_n(T;1) - \frac{2}{\pi}\log(n+1)\right] = \frac{2}{\pi}\left(\gamma + \log\frac{8}{\pi}\right) \approx 0.9625, \tag{2.168}$$

where

$$\gamma = \lim_{n\to\infty}\left[1 + \frac{1}{2} + \frac{1}{3} + \cdots + \frac{1}{n} - \log n\right] \approx 0.5772$$

is Euler's constant. The theorem that follows, which verifies Luttmann and Rivlin's conjecture, is due to Ehlich and Zeller [16] (see also Powell [46]). The proof given here is based on that given by Rivlin [48].

Theorem 2.5.6 The Lebesgue constants for the triangular array T, whose $(n+1)$th row consists of the zeros of T_{n+1}, are given by

$$\Lambda_n(T) = \lambda_n(T;1) = \frac{1}{n+1}\sum_{i=0}^{n}\cot\frac{\theta_i}{2} = \frac{1}{n+1}\sum_{i=0}^{n}\tan\frac{(2i+1)\pi}{4(n+1)}, \tag{2.169}$$

where θ_i is defined in (2.162).

Proof. We begin by rewriting (2.164) in the form (see Problem 2.5.2)

$$L_i^{(n)}(x) = (-1)^{n-i}\frac{\cos(n+1)\theta}{2(n+1)}\left(\cot\frac{\theta+\theta_i}{2} - \cot\frac{\theta-\theta_i}{2}\right),$$

and thus obtain

$$\lambda_n(T;x) = \frac{|\cos(n+1)\theta|}{2(n+1)}\sum_{i=0}^{n}\left|\cot\frac{\theta+\theta_i}{2} - \cot\frac{\theta-\theta_i}{2}\right|, \tag{2.170}$$

where $x = \cos\theta$. Let us write

$$C(\theta) = |\cos(n+1)\theta|\sum_{i=0}^{n}\left|\cot\frac{\theta+\theta_i}{2} - \cot\frac{\theta-\theta_i}{2}\right|. \tag{2.171}$$

Since $C(\theta) = C(-\theta)$, this is an even function that is periodic, with period 2π, and is positive for $-\infty < \theta < \infty$. If we replace θ by $\theta - k\pi/(n+1)$, where $1 \le k \le n+1$, then (see Problem 2.5.3) the only change that occurs on the right side of (2.171) is that the $2n+2$ cotangents are paired differently.

Let the maximum value of $C(\theta)$ on $[0,\pi]$ be attained at $\theta = \theta'$ such that

$$\frac{(2m-1)\pi}{2n+2} < \theta' \le \frac{(2m+1)\pi}{2n+2}, \tag{2.172}$$

where $0 \le m \le n$. We see from (2.172) that θ' belongs to an interval of width $\pi/(n+1)$ with midpoint $m\pi/(n+1)$, and to a smaller interval when $m = 0$ or n. In either case, we have

$$\left|\theta' - \frac{m\pi}{n+1}\right| \le \frac{\pi}{2(n+1)}.$$

Let us now define $\theta'' = |\theta' - m\pi/(n+1)|$, and then we have

$$0 \le \theta'' \le \frac{\pi}{2(n+1)}. \tag{2.173}$$

We know that $C(\theta'') = C(-\theta'')$ and that the expressions for $C(\theta'')$ and $C(\theta')$ differ only in the way the $2n+2$ cotangents are paired in (2.171).

It follows from (2.162) and (2.173) that

$$0 \le \frac{\theta'' + \theta_i}{2} \le \frac{\pi}{2} \quad \text{and} \quad 0 \le -\frac{\theta'' - \theta_i}{2} \le \frac{\pi}{2}, \quad 0 \le i \le n,$$

and thus

$$\cot\frac{\theta'' + \theta_i}{2} \ge 0 \quad \text{and} \quad \cot\frac{\theta'' - \theta_i}{2} \le 0, \quad 0 \le i \le n.$$

The latter inequalities make it clear that if we put $\theta = \theta''$ in (2.171), the resulting expression will have the same value for *any* permutation of the $2n+2$ cotangents. This shows that

$$\max_{0\le\theta\le\pi} C(\theta) = C(\theta') = C(\theta'') = \max_{0\le\theta\le\pi/(2n+2)} C(\theta).$$

To pursue this further, we see from (2.165), (2.170), and (2.171) that

$$C(\theta) = 2|\cos(n+1)\theta| \sum_{i=0}^{n} \frac{\sin\theta_i}{|\cos\theta - \cos\theta_i|}.$$

Then, for $0 \le \theta \le \pi/(2n+2)$, we have

$$C(\theta) = 2\sum_{i=0}^{n} \sin\theta_i \frac{\cos(n+1)\theta}{\cos\theta - \cos\theta_i},$$

and we see from Problem 2.5.4 that

$$C(\theta) = 2^{n+1} \sum_{i=0}^{n} \sin\theta_i \prod_{j\ne i} (\cos\theta - \cos\theta_j). \tag{2.174}$$

Since each factor $\cos\theta - \cos\theta_j$ is a nonnegative function of θ that decreases monotonically on $[0, \pi/(2n+2)]$, we see immediately that the maximum value of $C(\theta)$ is attained at $\theta = 0$. On comparing (2.171) and (2.170), and putting $\theta = 0$, we see that

$$\Lambda_n(T) = \lambda_n(T;1) = \frac{1}{n+1}\sum_{i=0}^{n} \cot\frac{\theta_i}{2} = \frac{1}{n+1}\sum_{i=0}^{n} \cot\frac{(2n+1-2i)\pi}{4(n+1)},$$

and (2.169) follows. This completes the proof. ■

It is not hard to deduce from (2.169) that the Lebesgue constants $\Lambda_n(T)$ grow logarithmically with n. Let

$$g(\theta) = \tan\theta - \theta,$$

so that $g(0) = 0$, and since

$$g'(\theta) = \sec^2\theta - 1 = \tan^2\theta,$$

we see that $g'(\theta) > 0$ for $0 < \theta < \frac{\pi}{2}$. It follows that

$$0 < \theta < \tan\theta, \qquad 0 < \theta < \frac{\pi}{2},$$

and thus

$$\frac{1}{\theta} - \cot\theta = \frac{1}{\theta} - \frac{1}{\tan\theta} > 0, \qquad 0 < \theta < \frac{\pi}{2}. \tag{2.175}$$

Let us now return to (2.169) and write

$$\Lambda_n(T) = \frac{1}{n+1}\sum_{i=0}^{n} \cot\frac{\theta_i}{2} = \Sigma_n - \Sigma_n^*, \tag{2.176}$$

say, where

$$\Sigma_n = \frac{1}{n+1}\sum_{i=0}^{n}\frac{2}{\theta_i} \qquad \text{and} \qquad \Sigma_n^* = \frac{1}{n+1}\sum_{i=0}^{n}\left(\frac{2}{\theta_i} - \cot\frac{\theta_i}{2}\right). \tag{2.177}$$

The inequality in (2.175) implies that $\Sigma_n^* > 0$, and it follows from (2.176) that

$$\Lambda_n(T) < \Sigma_n = \frac{1}{n+1}\sum_{i=0}^{n}\frac{2}{\psi_i}, \tag{2.178}$$

where

$$\psi_i = \theta_{n-i} = \frac{(2i+1)\pi}{2n+2},$$

giving the simple inequality

$$\Lambda_n(T) < \frac{4}{\pi}\sum_{i=0}^{n}\frac{1}{2i+1}.$$

On writing

$$S_n = \sum_{i=1}^{n}\frac{1}{i},$$

and using the result in Problem 2.5.5, we see that

$$\Lambda_n(T) < \frac{4}{\pi}(S_{2n+2} - \frac{1}{2}S_{n+1}) < \frac{4}{\pi}\left(1 + \log(2n+2) - \frac{1}{2}\log(n+2)\right).$$

Then, on applying the inequality in Problem 2.5.6, we can deduce that

$$\Lambda_n(T) < \frac{2}{\pi}\log n + \frac{4}{\pi}\left(1 + \frac{1}{2}\log\frac{16}{3}\right),$$

and thus

$$\Lambda_n(T) < \frac{2}{\pi}\log n + 3. \tag{2.179}$$

Having obtained this inequality for the Lebesgue constant $\Lambda_n(T)$, let us estimate how much we "gave away" when we discarded the positive quantity Σ_n^* from (2.176) to obtain the inequality in (2.178). An inspection of the expression for Σ_n^* in (2.177) reveals that

$$\lim_{n\to\infty} \Sigma_n^* = \frac{2}{\pi}\int_0^{\pi/2}\left(\frac{1}{\theta} - \cot\theta\right)d\theta = \frac{2}{\pi}\log(\pi/2) \approx 0.287. \tag{2.180}$$

In fact, Σ_n^* is the quantity obtained by applying the midpoint rule in composite form with $n+1$ subintervals to estimate the integral given as the limit of the sequence (Σ_n^*) in (2.180). (See Section 3.1.) A more detailed analysis can confirm what is suggested by the foregoing material, that Σ_n^* is smaller than the error incurred in estimating Σ_n.

Let T' denote the infinite triangular array whose $(n+1)$th row consists of the extreme points of the Chebyshev polynomial T_n, and thus

$$x_i^{(n)} = \cos\phi_{n-i}, \quad \text{where} \quad \phi_i = \cos\frac{i}{n}, \quad 0 \le i \le n. \tag{2.181}$$

Intuitively, in view of Theorem 2.5.4, one would expect the Lebesgue constants $\Lambda_n(T')$ to be small, since the abscissas in T' are distributed in a similar fashion to those in T, *and* each row of the array T' contains the endpoints ± 1. Using the methods employed in the proof of Theorem 2.5.6, it is not hard to show that

$$\lambda_n(T';x) = \frac{|\sin n\theta|}{2n}\sum_{i=0}^{n}{}'' \left|\cot\frac{\theta+\phi_i}{2} + \cot\frac{\theta-\phi_i}{2}\right|, \tag{2.182}$$

where $x = \cos\theta$ and $\sum''$ denotes a sum whose first and last terms are halved. It is also not difficult to verify that $\Lambda_n(T') \le \Lambda_{n-1}(T)$ for all $n \ge 2$, with equality when n is odd. More precisely, we have

$$\Lambda_n(T') = \Lambda_{n-1}(T) = \frac{1}{n}\sum_{i=1}^{n}\tan\frac{(2i-1)\pi}{4n}, \quad n \text{ odd}, \tag{2.183}$$

and

$$\sum_{i=2}^{n}\tan\frac{(2i-1)\pi}{4n} < \Lambda_n(T') < \frac{1}{n}\sum_{i=1}^{n}\tan\frac{(2i-1)\pi}{4n}, \quad n \text{ even}. \tag{2.184}$$

n	$\Lambda_n(T)$	$\Lambda_n(T')$	$\Lambda_n(T^*)$
1	1.414	1.000	1.000
2	1.667	1.250	1.250
3	1.848	1.667	1.430
4	1.989	1.799	1.570
5	2.104	1.989	1.685
6	2.202	2.083	1.783
7	2.287	2.202	1.867
8	2.362	2.275	1.942
9	2.429	2.362	2.008
10	2.489	2.421	2.069
20	2.901	2.868	2.479
50	3.466	3.453	3.043
100	3.901	3.894	3.478
200	4.339	4.336	3.916
500	4.920	4.919	4.497
1000	5.361	5.360	4.937

TABLE 2.1. Comparison of the Lebesgue constants $\Lambda_n(T)$, $\Lambda_n(T')$, and $\Lambda_n(T^*)$.

Let T^* denote the infinite triangular array that is derived from T, the array whose rows consist of the zeros of the Chebyshev polynomials, by dividing the numbers in the $(n+1)$th row of T by $\cos(\pi/(2n+2)$, for $n \geq 1$. To be definite, we will choose the sole number in the first row of T^* as zero. We call the numbers in T^*, from the second row onwards, the *stretched* zeros of the Chebyshev polynomials. Thus, if the numbers in the $(n+1)$th row of T^* are denoted by $\bar{x}_i^{(n)}$, we see that

$$\bar{x}_0^{(n)} = -1 \quad \text{and} \quad \bar{x}_0^{(n)} = 1.$$

We know from Theorem 2.5.4 that we must have $\Lambda_n(T^*) \leq \Lambda_n(T)$, for $n \geq 1$, and Table 2.1 strongly suggests that

$$\Lambda_n(T) < \Lambda_n(T') < \Lambda_n(T^*)$$

for $n > 2$.

Luttmann and Rivlin [36] show that for every triangular array X whose abscissas satisfy

$$-1 = x_0^{(n)} < x_1^{(n)} < \cdots < x_n^{(n)} = 1, \quad n \geq 1,$$

each Lebesgue function $\lambda_n(X;x)$ has a single maximum

$$M_{j,n}(X) = \max_{x_{j-1}^{(n)} \leq x \leq x_j^{(n)}} \lambda_n(X;x), \quad 1 \leq j \leq n,$$

on each of the n subintervals $\left[x_{j-1}^{(n)}, x_j^{(n)}\right]$. As early as 1931 Bernstein [4] conjectured that if there existed an array X for which

$$M_{1,n}(X) = M_{2,n}(X) = \cdots = M_{n,n}(X), \quad n \geq 2, \tag{2.185}$$

then the array X would be optimal. Subsequently, in 1958, Paul Erdős [17] further conjectured that there is a unique array X^* for which (2.185) holds, and that for *any* array X,

$$\min_{1\leq j\leq n} M_{j,n}(X) \leq \Lambda_n(X^*). \tag{2.186}$$

In 1978 two papers were published consecutively in the same issue of the *Journal of Approximation Theory*, one by T. A. Kilgore [29] and the other by C. de Boor and A. Pinkus [12], in which these conjectures were proved. It was shown by Lev Brutman [6] that there is little variation in the n numbers $M_{j,n}(T^*)$, since

$$\Lambda_n(T^*) \leq \min_{1\leq j\leq n} M_{j,n}(T^*) + 0.201,$$

and thus

$$\Lambda_n(T^*) \leq \Lambda_n(X^*) + 0.201.$$

We may conclude that even if we do not know the optimal array X^* explicitly, it suffices for all practical purposes to use the array of stretched Chebyshev zeros T^*.

Example 2.5.3 Let us find the Lebesgue function $\lambda_3(x)$ on $[-1, 1]$ based on the set of interpolating abscissas $\{-1, -t, t, 1\}$, where $0 < t < 1$. With a little simplification, we find that

$$\lambda_3(x) = \frac{|x^2 - t^2|}{1 - t^2} + \frac{(1 - x^2)|t - x|}{2t(1 - t^2)} + \frac{(1 - x^2)|t + x|}{2t(1 - t^2)},$$

so that

$$\lambda_3(x) = \begin{cases} \dfrac{1 + t^2 - 2x^2}{1 - t^2}, & 0 \leq x \leq t, \\[2ex] \dfrac{-t^3 + x + tx^2 - x^3}{t(1 - t^2)}, & t < x \leq 1, \end{cases}$$

and $\lambda_3(x)$ is *even* on $[-1, 1]$. It is obvious that $\lambda_3(x)$ has a local maximum value at $x = 0$, say $M(t)$, and that

$$M(t) = \frac{1 + t^2}{1 - t^2}.$$

With a little more effort, we find that $\lambda_3(x)$ has local maximum values, say $M^*(t)$, at $x = \pm x^*(t)$, where

$$x^*(t) = \frac{1}{3}\left[t + (t^2 + 3)^{1/2}\right] \quad \text{for} \quad 0 < t < 1.$$

We can verify that $t < x^*(t) < 1$, and that

$$M^*(t) = \frac{9t - 25t^3 + (t^2+3)^{1/2}(6+2t^2)}{27t(1-t^2)}.$$

On the interval $0 < t < 1$, $M(t)$ increases monotonically and $M^*(t)$ decreases monotonically, and we find that

$$M(t) = M^*(t) \approx 1.423 \quad \text{for} \quad t \approx 0.4178.$$

Thus, as we see from the line above and from Table 2.1,

$$\Lambda_3(X^*) \approx 1.423 < \Lambda_3(T^*) \approx 1.430,$$

where X^* and T^* denote respectively the optimal array and the array of stretched Chebyshev zeros. ■

Problem 2.5.1 Show that the Lebesgue function $\lambda_1(x)$ on the interval $[-1,1]$ for interpolation on the abscissas $\pm t$, where $0 < t \le 1$, is given by

$$\lambda_1(x) = \begin{cases} -x/t, & -1 \le x \le -t, \\ 1, & -t < x \le t, \\ x/t, & t < x \le 1, \end{cases}$$

and so verify the value given for $\Lambda_1(T)$ in Table 2.1.

Problem 2.5.2 Write

$$\cot\frac{\theta+\theta_i}{2} - \cot\frac{\theta-\theta_i}{2} = \frac{\sin\frac{\theta-\theta_i}{2}\cos\frac{\theta+\theta_i}{2} - \sin\frac{\theta+\theta_i}{2}\cos\frac{\theta-\theta_i}{2}}{\sin\frac{\theta+\theta_i}{2}\sin\frac{\theta-\theta_i}{2}},$$

and hence show that

$$\cot\frac{\theta+\theta_i}{2} - \cot\frac{\theta-\theta_i}{2} = \frac{2\sin\theta_i}{\cos\theta - \cos\theta_i}.$$

Problem 2.5.3 Following Rivlin [48], let us write

$$c_i^+ = \cot\frac{\theta+\phi_i}{2} \quad \text{and} \quad c_i^- = \cot\frac{\theta-\phi_i}{2},$$

where $\phi_i = (2i+1)\pi/(2n+2)$, for $0 \le i \le n$. Show that if θ is replaced by $\theta - k\pi/(n+1)$, where $1 \le k \le n+1$, we have

$$c_i^+ \to \begin{cases} c_{k-i-1}^-, & 0 \le i \le k-1, \\ \\ c_{i-k}^+, & k \le i \le n, \end{cases}$$

and

$$c_i^- \to \begin{cases} c_{i+k}^-, & 0 \le i \le n-k, \\ \\ c_{2n+1-i-k}^+, & n-k+1 \le i \le n. \end{cases}$$

Verify that as a result of the mapping $\theta \mapsto \theta - k\pi/(n+1)$, the *set* of $2n+2$ cotangents c_i^+ and c_i^- is mapped into itself.

To justify the above result for c_i^- with $i \ge n-k+1$, first show that

$$\cot(\phi + \tfrac{\pi}{2}) = \cot(\phi - \tfrac{\pi}{2}) = -\tan\phi.$$

Then write $\phi_i = \pi + \alpha_i$ and show that

$$c_i^- = \cot\frac{\theta - \pi - \alpha_i}{2} = \cot\frac{\theta + \pi - \alpha_i}{2} = c_{2n+1-i-k}^+.$$

Problem 2.5.4 Show that

$$\frac{1}{2^n}\frac{T_{n+1}(x)}{x - x_i^{(n)}} = \prod_{j \ne i}(x - x_j^{(n)}),$$

where the numbers $x_i^{(n)}$, $0 \le i \le n$, are the zeros of T_{n+1}. Put $x = \cos\theta$ and hence verify that

$$\frac{\cos(n+1)\theta}{\cos\theta - \cos\theta_i} = 2^n \prod_{j \ne i}(\cos\theta - \cos\theta_j),$$

where θ_i is defined in (2.162).

Problem 2.5.5 With $S_n = \sum_{i=1}^n 1/i$, deduce from the inequalities

$$\frac{1}{i+1} < \int_i^{i+1} \frac{dx}{x} < \frac{1}{i}, \qquad i \ge 1,$$

that

$$\int_1^{n+1} \frac{dx}{x} < S_n < 1 + \int_1^n \frac{dx}{x},$$

and hence

$$\log(n+1) < S_n < 1 + \log n, \qquad n \ge 1.$$

Problem 2.5.6 Verify that

$$\frac{(n+1)^2}{n(n+2)} = 1 + \frac{1}{n(n+2)}$$

decreases with n, and thus show that

$$\log(n+1) - \frac{1}{2}\log(n+2) = \frac{1}{2}\log n + \frac{1}{2}\log\frac{(n+1)^2}{n(n+2)} \le \frac{1}{2}\log n + \frac{1}{2}\log\frac{4}{3},$$

for all $n \ge 1$, with equality only for $n = 1$.

Problem 2.5.7 Show that $\Lambda_n(T)$, given by (2.169), may be expressed in the form

$$\Lambda_n(T) = \frac{1}{n+1} \sum_{i=0}^{n} \tan \frac{\theta_i}{2},$$

where θ_i is defined by (2.162).

2.6 The Modulus of Continuity

Definition 2.6.1 Given any function $f \in C[a,b]$, we define an associated function $\omega \in C[a,b]$ as

$$\omega(\delta) = \omega(f;[a,b];\delta) = \sup_{|x_1 - x_2| \le \delta} |f(x_1) - f(x_2)|. \tag{2.187}$$

We call $\omega(f;[a,b];\delta)$ the *modulus of continuity* of the function f. ■

We have written "sup" for *supremum*, meaning the least upper bound. We express the modulus of continuity in the simpler form $\omega(\delta)$ when it is clear which function f and interval $[a,b]$ are involved. It is not difficult to verify that $\omega \in C[a,b]$, given that $f \in C[a,b]$.

Example 2.6.1 For convenience, let us take the interval $[a,b]$ to be $[0,1]$. It is obvious from Definition 2.6.1 that

$$\omega(1;[0,1];\delta) = 0 \quad \text{and} \quad \omega(\delta;[0,1];\delta) = \delta.$$

To evaluate $\omega(\delta^2;\delta)$, let us write

$$x_1^2 - x_2^2 = (x_1 - x_2)(x_1 + x_2).$$

Thus, if $|x_1 - x_2| \le \delta$, we have

$$|x_1^2 - x_2^2| \le \delta |x_1 + x_2|,$$

and the right side of the above inequality is greatest when one of x_1, x_2 is 1 and the other is $1 - \delta$. It is then clear that

$$\omega(\delta^2;[0,1];\delta) = \delta(2 - \delta) = 1 - (1 - \delta)^2.$$

More generally, it is not hard to see that

$$\omega(\delta^n;[0,1];\delta) = 1 - (1 - \delta)^n$$

for all integers $n \ge 0$. ■

The above example helps give us some familiarity with the modulus of continuity, although the results obtained in it are of little intrinsic importance. It is not difficult to justify the following more substantial properties of the modulus of continuity.

Theorem 2.6.1 If $0 < \delta_1 \leq \delta_2$, then $\omega(\delta_1) \leq \omega(\delta_2)$.

Theorem 2.6.2 A function f is uniformly continuous on the interval $[a, b]$ if and only if

$$\lim_{\delta \to 0} \omega(\delta) = 0. \qquad \blacksquare$$

We conclude this chapter with the statement of two important results due to Dunham Jackson (1888–1946) that express the minimax error for a function f in terms of moduli of continuity.

Theorem 2.6.3 Let $f \in C[-1, 1]$ and let

$$E_n(f) = \|f - p\|_\infty, \tag{2.188}$$

where $\| \cdot \|_\infty$ denotes the maximum norm on $[-1, 1]$, and $p \in P_n$ is the minimax approximation for $f \in C[a, b]$. Then

$$E_n(f) \leq 6\omega\left(\frac{1}{n}\right). \qquad \blacksquare \tag{2.189}$$

The second result of Jackson that we cite is applicable to functions that belong to $C^k[a, b]$, and gives an inequality that relates $E_n(f)$ to the modulus of continuity of the kth derivative of f.

Theorem 2.6.4 If $f \in C^k[-1, 1]$, then

$$E_n(f) \leq \frac{c}{n^k}\, \omega_k\left(\frac{1}{n-k}\right) \tag{2.190}$$

for $n > k$, where ω_k is the modulus of continuity of $f^{(k)}$, and

$$c = \frac{6^{k+1} e^k}{1+k}. \qquad \blacksquare$$

For proofs of these two theorems of Jackson, see Rivlin [48].

In addition to the modulus of continuity, there are other moduli that measure the "smoothness" of a function. These include moduli concerned with kth differences of a function. See the text by Sendov and Popov [51].

Problem 2.6.1 Show that

$$\omega(e^\delta; [0, 1]; \delta) = e - e^{1-\delta}.$$

Problem 2.6.2 Verify that for the sine function on the interval $[0, \pi/2]$, we have

$$\omega(\delta) = \sin \delta.$$

Find a class of functions f and intervals $[a, b]$ for which

$$\omega(f; [a, b]; \delta) = f(\delta).$$

Problem 2.6.3 If $|f'(x)| \leq M$ for $a \leq x \leq b$, show that

$$E_n(f) \leq \frac{6M}{n},$$

where $E_n(f)$ is defined in (2.188).

3
Numerical Integration

3.1 Interpolatory Rules

Every student who takes a calculus course learns that differentiation and integration are inverse processes, and soon discovers that integration presents more difficulties than differentiation. This is because the derivatives of compositions of the standard functions, which traditionally include the polynomials, the rational functions, functions of the form x^α with α real, the circular and inverse circular functions, the exponential and logarithmic functions, the hyperbolic and inverse hyperbolic functions, are themselves compositions of the standard functions. On the other hand, the indefinite integral of a composition of standard functions does not necessarily belong to this class. For example,

$$F(x) = \int_0^x e^{t^2} dt \tag{3.1}$$

cannot be expressed as a finite composition of standard functions.

This is why we need numerical integration processes, which give us numerical approximations to integrals even in cases where we know very little about the integrand, that is, the function being integrated. However, in an individual case where we know more about the integrand, we can often find a better method. For example, if we wished to estimate the integral in (3.1) for a value of x in the interval $[0, 1]$, we could begin by integrating the Maclaurin series for the integrand term by term.

The commonest numerical integration processes are based on interpolation. Let us replace an integrand f by its interpolating polynomial p_n, based on the abscissas x_0, $x_1, \ldots, x_n$, and integrate over $[a, b]$. We obtain

$$\int_a^b f(x)dx \approx \int_a^b p_n(x)dx = R_n(f),$$

say, and it follows from the Lagrange form (1.10) that

$$R_n(f) = \sum_{i=0}^{n} w_i^{(n)} f_i, \tag{3.2}$$

where f_i denotes $f(x_i)$, and

$$w_i^{(n)} = \int_a^b L_i(x)dx. \tag{3.3}$$

We call R_n an *interpolatory integration rule*, and refer to the numbers $w_i^{(n)}$ as the *weights* and the numbers x_i as the *abscissas* of the rule. It is clear from (3.3) that the weights depend only on the abscissas $x_0, \ldots, x_n$, and are independent of the integrand f. Indeed, it is this very property that makes integration rules so useful. An obvious choice is to let the abscissas x_i be equally spaced, so that $x_i = x_0 + ih$, for some fixed nonzero value of h. We can then introduce a new variable s such that $x = x_0 + sh$. Then $x = x_i$ corresponds to $s = i$. If we now integrate over the interval $x_0 \le x \le x_n$, we obtain

$$w_i^{(n)} = \int_{x_0}^{x_n} \prod_{j \neq i} \frac{x - x_j}{x_i - x_j}\, dx = h \int_0^n \prod_{j \neq i} \frac{s - j}{i - j}\, ds. \tag{3.4}$$

Observe that the second integrand in (3.4) is just the fundamental polynomial corresponding to the abscissa i in the set of abscissas $\{0, 1, \ldots, n\}$. Thus, to within the multiplicative constant h, the weights $w_i^{(n)}$ depend only on i and n and are independent of the value of x_0. The rules with weights defined by (3.4) are called the closed Newton–Cotes rules, named after Isaac Newton and Roger Cotes (1682–1716). If we interpolate f at $x_1, \ldots, x_{n-1}$, and integrate over $[x_0, x_n]$, we obtain a sequence of formulas called the *open* Newton–Cotes rules. The simplest of these, corresponding to $n = 2$, is

$$\int_{x_0}^{x_2} f(x)dx \approx 2hf(x_1), \tag{3.5}$$

where $x_1 - x_0 = h$. This is called the *midpoint* rule.

Example 3.1.1 Let us evaluate the weights for the Newton–Cotes rules corresponding to $n = 1$ and $n = 2$. For $n = 1$, we obtain

$$w_0^{(1)} = h \int_0^1 \frac{s - 1}{0 - 1}\, ds = \frac{1}{2}h$$

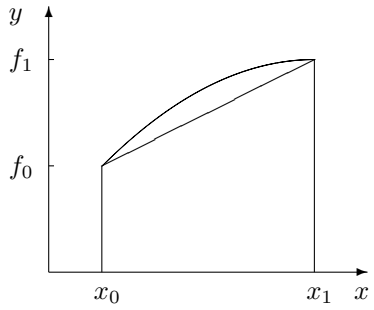

FIGURE 3.1. Area under a curve approximated by the area of a trapezoid.

and

$$w_1^{(1)} = h \int_0^1 \frac{s-0}{1-0}\, ds = \frac{1}{2}h.$$

This gives the Newton–Cotes rule of order one,

$$\int_{x_0}^{x_1} f(x)dx \approx \frac{1}{2}h(f_0 + f_1), \tag{3.6}$$

where $x_1 - x_0 = h$. This is known as the *trapezoidal rule*, because it approximates the integral by the area of the quadrilateral whose vertices are $(x_0, 0)$, (x_0, f_0), $(x_1, 0)$, and (x_1, f_1). This is a trapezoid, defined as a quadrilateral with one pair of opposite sides parallel. See Figure 3.1.

For $n = 2$, we have

$$w_0^{(2)} = h \int_0^2 \frac{(s-1)(s-2)}{(0-1)(0-2)} ds = \frac{1}{3}h,$$

and we similarly find that $w_2^{(2)} = \frac{1}{3}h$ and $w_1^{(2)} = \frac{4}{3}h$. Thus the Newton–Cotes rule of order two is

$$\int_{x_0}^{x_2} f(x)dx \approx \frac{1}{3}h(f_0 + 4f_1 + f_2), \tag{3.7}$$

where $x_2 - x_0 = 2h$. This is called Simpson's rule, after Thomas Simpson (1710–1761). ■

There is a symmetry in the weights of the trapezoidal rule, and the same holds for Simpson's rule. This symmetry holds for the general Newton–Cotes rule, where we have

$$w_i^{(n)} = w_{n-i}^{(n)}, \quad 0 \le i \le n. \tag{3.8}$$

To verify (3.8), let us begin with (3.4) and write

$$w_{n-i}^{(n)} = h \int_0^n \prod_{j \ne n-i} \frac{s-j}{n-i-j}\, ds.$$

If we make the substitution $s = n - u$, we obtain

$$w_{n-i}^{(n)} = h \int_0^n \prod_{j \neq n-i} \frac{n-u-j}{n-i-j}\, du,$$

and on writing $j = n - k$, it is clear that

$$w_{n-i}^{(n)} = h \int_0^n \prod_{k \neq i} \frac{k-u}{k-i}\, du = h \int_0^n \prod_{k \neq i} \frac{u-k}{i-k}\, du = w_i^{(n)},$$

which justifies (3.8).

If $f \in P_n$, it follows from the uniqueness of the interpolating polynomial that

$$\int_{x_0}^{x_n} f(x)dx = R_n(f) = \sum_{i=0}^{n} w_i^{(n)} f_i.$$

Thus, when the integrand is a polynomial of degree n or less, the Newton–Cotes rule of order n gives a value *equal* to that of the integral. We say that the rule is *exact* in this case. Again, let us make a linear change of variable and work with the interval $[0, n]$ in place of $[x_0, x_n]$, so that we have

$$\int_0^n g(s)ds = \sum_{i=0}^{n} w_i^{(n)} g(i), \tag{3.9}$$

whenever g is a polynomial in s of degree at most n. We can therefore determine the weights $w_i^{(n)}$ by putting $g(s)$ equal to s^j and writing down the system of linear equations

$$\sum_{i=0}^{n} w_i^{(n)}\, i^j = \frac{n^{j+1}}{j+1}, \quad 0 \leq j \leq n, \tag{3.10}$$

where the right side of the above equation is the result of integrating s^j over the interval $0 \leq s \leq n$.

Example 3.1.2 Let us derive the Newton–Cotes rule that is obtained by solving the linear system (3.10) when $n = 3$. On putting $j = 0$ and $j = 2$ in (3.10) with $n = 3$, and using the relations $w_3^{(3)} = w_0^{(3)}$ and $w_2^{(3)} = w_1^{(3)}$ obtained from (3.8), we obtain the equations

$$\begin{aligned} 2w_0^{(3)} + 2w_1^{(3)} &= 3, \\ 9w_0^{(3)} + 5w_1^{(3)} &= 9, \end{aligned}$$

whose solution is $w_0^{(3)} = \frac{3}{8}$, $w_1^{(3)} = \frac{9}{8}$. This gives the Newton–Cotes rule of order three,

$$\int_{x_0}^{x_3} f(x)dx \approx \frac{3h}{8}(f_0 + 3f_1 + 3f_2 + f_3), \tag{3.11}$$

where $x_3 - x_0 = 3h$. This is also called the three-eighths rule.

Note that we wrote down only two (corresponding to $j = 0$ and $j = 2$) of the four possible equations that we could have used, given that this rule is exact for the four monomials 1, x, x^2, and x^3. ■

As we have already deduced from (3.4), the Newton–Cotes weights $w_i^{(n)}$ for integrals over an arbitrary interval $[x_0, x_n]$ are a constant multiple of the Newton–Cotes weights of the same order for integrals over the interval $[0, n]$. Thus the weights of a given order for integrals over any two intervals differ only by a multiplicative constant. Let $w_i^{(n)}$, for $0 \le i \le n$, denote the Newton–Cotes weights for integrals over the interval $[-1, 1]$, so that

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=0}^{n} w_i^{(n)} f_i, \tag{3.12}$$

where $f_i = f(x_i)$, with $x_i = -1 + 2i/n$. The Newton–Cotes rules are exact when the integrand f is in P_n and, in particular, are exact for $f(x) = 1$. It then follows from (3.12) with $f(x) = 1$ that

$$\sum_{i=0}^{n} w_i^{(n)} = 2, \tag{3.13}$$

and so the Newton–Cotes rule of order n for integrals over a general interval $[a, b]$ is

$$\int_a^b f(x)dx \approx \frac{1}{2}(b-a) \sum_{i=0}^{n} w_i^{(n)} f_i,$$

where $f_i = f(x_i)$, with $x_i = a + i(b-a)/n$, and the weights $w_i^{(n)}$ satisfy (3.13). On the interval $[-1, 1]$, the Newton–Cotes abscissas are symmetrically placed with respect to the origin, with one abscissa at $x = 0$ when n is even. For we have

$$x_i = -1 + \frac{2i}{n} = \frac{1}{n}(2i - n),$$

and so

$$x_{n-i} = \frac{1}{n}(-2i + n) = -x_i.$$

It thus follows from (3.12) that each Newton–Cotes rule is exact for every odd function, that is, for every function f such that $f(-x) = -f(x)$, for then both sides of the approximate equality (3.12) are zero. In particular, when n is even, (3.12) is exact for the odd-order monomial x^{n+1}. Thus the Newton–Cotes rule of order $2n$ is exact not just for all integrands $f \in P_{2n}$, but for all $f \in P_{2n+1}$. For example, the Newton–Cotes rule of order two (Simpson's rule) and that of order three (the three-eighths rule) are both exact for $f \in P_3$.

We will now seek error terms for the Newton–Cotes rules by using the error formula of interpolation (see Theorem 1.1.2), thus obtaining an error term for the trapezoidal rule and, with a little more work, for Simpson's rule. First let us recall two results from analysis, both concerned with continuity, which we now state as Theorems 3.1.1 and 3.1.2. The first is the intuitively obvious result that a continuous function attains every value between its minimum and maximum values, and the second is the *mean value theorem* for integrals.

Theorem 3.1.1 If f is continuous on $[a, b]$ and y is any number such that

$$\min_{a\le x\le b} f(x) \le y \le \max_{a\le x\le b} f(x),$$

then there exists a number ξ, with $a \le \xi \le b$, such that $y = f(\xi)$. ■

Corollary 3.1.1 If F is continuous on $[a, b]$ and

$$a \le t_1 \le t_2 \le \cdots \le t_N \le b,$$

then there exists a number ξ in (a, b) such that

$$F(\xi) = \frac{1}{N}(F(t_1) + F(t_2) + \cdots + F(t_N)).$$

Proof of Corollary. Since the inequalities

$$\min_{a\le x\le b} F(x) \le F(t_r) \le \max_{a\le x\le b} F(x)$$

hold for $1 \le r \le N$, the same inequalities hold for the mean of the N numbers $F(t_r)$, and the corollary follows from Theorem 3.1.1. ■

Theorem 3.1.2 If F is continuous, and G is integrable and is nonnegative on $[a, b]$, then there exists a number ξ in (a, b) such that

$$\int_a^b F(x)G(x)dx = F(\xi)\int_a^b G(x)dx. \tag{3.14}$$

It is clear, on replacing G by $-G$, that the same result holds if $G(x) \le 0$ in (a, b). Proofs of Theorems 3.1.1 and 3.1.2 may be found in any text on elementary analysis. ■

To derive the error term for the trapezoidal rule, let us choose $n = 1$ in (1.25) and integrate over $[x_0, x_1]$, assuming that f'' is continuous on $[x_0, x_1]$. We obtain

$$\int_{x_0}^{x_1} f(x)dx - \frac{1}{2}h(f_0 + f_1) = E_T(f),$$

say, where the error term $E_T(f)$ is given by

$$E_T(f) = \frac{1}{2}\int_{x_0}^{x_1}(x-x_0)(x-x_1)f''(\xi_x)dx. \tag{3.15}$$

It follows from the error formula (1.25) that $f''(\xi_x)$ is a continuous function of x, and we also note that $(x-x_0)(x-x_1) \le 0$ on $[x_0, x_1]$. Thus we may apply Theorem 3.1.2, to give

$$E_T(f) = \frac{1}{2}f''(\xi_1)\int_{x_0}^{x_1}(x-x_0)(x-x_1)dx,$$

where $x_0 < \xi_1 < x_1$. On making the change of variable $x = x_0 + sh$, we obtain

$$E_T(f) = \frac{h^3}{2}f''(\xi)\int_0^1 s(s-1)ds = -\frac{h^3}{12}f''(\xi_1),$$

so that the error of the trapezoidal rule satisfies

$$\int_{x_0}^{x_1} f(x)dx - \frac{1}{2}h(f_0+f_1) = -\frac{h^3}{12}f''(\xi_1), \tag{3.16}$$

where $x_0 < \xi_1 < x_1$.

In practice, integration rules are used in *composite* form, where the interval of integration is split into a number of subintervals and the same basic rule is applied to each subinterval. Thus, if we write

$$\int_{x_0}^{x_N} f(x)dx = \sum_{r=1}^{N}\int_{x_{r-1}}^{x_r} f(x)dx$$

and apply the trapezoidal rule (3.6) to each of the integrals in the latter sum, we obtain

$$\int_{x_0}^{x_N} f(x)dx \approx \frac{1}{2}h(f_0 + 2f_1 + 2f_2 + \cdots + 2f_{N-1} + f_N) = T_N(f), \tag{3.17}$$

say. This is the composite form of the trapezoidal rule. It is equivalent to the sum of N numbers, each denoting the area of a trapezoid. Thus $T_N(f)$ is the integral of the function defined by the polygonal arc that connects the $N+1$ points (x_0, f_0), $(x_1, f_1), \ldots, (x_N, f_N)$. To obtain an error term for this composite rule, we can adapt (3.16) to give the basic trapezoidal rule plus error term for a general interval $[x_{r-1}, x_r]$,

$$\int_{x_{r-1}}^{x_r} f(x)dx - \frac{1}{2}h(f_{r-1}+f_r) = -\frac{h^3}{12}f''(\xi_r),$$

where $x_{r-1} < \xi_r < x_r$. Then, on summing this from $r = 1$ to N, we obtain

$$\int_{x_0}^{x_N} f(x)dx - T_N(f) = -\frac{h^3}{12}\sum_{r=1}^{N} f''(\xi_r).$$

Since f'' is continuous, it follows from Corollary 3.1.1 that

$$\int_{x_0}^{x_N} f(x)dx - T_N(f) = -\frac{Nh^3}{12}f''(\xi),$$

where $x_0 < \xi < x_N$. On writing $x_j = a + jh$, for $0 \le j \le N$, and putting $x_N = b$, so that $x_0 = a$ and $b - a = Nh$, we may express the composite trapezoidal rule plus error term in the form

$$\int_a^b f(x)dx - T_N(f) = -\frac{1}{12}(b-a)h^2 f''(\xi), \tag{3.18}$$

where $a < \xi < b$. Notice that if f'' is positive on $[a, b]$, the error term (3.18) shows that the trapezoidal approximant $T_N(f)$ is greater than the value of the integral. This is in agreement with what we should expect from a geometrical argument. For if $f''(x) > 0$, the function f is convex, and the polygonal arc that connects the points (x_0, f_0), $(x_1, f_1), \ldots, (x_N, f_N)$ lies *above* the graph of f, and consequently, the area under this polygonal arc, $T_N(f)$, is greater than the area under the graph of f. Likewise, if f'' is negative, then f is concave, the polygonal arc lies *below* the graph of f, and $T_N(f)$ underestimates the value of the integral of f.

There is a simple modification of the composite trapezoidal rule, in which $T_N(f)$ is replaced by $T'_N(f)$, defined by

$$T'_N(f) = T_N(f) - \frac{h^2}{12}(f'(b) - f'(a)), \tag{3.19}$$

that is valid for integrands f whose first derivative exists. The rule $T'_N(f)$ is called the trapezoidal rule with end correction. It is somewhat surprising that (see Problem 3.1.9) by making such a simple change to the trapezoidal rule, involving the values of the derivative of the integrand at the endpoints a and b only, we obtain a rule that, like Simpson's rule, is exact for $f \in P_3$.

Example 3.1.3 Let us illustrate our findings on the composite trapezoidal rule with the function $f(x) = e^{x^2}$ on $[0, 1]$. This is a useful test integral, because we can easily estimate it by integrating the Maclaurin series for e^{x^2} term by term, giving

$$\int_0^1 e^{x^2}dx = \int_0^1 \left(\sum_{r=0}^{\infty} \frac{x^{2r}}{r!}\right) dx = \sum_{r=0}^{\infty} \frac{1}{(2r+1)r!} \approx 1.462652. \tag{3.20}$$

In the table below, the values of $T_N(f)$ are rounded to four places after the decimal point. For comparison later with Simpson's rule (see Example 3.1.4), we also give the corresponding results for $T'_N(f)$, the composite trapezoidal rule with end correction, defined in (3.19).

N	2	4	10	20
$T_N(f)$	1.5716	1.4907	1.4672	1.4638
$T'_N(f)$	1.4583	1.4624	1.4626	1.4627

The values of $T_N(f)$ are all larger than the value of the integral, which is consistent (see (3.18)) with the fact that f'' is positive, since

$$f''(x) = \frac{d^2}{dx^2}e^{x^2} = (2 + 4x^2)e^{x^2} \geq 2$$

on $[0, 1]$. Since f'' is monotonically increasing, we have

$$2 \leq f''(x) \leq 6e$$

on $[0, 1]$, and so from (3.18) the error in $T_N(f)$ lies between $1/(6N^2)$ and $e/(2N^2)$. Thus, for example, N would have to be of the order of 1000 to estimate the integral to six decimal places. ■

We will now derive an error term for Simpson's rule. We begin by writing down (1.25) with $n = 3$,

$$f(x) - p_3(x) = \frac{1}{24}(x - x_0)(x - x_1)(x - x_2)(x - x_3)f^{(4)}(\xi_x), \tag{3.21}$$

and we will assume that $f^{(4)}$ is continuous. As we observed above, Simpson's rule is exact if the integrand belongs to P_3, and thus

$$\int_{x_0}^{x_2} p_3(x)dx = \frac{1}{3}h(p_3(x_0) + 4p_3(x_1) + p_3(x_2)). \tag{3.22}$$

Since p_3 interpolates f at x_0, x_1, x_2, and x_3, we deduce from (3.22) that

$$\int_{x_0}^{x_2} p_3(x)dx = \frac{1}{3}h(f_0 + 4f_1 + f_2).$$

If we now integrate (3.21) over the interval $[x_0, x_2]$, we obtain

$$\int_{x_0}^{x_2} f(x)dx - \frac{1}{3}h(f_0 + 4f_1 + f_2) = E_S(f), \tag{3.23}$$

say, where

$$E_S(f) = \frac{1}{24}\int_{x_0}^{x_2} (x - x_0)(x - x_1)(x - x_2)(x - x_3)f^{(4)}(\xi_x)dx. \tag{3.24}$$

The polynomial $(x - x_0)(x - x_1)(x - x_2)(x - x_3)$ obviously changes sign in the interval $[x_0, x_3]$, at $x = x_1$ and at $x = x_2$, and so we cannot apply Theorem 3.1.2 to simplify (3.24) further, as we did with (3.15) in obtaining the error of the trapezoidal rule.

We therefore turn to (1.32), the alternative error term for the interpolating polynomial that involves a divided difference rather than an $(n + 1)$th derivative. Let us use (1.32) with $n = 2$,

$$f(x) - p_2(x) = (x - x_0)(x - x_1)(x - x_2)f[x, x_0, x_1, x_2],$$

and integrate over the interval $[x_0, x_2]$. The integral of the interpolating polynomial p_2 just gives Simpson's rule, as we obtained by integrating p_3 in (3.21). Thus we obtain an alternative expression for the error of Simpson's rule, namely,

$$E_S(f) = \int_{x_0}^{x_2} (x - x_0)(x - x_1)(x - x_2) f[x, x_0, x_1, x_2] dx. \tag{3.25}$$

In preparation for using integration by parts on the latter integral, we find that

$$\frac{d}{dx}(x - x_0)^2(x - x_2)^2 = 2(x - x_0)(x - x_2)\,\frac{d}{dx}(x - x_0)(x - x_2),$$

and since

$$\frac{d}{dx}(x - x_0)(x - x_2) = 2x - x_0 - x_2 = 2(x - x_1),$$

we obtain

$$\frac{d}{dx}(x - x_0)^2(x - x_2)^2 = 4(x - x_0)(x - x_1)(x - x_2).$$

Thus, using integration by parts on the integral in (3.25), we find that

$$E_S(f) = -\frac{1}{4}\int_{x_0}^{x_2} (x - x_0)^2(x - x_2)^2 \frac{d}{dx} f[x, x_0, x_1, x_2] dx. \tag{3.26}$$

It is encouraging that we can now apply Theorem 3.1.2, since the integrand in (3.26) is of the form $F(x)G(x)$, where $G(x) = (x-x_0)^2(x-x_2)^2$ does not change sign on the interval of integration. The other factor in the integrand is

$$F(x) = \frac{d}{dx} f[x, x_0, x_1, x_2] = \lim_{\delta x \to 0} \frac{f[x + \delta x, x_0, x_1, x_2] - f[x, x_0, x_1, x_2]}{\delta x},$$

and using the recurrence relation (1.22), the divided differences simplify to give

$$F(x) = \lim_{\delta x \to 0} f[x, x + \delta x, x_0, x_1, x_2] = f[x, x, x_0, x_1, x_2].$$

Thus, on applying Theorem 3.1.2 to (3.26), we obtain

$$E_S(f) = -\frac{1}{4} f[\eta, \eta, x_0, x_1, x_2] \int_{x_0}^{x_2} (x - x_0)^2(x - x_2)^2 dx,$$

say, where $x_0 < \eta < x_2$. Finally, we can replace the fourth-order divided difference by a fourth-order derivative, using (1.33), and integrate the latter integral by means of the substitution $x = x_0 + sh$, to give

$$E_S(f) = -\frac{1}{4} h^5 \frac{f^{(4)}(\xi)}{4!} \int_0^2 s^2(s - 2)^2 ds = -\frac{h^5}{90} f^{(4)}(\xi). \tag{3.27}$$

By pursuing a similar method to that used above for Simpson's rule, we can derive an error term for the midpoint rule (3.5) of the form

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + \frac{h^3}{3}f''(\xi), \tag{3.28}$$

where $\xi \in (x_0, x_2)$. For more details, see Problem 3.1.3.

We can now write down a composite form of Simpson's rule. Beginning with

$$\int_{x_0}^{x_{2N}} f(x)dx = \sum_{r=1}^{N} \int_{x_{2r-2}}^{x_{2r}} f(x)dx$$

and applying Simpson's rule (3.7) to each of the integrals in the latter sum, we obtain the composite form of Simpson's rule,

$$\int_{x_0}^{x_{2N}} f(x)dx \approx S_N, \tag{3.29}$$

where

$$S_N = \frac{1}{3}h(f_0 + 4f_1 + 2f_2 + \cdots + 2f_{2N-2} + 4f_{2N-1} + f_{2N}), \tag{3.30}$$

and the values of f_j are multiplied by 4 and 2 alternately, for $1 \le j \le 2N-1$. To obtain an error term for the composite form of Simpson's rule, we begin with (3.23) and (3.27) and adapt these formulas to the interval $[x_{2r-2}, x_{2r}]$, to give

$$\int_{x_{2r-2}}^{x_{2r}} f(x)dx - \frac{1}{3}h(f_{2r-2} + 4f_{2r-1} + f_{2r}) = -\frac{h^5}{90}f^{(4)}(\xi_r),$$

where $x_{2r-2} < \xi_r < x_{2r}$. We now sum from $r = 1$ to N, so that

$$\int_{x_0}^{x_{2N}} f(x)dx - S_N = -\frac{h^5}{90}\sum_{r=1}^{N} f^{(4)}(\xi_r),$$

and assuming continuity of $f^{(4)}$, the application of Corollary 3.1.1 yields

$$\int_{x_0}^{x_{2N}} f(x)dx - S_N = -\frac{Nh^5}{90}f^{(4)}(\xi),$$

where $x_0 < \xi < x_{2N}$. Finally, writing $x_j = a + jh$, for $0 \le j \le 2N$, and putting $x_{2N} = b$, so that $x_0 = a$ and $b - a = 2Nh$, we may express the composite form of Simpson's plus error term as

$$\int_a^b f(x)dx - S_N = -\frac{1}{180}(b-a)h^4 f^{(4)}(\xi), \tag{3.31}$$

where $a < \xi < b$.

Example 3.1.4 We will use the function $f(x) = e^{x^2}$ on $[0,1]$ to illustrate the results given above on Simpson's rule, for comparison with our findings in Example 3.1.3 on the trapezoidal rule. In the table below, the values of $S_N(f)$ are rounded to four places after the decimal point. In making a comparison of the composite rules $T_N(f)$ and $S_N(f)$, we need to take into account the fact that these rules require respectively $N+1$ and $2N+1$ evaluations of the integrand f. Thus corresponding entries in the table below and the table in Example 3.1.3 are computed using the same number of values of the integrand, namely, 3, 5, 11, and 21, respectively.

N	1	2	5	10
$S_N(f)$	1.4757	1.4637	1.4627	1.4627

We see from the above table and the table in Example 3.1.3 that the results from Simpson's rule are much more accurate than those from the trapezoidal rule, and are comparable with those from the trapezoidal rule with end correction. The error term (3.31) requires $f^{(4)}$, and we find that

$$\frac{d^4}{dx^4}e^{x^2} = (12 + 48x^2 + 16x^4)e^{x^2}.$$

This fourth derivative is positive on $[0,1]$, and thus the above estimates obtained by Simpson's rule are all greater than the value of the integral. Since $f^{(4)}$ is monotonically increasing,

$$12 \le f^{(4)}(x) \le 76e$$

on $[0,1]$, and so from (3.31) the error in $S_N(f)$ satisfies

$$\frac{1}{240N^4} \le \left| \int_a^b f(x)dx - S_N \right| \le \frac{19e}{720N^4}.$$

If we compute the numbers in the above table with greater precision, we find that $S_{20} \approx 1.462652$, and as we see from (3.20), this is correct to six decimal places. ■

It is not difficult to show that if f is integrable, for example, if f is continuous, then $T_N(f)$ converges to the value of the integral as $N \to \infty$. The same is true of $S_N(f)$, and indeed, for the composite form of any Newton–Cotes rule. Examples 3.1.3 and 3.1.4 illustrate the point that unless $f^{(4)}$ is very much larger than f'', Simpson's rule is greatly to be preferred to the trapezoidal rule. However, we will see in Section 3.2 that an adaptation of the trapezoidal rule, Romberg's method, can be extremely efficient when the integrand is many times differentiable. Finally, the following example shows that we should not *always* expect Simpson's rule to be very much better than the trapezoidal rule.

Example 3.1.5 Let us apply the trapezoidal rule and Simpson's rule to estimate the integral of $x^{1/2}$ over the interval $[0, 1]$. Since all the derivatives of the integrand are undefined (since they are infinite) at $x = 0$, the error estimates for both $T_N(f)$ and $S_N(f)$ tell us nothing, and we cannot use the trapezoidal rule with end correction. The table below gives a comparison of $T_{2N}(f)$ and $S_N(f)$, both of which require $2N + 1$ evaluations of the integrand, for this integral, whose value is $\frac{2}{3}$.

N	1	2	5	10
$T_{2N}(f)$	0.6036	0.6433	0.6605	0.6644
$S_N(f)$	0.6381	0.6565	0.6641	0.6658

The table shows that although $S_N(f)$ does give better results than $T_{2N}(f)$ for all the values of N used, there is not a dramatic difference in the performance of the two rules such as we saw for the integrand e^{x^2}. ■

We were able to use Theorem 3.1.2 directly to obtain an error term for the trapezoidal rule, and with the aid of a clever argument involving integration by parts, we were able to use it to give an error term for Simpson's rule. We need to turn to other methods to find error terms for higher-order Newton–Cotes rules. In the next chapter we will show how Peano kernel theory can be used for this purpose.

Problem 3.1.1 Replace the integrand $f(x)$ by the approximation $f(x_0)$ on $[x_0, x_1]$, and so derive the integration rule

$$\int_{x_0}^{x_1} f(x)dx \approx hf(x_0).$$

This is called the *rectangular rule.*

Problem 3.1.2 Write $f(x) = f(x_0) + (x - x_0)f'(\xi_x)$, where $\xi_x \in (x_0, x_1)$, and integrate over $[x_0, x_1]$ to give an error term for the rectangular rule.

Problem 3.1.3 To derive the error term (3.28) for the midpoint rule, begin with the expression

$$f(x) = f(x_1) + (x - x_1)f[x, x_1],$$

obtained by putting $n = 0$ in (1.32) and replacing x_0 by x_1. Then integrate $(x - x_1)f[x, x_1]$ over the interval $[x_0, x_2]$, and show that

$$\int_{x_0}^{x_2} (x - x_1)f[x, x_1]\, dx = \frac{1}{2}\int_{x_0}^{x_2} \frac{d}{dx}\{(x - x_0)(x - x_2)\}\, f[x, x_1]\, dx.$$

Complete the derivation of (3.28) by following the method that we used in deriving (3.26), using integration by parts and Theorem 3.1.2.

Problem 3.1.4 Show that Simpson's rule S_N may be expressed as the following linear combination of the two trapezoidal rules T_N and T_{2N}:

$$S_N(f) = \frac{1}{3}(4T_{2N}(f) - T_N(f)).$$

Problem 3.1.5 Verify that

$$S_N(f) = \frac{1}{3}(2T_{2N} + M_N),$$

where S_N and T_{2N} denote Simpson's rule and the trapezoidal rule in composite form, each requiring $2N + 1$ evaluations of the integrand, and M_N denotes the composite form of the midpoint rule that requires N evaluations of the integrand.

Problem 3.1.6 Let us apply the composite trapezoidal rule

$$T_2(f) = \frac{1}{2}(f(-1) + 2f(0) + f(1))$$

to estimate the integral $\int_{-1}^{1} f(x)dx$. Find an integrand f for which $T_2(f)$ is exact and for which Simpson's rule, based on the same three abscissas, is not exact.

Problem 3.1.7 Derive the open Newton–Cotes rule of the form

$$\int_0^3 f(x)dx \approx w_1 f(1) + w_2 f(2)$$

that is exact for $f \in P_3$.

Problem 3.1.8 Derive the open Newton–Cotes rule of the form

$$\int_{-2}^{2} f(x)dx \approx w_{-1} f(-1) + w_0 f(0) + w_1 f(1)$$

that is exact for $f \in P_3$.

Problem 3.1.9 Show that the integration rule

$$\int_{x_0}^{x_1} f(x)dx \approx \frac{h}{2}(f(x_0) + f(x_1)) - \frac{h^2}{12}(f'(x_1) - f'(x_0)),$$

where $h = x_1 - x_0$, which may be applied to any integrand f whose first derivative exists, is exact for $f \in P_3$. Deduce that the composite form of the above rule, the trapezoidal rule with end correction, is also exact for $f \in P_3$.

3.2 The Euler–Maclaurin Formula

The trapezoidal rule with end correction (3.19) is a special case of the Euler–Maclaurin formula, which is named after Leonhard Euler (1707–1783) and Colin Maclaurin (1698–1746). Assuming that f is sufficiently differentiable, let us write

$$\int_{-h/2}^{h/2} f(x)dx \approx \frac{h}{2}(f(-h/2)+f(h/2)) - \sum_{r=1}^{m} h^{2r}c_{2r}\left(f^{(2r-1)}(h/2)-f^{(2r-1)}(-h/2)\right), \quad (3.32)$$

for any positive integer m. First we note that (3.32) is exact when $f(x)=1$ and when $f(x)=x^{2n-1}$, for any integer $n \geq 1$. We next observe that (3.32) is exact for all $f \in P_{2m+1}$ if and only if it is exact for $f(x)=x^2, x^4, \ldots, x^{2m}$. On substituting $f(x)=x^{2s}$ into (3.32), and dividing by $h^{2s+1}/2^{2s}$, we have equality if

$$\frac{1}{2s+1} = 1 - \sum_{r=1}^{s} 2^{2r}2s(2s-1)\cdots(2s-2r+2)c_{2r}, \quad 1 \leq s \leq m. \quad (3.33)$$

Since this is a nonsingular lower triangular system, we can solve these equations by forward substitution to determine the c_{2r} uniquely, and then (3.32) will be exact for all $f \in P_{2m+1}$. Note that the coefficients c_{2r} do not depend on m. On solving these equations, we find that the first few values of the coefficients c_{2r} are

$$c_2 = \frac{1}{12},\ c_4 = -\frac{1}{720},\ c_6 = \frac{1}{30240},\ c_8 = -\frac{1}{1209600},\ c_{10} = \frac{1}{47900160}.$$

The value obtained for c_2 is that already found in (3.19), the trapezoidal rule with end correction. We now make a linear change of variable, mapping $[-h/2, h/2]$ onto $[x_j, x_{j+1}]$, where $x_j = x_0 + jh$. Under this transformation, (3.32) becomes

$$\int_{x_j}^{x_{j+1}} f(x)dx \approx \frac{h}{2}(f(x_j)+f(x_{j+1})) - \sum_{r=1}^{m} h^{2r}c_{2r}\left(f^{(2r-1)}(x_{j+1})-f^{(2r-1)}(x_j)\right), \quad (3.34)$$

and with the values of c_{2r} as derived above, this equation is exact for all $f \in P_{2m+1}$. If we evaluate (3.34) with $j=0$, $x_0=0$, $h=1$, and $f(x)=x^{2s}$, we obtain

$$\frac{1}{2s+1} = \frac{1}{2} - \sum_{r=1}^{s} 2s(2s-1)\cdots(2s-2r+2)c_{2r}, \quad (3.35)$$

giving an alternative linear system to (3.33) for determining the c_{2r}.

We now show that for all $r \geq 1$,

$$c_{2r} = \frac{B_{2r}}{(2r)!}, \tag{3.36}$$

where the B_{2r} are the Bernoulli numbers, named after Jacob Bernoulli (1654–1705). These are defined by

$$\frac{x}{e^x - 1} = \sum_{r=0}^{\infty} \frac{B_r x^r}{r!}\,. \tag{3.37}$$

Thus $B_0 = 1, B_1 = -\frac{1}{2}$, and (see Problem 3.2.1) $B_{2r-1} = 0$ for all $r > 1$. On multiplying (3.37) throughout by $e^x - 1$, we obtain

$$x = (e^x - 1)\left(1 - \frac{1}{2}x + \sum_{r=1}^{\infty} \frac{B_{2r} x^{2r}}{(2r)!}\right). \tag{3.38}$$

If we now compare the coefficient of x^{2s+1} on both sides of (3.38) and multiply throughout by $(2s)!$, we obtain the same equation as (3.35), with c_{2r} replaced by $B_{2r}/(2r)!$, thus justifying (3.36). The next few nonzero values of the Bernoulli numbers after B_1 are

$$B_2 = \frac{1}{6}, \quad B_4 = -\frac{1}{30}, \quad B_6 = \frac{1}{42}, \quad B_8 = -\frac{1}{30}, \quad B_{10} = \frac{5}{66}. \tag{3.39}$$

Now let us sum (3.34) from $j = 0$ to $N - 1$, and use (3.36) to express the coefficients c_{2r} in terms of the Bernoulli numbers. Finally, we replace x_0 by a and x_N by b, to give

$$\int_a^b f(x)dx \approx T_N(f) - \sum_{r=1}^{m} h^{2r} \frac{B_{2r}}{(2r)!}\left(f^{(2r-1)}(b) - f^{(2r-1)}(a)\right), \tag{3.40}$$

where $T_N(f)$ denotes the trapezoidal sum, defined in (3.17). We note that (3.40) is exact for all $f \in P_{2m+1}$.

If $f \in C^1[a, b]$, we can use (3.40) with the series on the right truncated after one term. As we have already noted, this is just the trapezoidal rule with end correction. If $f \in C^3[a, b]$, we can use (3.40) with the series on the right truncated after two terms, and so on. Note that although we have introduced the Euler–Maclaurin formula as a means of estimating an integral, we could equally use it to express a sum as an integral plus correction terms involving odd-order derivatives of the integrand at the endpoints of the interval of integration.

Having derived the Euler–Maclaurin formula in a constructive manner, as we have done above, we can appreciate so much more the following very elegant way of building it up by the repeated use of integration by parts.

This approach has the advantage of leading us naturally to an error term, and to a deeper understanding of the Euler–Maclaurin formula. Otherwise, it would not seem very useful, since we need to know that it has the form given by (3.40) in order to follow this route. For simplicity of presentation, we will take $x_0 = 0$ and $h = 1$. We define a sequence of functions (p_n), for $n \geq 1$ such that each $p_n(x)$ is a polynomial of degree n for $0 \leq x < 1$ and is periodic elsewhere in $(-\infty, \infty)$, with period 1, so that

$$p_n(x+1) = p_n(x), \quad -\infty < x < \infty, \tag{3.41}$$

for all $n \geq 1$. We define, for $n \geq 1$,

$$p_n(x) = \frac{1}{n!} \sum_{s=0}^{n} \binom{n}{s} B_s x^{n-s}, \quad 0 \leq x < 1, \tag{3.42}$$

and we can show that

$$p'_{n+1}(x) = p_n(x), \quad 0 \leq x < 1, \tag{3.43}$$

for all $n \geq 1$. The polynomial $p_n(x)$ is commonly written as $B_n(x)/n!$, where $B_n(x)$ is called the Bernoulli polynomial of degree n. However, we need to be careful with references to the Bernoulli polynomials, because there is more than one definition of these, as indeed there is for the Bernoulli numbers. It follows from (3.42) that

$$p_1(x) = x - \frac{1}{2}, \quad 0 \leq x < 1. \tag{3.44}$$

It is also clear from (3.42) that for all $r \geq 1$,

$$p_{2r}(0) = \frac{B_{2r}}{(2r)!} \quad \text{and} \quad p_{2r+1}(0) = 0. \tag{3.45}$$

The first few members of the sequence (p_n), following p_1, are defined by

$$\begin{aligned}
p_2(x) &= \frac{1}{2}x^2 - \frac{1}{2}x + \frac{1}{12}, \\
p_3(x) &= \frac{1}{6}x^3 - \frac{1}{4}x^2 + \frac{1}{12}x, \\
p_4(x) &= \frac{1}{24}x^4 - \frac{1}{12}x^3 + \frac{1}{24}x^2 - \frac{1}{720}, \\
p_5(x) &= \frac{1}{120}x^5 - \frac{1}{48}x^4 + \frac{1}{72}x^3 - \frac{1}{720}x,
\end{aligned}$$

for $0 \leq x < 1$, and they satisfy the periodicity condition (3.41) elsewhere in $(-\infty, \infty)$. These functions are called piecewise polynomials or splines, about which we have more to say in Chapter 6. We see that $p_1(x)$ is a "saw-tooth" function that is discontinuous at every integer value of x. However,

as we see from its above explicit form for $0 \le x < 1$ and the periodicity condition, the function p_2 is continuous on the whole real line, and an induction argument based on (3.43) shows that p_n is continuous on the whole real line for all $n \ge 2$. Indeed, such an argument reveals that these functions become increasingly smooth as n is increased, and we can deduce that $p_n \in C^{n-2}(-\infty, \infty)$ for $n \ge 2$. From the periodicity property (3.41) and the continuity of p_n for $n \ge 2$, we see from (3.45) that

$$p_{2r}(0) = p_{2r}(1) = \frac{B_{2r}}{(2r)!} \quad \text{and} \quad p_{2r+1}(0) = p_{2r+1}(1) = 0, \quad r \ge 1. \tag{3.46}$$

On $[0,1]$, the polynomial p_n is symmetric about the midpoint $x = \frac{1}{2}$, when n is even, and is antisymmetric about $x = \frac{1}{2}$, when n is odd; that is,

$$p_n(1-x) = (-1)^n p_n(x), \quad 0 \le x \le 1, \tag{3.47}$$

for all $n \ge 2$. (See Problem 3.2.4.) It follows from (3.47) that $p_{2r+1}(\frac{1}{2}) = 0$ for all $r \ge 0$.

Let us use integration by parts. Because p_1 is discontinuous, we will work with the interval $[0, 1-\epsilon]$, where $0 < \epsilon < 1$, instead of $[0,1]$, and write

$$\int_0^{1-\epsilon} p_1(x) f'(x) dx = [p_1(x) f(x)]_0^{1-\epsilon} - \int_0^{1-\epsilon} f(x) dx,$$

since $p_1'(x) = 1$ on $[0, 1-\epsilon]$. Now we let ϵ tend to zero from above, which we write as $\epsilon \to 0_+$. We have

$$\lim_{\epsilon \to 0_+} p_1(1-\epsilon) = \lim_{\epsilon \to 0_+} (1 - \epsilon - \tfrac{1}{2}) = \frac{1}{2},$$

and as $\epsilon \to 0_+$, we obtain

$$\int_0^1 p_1(x) f'(x) dx = \frac{1}{2}(f(0) + f(1)) - \int_0^1 f(x) dx.$$

It then follows from the periodicity of p_1 that

$$\int_j^{j+1} p_1(x) f'(x) dx = \frac{1}{2}(f(j) + f(j+1)) - \int_j^{j+1} f(x) dx, \tag{3.48}$$

for any integer j. Now, for any integer $n \ge 1$, we have

$$\int_j^{j+1} p_n(x) f^{(n)}(x) dx = \Big[p_{n+1}(x) f^{(n)}(x) \Big]_j^{j+1} - \int_j^{j+1} p_{n+1}(x) f^{(n+1)}(x) dx,$$

and if we write down the latter equation with n replaced by $2r-1$ and $2r$, in turn, apply (3.46), and add those two equations together, we obtain

$$\int_j^{j+1} p_{2r-1}(x) f^{(2r-1)}(x) dx = \frac{B_{2r}}{(2r)!} \Big(f^{(2r-1)}(j+1) - f^{(2r-1)}(j) \Big) + \int_j^{j+1} p_{2r+1}(x) f^{(2r+1)}(x) dx.$$

If we now use (3.48) and the latter equation with $r = 1, \ldots, m$, we obtain

$$\int_j^{j+1} f(x)dx = \frac{1}{2}(f(j) + f(j+1)) - \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}\left(f^{(2r-1)}(j+1) - f^{(2r-1)}(j)\right) - \int_j^{j+1} p_{2m+1}(x) f^{(2m+1)}(x)dx, \tag{3.49}$$

which holds for all functions f that are sufficiently differentiable. If we now sum (3.49) from $j = 0$ to $N - 1$, we obtain

$$\int_0^N f(x)dx = T_N(f) - \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}(f^{(2r-1)}(N) - f^{(2r-1)}(0)) - \int_0^N p_{2m+1}(x) f^{(2m+1)}(x)dx, \tag{3.50}$$

where, in the trapezoidal sum $T_N(f)$, the function f is evaluated at the integers $0, 1, \ldots, N$. Finally, we obtain (3.40) plus an error term from (3.50) by making the linear transformation that maps $[0, N]$ onto $[x_0, x_0 + Nh]$. An alternative derivation of the error term is given in Section 4.2.

Before leaving this topic, we note that $p_1(x)$ has the Fourier expansion

$$p_1(x) = -\sum_{n=1}^{\infty} \frac{2 \sin 2\pi nx}{2\pi n}. \tag{3.51}$$

(See Problem 3.2.5.) Let us repeatedly integrate this series formally, term by term. We note from (3.47) that each p_{2r} must have a Fourier expansion containing only cosine terms, and the expansion for each p_{2r+1} contains only sine terms. We obtain

$$p_{2r}(x) = (-1)^{r-1} \sum_{n=1}^{\infty} \frac{2 \cos 2\pi nx}{(2\pi n)^{2r}}, \tag{3.52}$$

$$p_{2r+1}(x) = (-1)^{r-1} \sum_{n=1}^{\infty} \frac{2 \sin 2\pi nx}{(2\pi n)^{2r+1}}. \tag{3.53}$$

It is clear that the Fourier expansions in (3.52) and (3.53), for $r \geq 1$, converge uniformly on $[0, 1]$ and so, by periodicity, converge uniformly everywhere. Then, by the well-known result on the uniform convergence of Fourier expansions (see, for example, Davis [10]), since for $n > 1$, p_n is continuous and is periodic with period 1 and its Fourier series is uniformly convergent, the Fourier series converges to p_n. Note that this does not hold for p_1, since

$$p_1(0) = p_1(1) = -\frac{1}{2} \quad \text{and} \quad \lim_{\epsilon \to 0_+} p_1(1 - \epsilon) = \frac{1}{2},$$

whereas its Fourier expansion assumes the value 0 at both $x = 0$ and 1.

If we put $x = 0$ in (3.52) and evaluate $p_{2r}(0)$ from (3.45), we obtain

$$\sum_{n=1}^{\infty} \frac{1}{n^{2r}} = (-1)^{r-1} 2^{2r-1} \pi^{2r} \frac{B_{2r}}{(2r)!}, \quad r \geq 1. \tag{3.54}$$

The first few such sums are

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6}, \quad \sum_{n=1}^{\infty} \frac{1}{n^4} = \frac{\pi^4}{90}, \quad \sum_{n=1}^{\infty} \frac{1}{n^6} = \frac{\pi^6}{945}, \quad \sum_{n=1}^{\infty} \frac{1}{n^8} = \frac{\pi^8}{9450}.$$

The relation (3.54) also shows us that the Bernoulli numbers B_{2r} alternate in sign, as the reader may have suspected from (3.39), and thus, from (3.45), the values of $p_{2r}(0)$ alternate in sign.

Example 3.2.1 Let us use the Euler–Maclaurin formula to compute the sum $S = \sum_{n=1}^{\infty} 1/n^2 = \pi^2/6$. Although the value of S is known from (3.54), this example will serve to demonstrate the great power of this method. We will write $S = S_{N-1} + R_N$, where

$$S_{N-1} = \sum_{n=1}^{N-1} \frac{1}{n^2} \qquad \text{and} \qquad R_N = \sum_{n=N}^{\infty} \frac{1}{n^2},$$

and use the Euler–Maclaurin formula to estimate the "tail" of the series, R_N. We begin with (3.40), letting $b \to \infty$ and setting $h = 1$ and $a = N$. Finally, we let $m \to \infty$, to give

$$R_N = \int_N^{\infty} f(x)dx + \frac{1}{2} f(N) - \sum_{r=1}^{\infty} \frac{B_{2r}}{(2r)!} f^{(2r-1)}(N),$$

where $f(x) = 1/x^2$. Note that the term $\frac{1}{2} f(N)$ is required, since the Euler–Maclaurin formula involves a "trapezoidal" sum, where the first and last terms are halved. Note that *all* the derivatives of the function $1/x^2$ exist at $x = 1$ and all tend to zero as $x \to \infty$. Since for $f(x) = 1/x^2$, we have $f^{(2r-1)}(N)/(2r)! = -1$, for all r, we obtain

$$R_N = \frac{1}{N} + \frac{1}{2N^2} + \frac{1}{6N^3} - \frac{1}{30N^5} + \frac{1}{42N^7} - \frac{1}{30N^9} + O\left(\frac{1}{N^{11}}\right).$$

With $N = 10$, for example, we estimate R_{10} from the line above and add it to the sum S_9, to give the following estimate of S, with the correct value below it:

$$S \approx 1.64493\,40668\,47,$$
$$\frac{1}{6}\pi^2 \approx 1.64493\,40668\,48.$$

Both numbers are rounded to 12 places after the decimal point. ■

There is a finite difference analogue of the Euler–Maclaurin formula, where the correction terms at the endpoints of the interval of integration are expressed in terms of differences rather than derivatives. This is Gregory's formula,

$$\int_{x_0}^{x_N} f(x)dx \approx T_N(f) - h\sum_{r=1}^{2m} a_r \left(\Delta^r f_0 + (-1)^r \nabla^r f_N\right), \tag{3.55}$$

obtained by James Gregory before Euler and Maclaurin were born. The coefficients a_r are given by

$$a_r = \int_0^1 \binom{s}{r+1} ds. \tag{3.56}$$

Let us now recast the Euler–Maclaurin formula (3.40), writing the trapezoidal sum $T_N(f)$ as $T(h)$, and writing

$$\frac{B_{2r}}{(2r)!}(f^{(2r-1)}(b) - f^{(2r-1)}(a)) = -E_{2r}.$$

Then (3.40) is now expressed in the form

$$\int_a^b f(x)dx \approx T(h) + \sum_{r=1}^{m} h^{2r} E_{2r}, \tag{3.57}$$

which emphasizes its dependence on the parameter h. We will use (3.57) to justify a numerical integration process known as Romberg's method, named after Werner Romberg (born 1909). In the account that follows, we will see that we do not need to know the values of the coefficients E_{2r} in (3.57) in order to implement Romberg's method.

To concentrate our attention on the first of the terms in the sum on the right of (3.57), let us write

$$\int_a^b f(x)dx = T(h) + h^2 E_2 + O(h^4). \tag{3.58}$$

If we now also write down the latter equation with h replaced by $h/2$, we obtain

$$\int_a^b f(x)dx = T(h/2) + \frac{1}{4}h^2 E_2 + O(h^4). \tag{3.59}$$

Let us now *eliminate* the principal error term, involving h^2, between the last two expressions: We multiply (3.59) by 4, subtract (3.58), and divide by 3 to give

$$\int_a^b f(x)dx = T^{(1)}(h) + O(h^4), \tag{3.60}$$

where

$$T^{(1)}(h) = \frac{4T(h/2) - T(h)}{3}.$$

The process of eliminating the term in h^2 is called *extrapolation to the limit*, or *Richardson extrapolation*, after L. F. Richardson (1881–1953). Note that we do not have to know the value of E_2 in order to carry out this process, nor do we have to know the values of any of the error coefficients E_{2j} in the repeated extrapolations that follow. The first extrapolation gives us nothing new, for $T^{(1)}(h)$ is just the estimate of the integral given by Simpson's rule. However, we can continue by eliminating the error term involving h^4, then that involving h^6, and so on, assuming that the value of m in (3.57) is sufficiently large. When we have eliminated the error term involving h^4, we obtain

$$\int_a^b f(x)dx = T^{(2)}(h) + O(h^6),$$

say, where

$$T^{(2)}(h) = \frac{4^2 T^{(1)}(h/2) - T^{(1)}(h)}{4^2 - 1}.$$

Note that in order to compute $T^{(2)}(h)$, we need the numbers $T(h)$, $T(h/2)$, and $T(h/4)$. In general, we compute, recursively,

$$T^{(k)}(h) = \frac{4^k T^{(k-1)}(h/2) - T^{(k-1)}(h)}{4^k - 1},$$

and we have

$$\int_a^b f(x)dx = T^{(k)}(h) + O(h^{2k+2}).$$

This is valid, provided that the integrand f is sufficiently differentiable. Romberg's method gives the accuracy of the Euler–Maclaurin formula without the need to evaluate any derivatives of the integrand.

Example 3.2.2 Let us use Romberg integration to estimate the integral of e^{x^2} over the interval $[-1, 1]$. The results are displayed in the following table.

h	$T(h)$	$T^{(1)}(h)$	$T^{(2)}(h)$	$T^{(3)}(h)$
1	1.859141	1.475731	1.462909	1.462654
$\frac{1}{2}$	1.571583	1.463711	1.462658	
$\frac{1}{4}$	1.490679	1.462723		
$\frac{1}{8}$	1.469712			

All the numbers in the above Romberg table overestimate the value of the integral, which is 1.462652, rounded to six decimal places. Compare $T^{(3)}(1) = 1.462654$, with error $2 \cdot 10^{-6}$, and the best trapezoidal approximant in the table, 1.469712, which has an error of 7 units in the third decimal place. ■

In the above example on Romberg integration, all derivatives of the integrand exist, and we obtained excellent results. We should not expect Romberg's method to be as successful when the integrand has a singularity in a low-order derivative.

Example 3.2.3 Let us use Romberg integration to estimate the integral of $x^{1/2}$ over the interval $[0, 1]$. The results are displayed in the following table.

h	$T(h)$	$T^{(1)}(h)$	$T^{(2)}(h)$	$T^{(3)}(h)$
1	0.500000	0.638071	0.657757	0.663608
$\frac{1}{2}$	0.603553	0.656526	663516	
$\frac{1}{4}$	0.643283	0.663079		
$\frac{1}{8}$	0.658130			

This is only a test example, for we know the value of the integral to be $\frac{2}{3}$. All entries in the table underestimate the integral. The closest estimate, with an error of more than 0.003, shows an improvement of only one decimal place of accuracy on the best trapezoidal approximant. All derivatives of the integrand are singular at $x = 0$, and thus the Euler–Maclaurin formula, on which Romberg's method depends, is not applicable. ■

Problem 3.2.1 Show that

$$\frac{x}{e^x - 1} + \frac{1}{2}x$$

is an even function, and deduce from (3.37) that

$$\frac{x}{e^x - 1} = 1 - \frac{1}{2}x + \sum_{j=1}^{\infty} \frac{B_{2j} x^{2j}}{(2j)!}.$$

Problem 3.2.2 Express p_{n+1} as a sum, using (3.42) , and differentiate the sum term by term to verify (3.43).

Problem 3.2.3 Deduce from (3.43) and (3.46) that

$$\int_0^1 p_n(x)dx = 0, \quad n \geq 2,$$

and show also that

$$\lim_{\epsilon \to 0_+} \int_0^{1-\epsilon} p_1(x)dx = 0.$$

Problem 3.2.4 Verify that $p_2(1-x) = p_2(x)$ on $[0,1]$, and prove by induction on n, using (3.43) and the result in Problem 3.2.3, that (3.47) holds for all $n \geq 2$.

Problem 3.2.5 The Fourier expansion for a function f defined on $[0,1]$ is given by

$$f(x) \sim \frac{1}{2}a_0 + \sum_{n=1}^{\infty}(a_n \cos 2\pi nx + b_n \sin 2\pi nx),$$

where

$$a_n = 2\int_0^1 f(x)\cos 2\pi nx\, dx, \quad n \geq 0,$$
$$b_n = 2\int_0^1 f(x)\sin 2\pi nx\, dx, \quad n \geq 1.$$

For $f(x) = p_1(x)$, defined by (3.44), show that every a_n is zero and

$$b_n = 2\int_0^1 x \sin 2\pi nx\, dx = -\frac{1}{\pi n}, \quad n \geq 1,$$

as given in (3.51).

Problem 3.2.6 Using the identity

$$1 - \cos 2\pi nx = 2\sin^2 \pi nx,$$

deduce from (3.52) that

$$p_{2r}(x) - p_{2r}(0) = (-1)^r \sum_{n=1}^{\infty} \frac{4\sin^2 \pi nx}{(2\pi n)^{2r}},$$

and show that $p_{2r}(x) - p_{2r}(0)$ does not change sign.

Problem 3.2.7 Compute the sum $S = \sum_{n=1}^{\infty} 1/n^3$ by applying the Euler–Maclaurin formula, following the method used for estimating the sum in Example 3.2.1.

Problem 3.2.8 Use Romberg's method, with three repeated extrapolations, to estimate the integral of $1/(1+x)$ over the interval $[0,1]$.

3.3 Gaussian Rules

The best-known interpolatory integration rules are the Newton–Cotes rules, which we discussed in Section 3.1, and the Gaussian rules, named after C. F. Gauss. As we will see, these are naturally defined on the interval $[-1, 1]$, and can be used on any finite interval $[a, b]$ by making a linear change of variable. The abscissas of these rules are the zeros of the Legendre polynomials. The Gaussian integration rules are obtained by seeking an interpolatory rule on the interval $[-1, 1]$ whose abscissas $x_1, \ldots, x_n$ are chosen so that the rule is exact for all integrands in P_m, where m is as large as possible. One might expect that the greatest possible value of m is $2n - 1$, since a polynomial of degree $2n - 1$ has $2n$ coefficients, and a rule based on n abscissas has $2n$ parameters, since it has n weights and n abscissas. This turns out to be correct. For as we will show, the rule obtained by integrating the interpolating polynomial based on the zeros of the Legendre polynomial Q_n over the interval $[-1, 1]$ is exact for all integrands in P_{2n-1}. This is called a Gaussian rule or a Gauss–Legendre rule. Thus we find that the one-point Gaussian rule is

$$\int_{-1}^{1} f(x)dx \approx 2f(0). \tag{3.61}$$

This is the midpoint rule, the simplest open Newton–Cotes rule, which we have already met in Section 3.1. (See Problem 3.1.3 for the derivation of an error term for this rule.) The two-point and three-point Gaussian rules are as follows:

$$\int_{-1}^{1} f(x)dx \approx f\left(-\frac{1}{\sqrt{3}}\right) + f\left(\frac{1}{\sqrt{3}}\right), \tag{3.62}$$

$$\int_{-1}^{1} f(x)dx \approx \frac{5}{9} f\left(-\sqrt{\frac{3}{5}}\right) + \frac{8}{9} f(0) + \frac{5}{9} f\left(\sqrt{\frac{3}{5}}\right). \tag{3.63}$$

It can be shown (see Davis and Rabinowitz [11]) that the n-point Gaussian rule plus error term is of the form

$$\int_{-1}^{1} f(x)dx = \sum_{i=1}^{n} w_i^{(n)} f\left(x_i^{(n)}\right) + \frac{2^{2n+1}(n!)^4}{(2n+1)((2n)!)^3} f^{(2n)}(\xi_n), \tag{3.64}$$

where $-1 < \xi_n < 1$, the abscissas $x_i^{(n)}$ are the zeros of the Legendre polynomial Q_n, and the weights $w_i^{(n)}$ are all positive.

We will now justify our above statement that the n-point Gaussian rule is exact for all integrands in P_{2n-1}. As a first step, we state and prove the following lemma.

Lemma 3.3.1 If the divided difference $f[x, x_1, \ldots, x_k]$ is a polynomial in x of degree $m > 0$, then $f[x, x_1, \ldots, x_{k+1}]$ is a polynomial of degree $m - 1$.

Proof. From the recurrence relation (1.22), we have

$$f[x, x_1, \ldots, x_{k+1}] = \frac{f[x_1, \ldots, x_{k+1}] - f[x, x_1, \ldots, x_k]}{x_{k+1} - x}.$$

The lemma then follows from the fact that the numerator on the right side of the latter equation is a polynomial of degree m, and the denominator is a factor of the numerator. The last assertion follows from the observation that the numerator is zero when we put $x = x_{k+1}$, for the numerator then becomes

$$f[x_1, \ldots, x_{k+1}] - f[x_{k+1}, x_1, \ldots, x_k].$$

This is indeed zero, since a divided difference is unchanged if we rearrange the order of its arguments. ■

Theorem 3.3.1 The n-point Gaussian rule,

$$\int_{-1}^{1} f(x)dx \approx \sum_{i=1}^{n} w_i^{(n)} f\left(x_i^{(n)}\right) = G_n(f),$$

say, where $x_1^{(n)}, x_2^{(n)}, \ldots, x_n^{(n)}$ denote the zeros of the Legendre polynomial Q_n, is *exact* if $f \in P_{2n-1}$.

Proof. Let p_{n-1} denote the interpolating polynomial for a given function f on the zeros of the Legendre polynomial Q_n, and write the error of interpolation in the divided difference form

$$f(x) - p_{n-1}(x) = q_n(x)\, f\left[x, x_1^{(n)}, \ldots, x_n^{(n)}\right], \tag{3.65}$$

where

$$q_n(x) = \left(x - x_1^{(n)}\right) \cdots \left(x - x_n^{(n)}\right).$$

Then, from (2.29), we have

$$q_n(x) = \frac{1}{\mu_n} Q_n(x),$$

where

$$\mu_n = \frac{1}{q_n(1)} = \frac{1}{2^n} \left(\begin{array}{c} 2n \\ n \end{array} \right).$$

If we integrate (3.65) over $[-1, 1]$, we obtain an error term for the n-point Gaussian rule,

$$\int_{-1}^{1} f(x)dx - G_n(f) = \frac{1}{\mu_n} \int_{-1}^{1} Q_n(x)\, f\left[x, x_1^{(n)}, \ldots, x_n^{(n)}\right] dx. \tag{3.66}$$

If $f(x)$ is a polynomial in x of degree $2n-1$, an argument like that used in Lemma 3.3.1 shows that the first divided difference $f\left[x, x_1^{(n)}\right]$ is a polynomial of degree $2n-2$, and the repeated use of the lemma shows that

$f\left[x, x_1^{(n)}, \ldots, x_n^{(n)}\right]$ is a polynomial of degree $n-1$. Since the Legendre polynomial Q_n is orthogonal to all polynomials in P_{n-1}, the integral on the right side of (3.66) is zero, and this completes the proof. ■

We have from Theorem 2.1.2 that $Q_n(-x) = (-1)^n Q_n(x)$. Thus, if we order the zeros of Q_n so that

$$-1 < x_1^{(n)} < x_2^{(n)} < \cdots < x_n^{(n)} < 1,$$

we have

$$x_{n-i}^{(n)} = -x_i^{(n)}, \quad 1 \le i \le n.$$

In particular, $x = 0$ is a zero when n is odd. Then, in the same way that we verified the symmetry in the weights of the Newton–Cotes rules, we can show that the weights of the n-point Gaussian rule satisfy

$$w_{n-i}^{(n)} = w_i^{(n)}, \quad 1 \le i \le n.$$

We can deduce from this symmetry in abscissas and weights, as we did with the Newton–Cotes rules, that the Gaussian rules are exact for all integrands that are odd functions. A Gaussian rule may be applied on any finite interval, by using a linear transformation to map the given interval onto $[-1, 1]$, and like the Newton–Cotes rules, Gaussian rules are generally used in composite form.

We conclude this section by mentioning a simple generalization of the Gaussian rules. As we saw in Section 2.2, corresponding to any integrable function ω that is nonnegative on $[-1, 1]$, there exists a sequence of polynomials (q_n^ω) that satisfy

$$\int_{-1}^{1} \omega(x) x^r q_n^\omega(x) dx = 0, \quad 0 \le r < n. \tag{3.67}$$

The polynomials q_n^ω are said to be orthogonal on $[-1, 1]$ with respect to ω, which is called a *weight function*, and the scaled Legendre polynomials are recovered by putting $\omega(x) = 1$. The generalized orthogonal polynomials q_n^ω, like the Legendre polynomials, satisfy a recurrence relation of the form (2.18), where the coefficients α_n and β_n are given by (2.19) and (2.20), amended by inserting the factor $\omega(x)$ into each integrand. Further, if the weight function ω is even, these orthogonal polynomials satisfy Theorem 2.1.2, where again we need to insert the factor $\omega(x)$ into both integrands in (2.20), which defines β_n. Corresponding to any given weight function ω there exists a system of orthogonal polynomials (q_n^ω), and each system yields a sequence of Gaussian-type rules of the form

$$\int_{-1}^{1} \omega(x) f(x) dx \approx \sum_{i=1}^{n} w_i^{(n)} f\left(x_i^{(n)}\right) = G_n^\omega(f), \tag{3.68}$$

say, where $x_1^{(n)}, x_2^{(n)}, \ldots, x_n^{(n)}$ denote the zeros of the polynomial q_n^{ω}. This generalized Gaussian rule is *exact* if $f \in P_{2n-1}$.

Of all the Jacobi polynomials, those whose related integration rules are the most widely known are the Legendre polynomials, which we discussed above, and the Chebyshev polynomials, whose related rules are

$$\int_{-1}^{1} (1-x^2)^{-1/2} f(x)dx \approx \frac{\pi}{n} \sum_{i=1}^{n} f\left(x_i^{(n)}\right), \tag{3.69}$$

where the $x_i^{(n)}$ are the zeros of the Chebyshev polynomial T_n. (See (2.76).) Note that in (3.69), which is called the Gauss–Chebyshev rule of order n, all the weights are equal.

Problem 3.3.1 Verify directly that the two-point Gaussian rule, given in (3.62), is exact for integrands 1 and x^2, and thus for all integrands in P_3.

Problem 3.3.2 Verify directly that the three-point Gaussian rule, given in (3.63), is exact for integrands 1, x^2, and x^4, and thus for all integrands in P_5.

Problem 3.3.3 Assuming that the error of the n-point Gaussian rule is of the form $c_n f^{(2n)}(\xi_n)$, determine the values of c_1, c_2, and c_3 by putting $f(x) = x^2$, x^4, and x^6 in (3.61), (3.62), and (3.63), respectively. Verify that the resulting error terms agree with those given by (3.64).

Problem 3.3.4 Verify that the abscissas of the four-point Gaussian rule are $\pm\left(15 \pm 2\sqrt{30}\right)^{1/2}/\sqrt{35}$, and hence find its weights, so that the rule is exact for the integrands 1, x^2, x^4, and x^6. Apply the rule to the integrand x^8 and, making the same assumption as we did in Problem 3.3.3, show that the error is given by (3.64) with $n = 4$.

Problem 3.3.5 By making the substitution $x = \cos\theta$, show that the Gauss–Chebyshev rule (3.69) becomes

$$\int_0^{\pi} f(\cos\theta)d\theta \approx \frac{\pi}{n} \sum_{i=1}^{n} f(\cos\theta_i),$$

where $\theta_i = (2i-1)\pi/(2n)$. Note that this is simply the composite midpoint rule in the variable θ.

Problem 3.3.6 Show that the Gauss–Chebyshev integration rule of order three is

$$\int_{-1}^{1} (1-x^2)^{-1/2} f(x)dx \approx \frac{\pi}{3}\left(f\left(-\sqrt{3}/2\right) + f(0) + f\left(\sqrt{3}/2\right)\right),$$

and verify that it is exact for all integrands in P_5.

4

Peano's Theorem and Applications

4.1 Peano Kernels

We begin with a verification of the expansion of $f(x)$ as a Taylor polynomial plus an error term expressed as an integral. If $f^{(n+1)}$ exists on $[a, b]$, then

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(f), \qquad (4.1)$$

for $a \le x \le b$, where

$$R_n(f) = \frac{1}{n!}\int_a^x f^{(n+1)}(t)(x-t)^n dt. \qquad (4.2)$$

To justify (4.1) and (4.2), we use integration by parts in (4.2) to obtain

$$R_n(f) = -\frac{f^{(n)}(a)}{n!}(x-a)^n + R_{n-1}(f).$$

A second application of this recurrence relation yields

$$R_n(f) = -\frac{f^{(n)}(a)}{n!}(x-a)^n - \frac{f^{(n-1)}(a)}{(n-1)!}(x-a)^{n-1} + R_{n-2}(f).$$

The required result is then established by applying the recurrence relation n times, and noting that

$$R_0(f) = \int_a^x f'(t)dt = f(x) - f(a).$$

If $f^{(n+1)}$ is continuous, we can apply Theorem 3.1.2 to (4.2), since $(x-t)^n$ does not change sign over the interval of integration, to give

$$R_n(f) = \frac{f^{(n+1)}(\xi_x)}{n!}\int_a^x (x-t)^n dt = \frac{f^{(n+1)}(\xi_x)}{(n+1)!}(x-a)^{n+1},$$

where $a < \xi_x < x$. This is the more familiar version of the error in representing a function by its Taylor polynomial.

When $n > 0$ the factor $(x-t)^n$, regarded as a function of t, is zero for $t = x$, the upper limit of integration in (4.2). It is useful to define a function of t that coincides with $(x-t)^n$ for $x - t \geq 0$, where x is fixed, and is zero for $x - t < 0$. We now give a formal definition of this function.

Definition 4.1.1 For any fixed real number x and any nonnegative integer n, we write $(x-t)_+^n$ to denote the function of t defined for $-\infty < t < \infty$ as follows:

$$(x-t)_+^n = \begin{cases} (x-t)^n, & -\infty < t \leq x, \\ 0, & t > x. \end{cases} \tag{4.3}$$

This is called a *truncated power* function. ■

The function $(x-t)_+^0$ has the value 1 for $-\infty < t \leq x$, and is zero for $t > x$, and so is not continuous at $t = x$. For $n = 1$, we usually write $(x-t)_+^1$ more simply as $(x-t)_+$. As n is increased, the truncated power functions become increasingly smooth, in the sense that derivatives of increasingly higher order are continuous. It is not difficult to prove directly from the definition of derivative that for $n \geq 1$,

$$\frac{d}{dt}(x-t)_+^n = -n(x-t)_+^{n-1}, \tag{4.4}$$

and thus the derivative of a truncated power is just a multiple of the truncated power function of one order less. For $n > 1$, the application of (4.4) $n-1$ times shows that the $(n-1)$th derivative of $(x-t)_+^n$ is a multiple of $(x-t)_+^1$, which is continuous on $-\infty < t < \infty$. However, the nth derivative of $(x-t)_+^n$ behaves like a multiple of $(x-t)_+^0$, which is not continuous at $t = x$. We say that $(x-t)_+^n$, for $n \geq 1$, is a *spline* in $C^{n-1}(-\infty, \infty)$. (We will have much more to say about splines in Chapter 6.) Thus $(x-t)_+^1$ is continuous, and its first derivative is not; $(x-t)_+^2$ and its first derivative are continuous, and its second derivative is not; and so on.

With the introduction of the truncated power function, the expansion of f as a Taylor polynomial plus remainder may be written in the form

$$f(x) = f(a) + f'(a)(x-a) + \cdots + \frac{f^{(n)}(a)}{n!}(x-a)^n + R_n(f),$$

where

$$R_n(f) = \frac{1}{n!}\int_a^b f^{(n+1)}(t)(x-t)_+^n dt. \tag{4.5}$$

Thus, by merely substituting the truncated power function for the factor $(x-t)^n$, we have obtained an integral form of the remainder in which the limits of integration are independent of x. We need one more definition, and then we will be ready to apply (4.5).

Definition 4.1.2 A *linear functional* L is a mapping from the space of functions f defined on $[a,b]$ to the real numbers such that

$$L(\alpha f + \beta g) = \alpha L(f) + \beta L(g),$$

where f, g are any functions, and α, β are any real numbers. ■

Example 4.1.1 Let $p_n \in P_n$ denote the interpolating polynomial for a given function f on $n+1$ distinct abscissas. Let L_x map f to the real number $p_n(x)$. We have included the suffix x in L_x to remind ourselves that this mapping depends on x. As we can easily verify by writing down the Lagrange form of the interpolating polynomial, L_x is a linear functional. Two further examples of linear functionals are

$$L(f) = \int_a^b f(x)dx \qquad \text{and} \qquad L(f) = \sum_{i=1}^{n} w_i f(x_i),$$

where the w_i and x_i are any real numbers. ■

We can now state and prove the main result of this section, the well-known theorem named after Giuseppe Peano (1858–1932).

Theorem 4.1.1 Let us define

$$g_t(x) = (x-t)_+^n, \tag{4.6}$$

and let L denote a linear functional that commutes with the operation of integration, and satisfies the further conditions that $L(g_t)$ is defined and that $L(f) = 0$ for all $f \in P_n$. Then, for all $f \in C^{n+1}[a,b]$,

$$L(f) = \int_a^b f^{(n+1)}(t)K(t)dt, \tag{4.7}$$

where

$$K(t) = K_g(t) = \frac{1}{n!}L(g_t). \tag{4.8}$$

Proof. We need to be rather careful here. For we have been treating the truncated power function $(x-t)_+^n$ as a function of t, with x behaving as a parameter. In the statement of this theorem we have written $L(g_t)$, with $g_t(x) = (x-t)_+^n$, to emphasize that L is applied to the truncated power function $(x-t)_+^n$, regarded as a function of x, with t as a parameter. Thus the linear functional L maps $(x-t)_+^n$ to a real number that depends on t,

that is, to a *function* of t. The functional L is said to *annihilate* functions belonging to P_n. Let us now write $f(x)$ in the form (4.1), with the remainder $R_n(f)$ expressed as in (4.5), which involves the truncated power function. If we now apply the linear functional L to (4.1), because L is linear and annihilates the Taylor polynomial, we obtain

$$L(f) = \frac{1}{n!} L \int_a^b f^{(n+1)}(t)(x-t)_+^n dt.$$

Since the linear functional L commutes with the operation of integration, we have

$$L(f) = \frac{1}{n!} \int_a^b f^{(n+1)}(t)\, L\left((x-t)_+^n\right) dt,$$

and we need to keep in mind that the linear functional L is applied to $(x-t)_+^n$ as a function of x. This completes the proof. ■

Corollary 4.1.1 If, in addition to the conditions stated in Theorem 4.1.1, the kernel $K(t)$ does not change sign on $[a, b]$, then

$$L(f) = \frac{f^{(n+1)}(\xi)}{(n+1)!} L\left(x^{n+1}\right), \tag{4.9}$$

where $a < \xi < b$.

Proof. Since $f^{(n+1)}(t)$ is continuous and $K(t)$ does not change sign on $[a, b]$, we can apply Theorem 3.1.2 to (4.7) to give

$$L(f) = f^{(n+1)}(\xi) \int_a^b K(t)dt, \quad a < \xi < b.$$

This holds for all $f \in C^{n+1}[a, b]$. In particular, on replacing $f(x)$ by x^{n+1} in the latter equation, we obtain

$$L\left(x^{n+1}\right) = (n+1)! \int_a^b K(t)dt,$$

which completes the proof. ■

Remark 1 Suppose that L is a linear functional that commutes with the operation of integration, and that $L(f) = 0$ for all f in, say, P_3. Then, in seeking a kernel K for this linear functional, we would usually define K by (4.8) where g_t is given by (4.6) with the largest admissible value of n for which $L(g_t)$ is defined. But it is worth emphasizing that we could choose any value of $n \leq 3$ for which $L(g_t)$ is defined. ■

Remark 2 The above theorem shows that if K is defined by (4.8), then it satisfies the equation (4.7). Later (see (4.15)), we will find another function K that also satisfies (4.7). This prompts an obvious question: Is there a

unique function K that satisfies (4.7)? To answer this, let K_g denote the function defined by (4.8). Then, if K is *any* function that satisfies (4.7), we have

$$\int_a^b f^{(n+1)}(t)(K(t) - K_g(t))dt = 0$$

for *all* $f \in C^{n+1}[a,b]$. Thus $K(t) - K_g(t)$ must be zero on $[a,b]$, except possibly on a set of measure zero. For example, if we change $K(t)$ at a finite number of points, the value of the integral in (4.7) will be unchanged. In particular, if K_g is continuous on $[a,b]$, then $K = K_g$ is the only continuous function that satisfies (4.7). ■

In the example and problems that follow in this section, the linear functional is of the form

$$L(f) = \int_a^b f(x)dx + \sum_{i=1}^{N} c_i f(x_i) + \sum_{i=1}^{N'} c_i' f'(x_i'), \tag{4.10}$$

where the term involving the integral may be absent, as in Problems 4.1.1 and 4.1.2. We will say that the x_i and x_i' in (4.10) are *abscissas* of the linear functional L. Such linear functionals commute with the operation of integration, and the same is true of linear functionals obtained by adding terms involving evaluations of higher-order derivatives of f to the right side of (4.10).

Example 4.1.2 Let us find the error term for the trapezoidal rule, which we obtained by other means in Section 3.1. We begin with

$$L(f) = \int_a^b f(x)dx - \frac{h}{2}(f(a) + f(b)),$$

where $h = b - a$. This linear functional L annihilates all functions f in P_1, and we may choose $g_t(x) = (x-t)_+$ in (4.8), since $L(g_t)$ exists. Then

$$K(t) = L(g_t) = \int_a^b (x-t)_+ \, dx - \frac{h}{2}((a-t)_+ + (b-t)_+).$$

Note that L is applied to $(x-t)_+$, regarded as a function of x. For $a \le t \le b$, we have $(a-t)_+ = 0$ and $(b-t)_+ = b - t$. Since $(x-t)_+ = 0$ for $x < t$ in the above integral, we may write

$$K(t) = \int_t^b (x-t)dx - \frac{h}{2}(b-t) = \frac{1}{2}(b-t)^2 - \frac{h}{2}(b-t),$$

and since $h = b - a$, this simplifies to give

$$K(t) = \frac{1}{2}(b-t)(a-t) \le 0 \quad \text{for} \quad a \le t \le b.$$

Since the kernel $K(t)$ does not change sign on $[a,b]$, we can apply Corollary 4.1.1, assuming the continuity of f''. We thus need to evaluate

$$L(x^2) = \int_a^b x^2 dx - \frac{h}{2}(a^2+b^2) = \frac{1}{3}(b^3-a^3) - \frac{1}{2}(b-a)(b^2+a^2),$$

which simplifies to give $L(x^2) = -h^3/6$. Finally, we apply (4.9) to give

$$L(f) = \int_a^b f(x)dx - \frac{h}{2}(f(a)+f(b)) = -\frac{h^3}{12}f''(\xi),$$

where $a < \xi < b$. ■

Problem 4.1.1 Let L denote the linear functional

$$L(f) = f(x) - f(-1) - \frac{1}{2}(x+1)[f(1)-f(-1)],$$

which gives the error of linear interpolation at the abscissas -1 and 1. Show that

$$L(f) = \int_{-1}^1 f''(t)K(t)dt,$$

where

$$K(t) = \begin{cases} \frac{1}{2}(x-1)(t+1), & t \le x, \\ \\ \frac{1}{2}(x+1)(t-1), & t > x. \end{cases}$$

Problem 4.1.2 Let L denote the linear functional defined by

$$L(f) = f'(0) - \frac{1}{2h}(f(h)-f(-h)).$$

Show that L annihilates all functions $f \in P_2$, and deduce that, with $n=2$ in (4.6), the Peano kernel satisfies

$$2!K(t) = L((x-t)_+^2) = \begin{cases} -\frac{1}{2h}(h+t)^2, & -h \le t \le 0, \\ \\ -\frac{1}{2h}(h-t)^2, & 0 < t \le h. \end{cases}$$

Verify that K is an even function that does not change sign on $[-h,h]$, and apply Corollary 4.1.1 to show that

$$f'(0) - \frac{1}{2h}(f(h)-f(-h)) = -\frac{h^2}{6}f^{(3)}(\xi), \quad -h < \xi < h.$$

Problem 4.1.3 Apply the Peano kernel theory to find the error term for the midpoint rule, given by

$$\int_{x_0}^{x_2} f(x)dx = 2hf(x_1) + \frac{h^3}{3}f''(\xi),$$

where $\xi \in (x_0, x_2)$, which we obtained by other means in Problem 3.1.3.

Problem 4.1.4 The linear functional that is associated with the two-point Gaussian rule is

$$L(f) = \int_{-1}^{1} f(x)dx - f\left(-1/\sqrt{3}\right) - f\left(1/\sqrt{3}\right),$$

and L annihilates all $f \in P_3$. Show that, with $n = 3$ in (4.6), the Peano kernel for L is the even function K defined by

$$3!K(t) = \begin{cases} a - \frac{1}{4}t^2(b - t^2), & 0 \le t \le \frac{1}{\sqrt{3}}, \\ \\ \frac{1}{4}(1-t)^4, & \frac{1}{\sqrt{3}} < t \le 1, \end{cases}$$

where

$$a = \frac{1}{4} - \frac{\sqrt{3}}{9} > 0 \quad \text{and} \quad b = 4\sqrt{3} - 6 > t^2, \ \ 0 \le t \le \frac{1}{\sqrt{3}}.$$

Verify that on $[-1, 1]$, $K(t)$ is continuous, nonnegative, and attains its maximum modulus at $t = 0$. Hence apply Corollary 4.1.1 to justify the error term for the two-point Gaussian rule given by (3.64) with $n = 2$.

Problem 4.1.5 The trapezoidal rule with end correction on $[-1, 1]$ is

$$\int_{-1}^{1} f(x)dx \approx f(-1) + f(1) - \frac{1}{3}(f'(1) - f'(-1)).$$

This rule is exact (see Problem 3.1.9) for all $f \in P_3$. Apply (4.8) where g_t is defined by (4.6) with $n = 2$, and obtain the kernel

$$K(t) = \frac{1}{3}t(1 - t^2), \quad -1 \le t \le 1.$$

4.2 Further Properties

The Peano kernel in Example 4.1.2,

$$K(t) = \frac{1}{2}(b - t)(a - t), \quad a \le t \le b,$$

is zero at both endpoints of the interval on which it is defined. We now show that this condition holds for a very large class of kernels.

Theorem 4.2.1 Let L denote a linear functional of the form (4.10) such that $L(f) = 0$ for all $f \in P_n$, where the abscissas x_i and x_i' all belong to the interval $[a, b]$, and let the Peano kernel K be defined by (4.8), with g_t defined by (4.6). Then if $n \ge 2$, the kernel $K(t)$ is zero at both endpoints $t = a$ and $t = b$. Further, if all coefficients c_i' in (4.10) are zero, then the above result holds for $n \ge 1$.

Proof. If $f(x) = (x-t)_+^n$, the integral on the right of (4.10) is a continuous function of t. Also, for any fixed x, both $(x-t)_+^n$ and its derivative are continuous functions of t if $n \geq 2$. Thus, if $n \geq 2$ and $f(x) = (x-t)_+^n$, we see from (4.10) that $L(f)$ is a continuous function of t, and so from (4.8) and (4.6), the kernel $K(t)$ is continuous. For any linear functional L that satisfies the conditions of the theorem, with $n \geq 2$, let us evaluate $L\left((x-t)_+^n\right)$ at $t = a$, which we will write as $L\left((x-a)_+^n\right)$. Then we observe that for $a \leq x \leq b$,

$$L\left((x-a)_+^n\right) = L\left((x-a)^n\right) = 0,$$

since L annihilates all polynomials in P_n. We also observe that

$$L\left((x-b)_+^n\right) = L(0) = 0,$$

where $L(0)$ denotes the result of applying the linear functional L to the zero function. It then follows from (4.8) that when $n \geq 2$,

$$K(a) = 0 \quad \text{and} \quad K(b) = 0.$$

This result is obviously valid when $n = 1$ if all coefficients c_i' are zero in (4.10). This completes the proof. ■

Example 4.2.1 Consider the three-eighths rule on $[-3, 3]$. (See Example 3.1.2.) The corresponding linear functional is

$$L(f) = \int_{-3}^{3} f(x)dx - \frac{3}{4}(f(-3) + 3f(-1) + 3f(1) + f(3)).$$

Note that L annihilates all f in P_3. Let us define K by (4.8), where g_t is given by (4.6) with $n = 3$. Thus we obtain

$$3!K(t) = \int_{-3}^{3} (x-t)_+^3 dx - \frac{3}{4}\left(3(-1-t)_+^3 + 3(1-t)_+^3 + (3-t)_+^3\right),$$

since $(-3-t)_+^3 = 0$ on $[-3, 3]$. Now

$$\int_{-3}^{3} (x-t)_+^3 dx = \int_{t}^{3} (x-t)^3 dx = \frac{1}{4}(3-t)^4,$$

and we find that $K(t)$ is a polynomial of degree 4 in t in each of the intervals $[-3, -1]$, $[-1, 1]$, and $[1, 3]$. On the first interval, $[-3, -1]$, we have

$$3!K(t) = \frac{1}{4}(3-t)^4 - \frac{3}{4}\left(3(-1-t)^3 + 3(1-t)^3 + (3-t)^3\right) = \frac{1}{4}t(3+t)^3.$$

On the second interval, $[-1, 1]$, we find that

$$3!K(t) = \frac{1}{4}(3-t)^4 - \frac{3}{4}\left(3(1-t)^3 + (3-t)^3\right) = \frac{1}{4}(t^2-3)(t^2+3),$$

and on the third interval, $[1,3]$, we have

$$3!K(t) = \frac{1}{4}(3-t)^4 - \frac{3}{4}(3-t)^3 = -\frac{1}{4}t(3-t)^3.$$

On combining the above results, we find that the kernel is given by

$$K(t) = \begin{cases} \frac{1}{24}t(3+t)^3, & -3 \le t \le -1, \\ \frac{1}{24}(t^2-3)(t^2+3), & -1 \le t \le 1, \\ -\frac{1}{24}t(3-t)^3, & 1 \le t \le 3. \end{cases}$$

As predicted by Theorem 4.2.1, $K(-3) = K(3) = 0$. Let us extend the definition of K to $(-\infty, \infty)$ by setting $K(t) = 0$ outside the interval $[-3,3]$. We can then easily adapt the above representation of K on $[-3,3]$ to $(-\infty,\infty)$ by using truncated power functions, expressing $K(t)$ by $\frac{1}{24}t(3+t)_+^3$ for $-\infty \le t \le -1$, and by $-\frac{1}{24}t(3-t)_+^3$ for $1 \le t \le \infty$. We see that K is an even function, that is, $K(-t) = K(t)$, and $K(t) \le 0$ for all real t. Also, by examining the continuity of K and its derivatives at $t = \pm 1$ and $t = \pm 3$, we see that K belongs to $C^2(-\infty,\infty)$. On applying Corollary 4.1.1, assuming continuity of $f^{(4)}$, we find that $L(x^4) = -144/5$, and thus

$$\int_{-3}^{3} f(x)dx = \frac{3}{4}(f(-3) + 3f(-1) + 3f(1) + f(3)) - \frac{6}{5}f^4(\xi). \qquad (4.11)$$

If we make the linear change of variable $x = 2(u - x_0)/h - 3$, the interval $-3 \le x \le 3$ is mapped onto $x_0 \le u \le x_3$, where $x_j = x_0 + jh$, and then (4.11) gives the form of the three-eighths rule in (3.11), with error term $-3h^5 f^{(4)}(\eta)/80$, where $x_0 < \eta < x_3$. ■

Perhaps it is not surprising that the Peano kernel in Example 4.2.1 is an even function, given that the linear functional is symmetric with respect to $x = 0$. Also, in Example 4.1.2, the Peano kernel $\frac{1}{2}(b-t)(a-t)$ becomes the even function $-\frac{1}{2}(1-t^2)$ when we transform $[a,b]$ to the interval $[-1,1]$, and the linear functional of Example 4.1.2 is then symmetric about $x = 0$. We will now show that for a large class of linear functionals with this kind of symmetry, the corresponding Peano kernels are even or odd, depending on whether n is respectively odd or even in (4.6) and (4.8). We begin by finding an alternative form of the error term for the approximation of $f(x)$ by its Taylor polynomial, given by (4.1) and (4.2). If $f^{(n+1)}$ exists on $[a,b]$, we can write

$$f(x) = f(b) + f'(b)(x-b) + \cdots + \frac{f^{(n)}(b)}{n!}(x-b)^n + R_n'(f), \qquad (4.12)$$

for $a \le x \le b$, where

$$R_n'(f) = (-1)^{n+1}\frac{1}{n!}\int_x^b f^{(n+1)}(t)(t-x)^n dt.$$

We can use integration by parts to show that

$$R'_n(f) = -\frac{f^{(n)}(b)}{n!}(x-b)^n + R'_{n-1}(f),$$

and thus verify (4.12) in the same way that we justified (4.1) and (4.2). Then we can express $R'_n(f)$ in terms of a truncated power function, in the form

$$R'_n(f) = (-1)^{n+1}\frac{1}{n!}\int_a^b f^{(n+1)}(t)(t-x)_+^n dt. \tag{4.13}$$

The truncated power function $(t-x)_+^n$ has the value $(t-x)^n$ when $t-x \geq 0$ and is zero otherwise. This is consistent with Definition 4.1.1.

Now define

$$h_t(x) = (-1)^{n+1}(t-x)_+^n, \tag{4.14}$$

and let L denote a linear functional that commutes with the operation of integration, and satisfies the conditions that $L(f) = 0$ for all $f \in P_n$ and that $L(h_t)$ is defined. Let us write $f(x)$ as in (4.12), where $R'_n(f)$ is given by (4.13), and evaluate $L(f)$. Then, in the same way as we derived (4.7), we obtain

$$L(f) = \int_a^b f^{(n+1)}(t)K(t)dt,$$

where the kernel K is given by

$$K(t) = K_h(t) = \frac{1}{n!}L(h_t). \tag{4.15}$$

Thus we now have two expressions for the Peano kernel: $K = K_h$, defined by (4.15), involving the truncated power function $(t-x)_+^n$, and $K = K_g$, defined by (4.8), involving the truncated power function $(x-t)_+^n$. We know (see Remark 2 after the proof of Theorem 4.1.1) that $K_g(t)$ and $K_h(t)$ must be equal on $[a,b]$, except possibly on a set of measure zero. In the following theorem we will refine this result for a class of symmetric linear functionals, and give conditions that ensure that the associated Peano kernels are even or odd functions.

Theorem 4.2.2 Let L be a linear functional of the form

$$L(f) = \int_{-a}^{a} f(x)dx + c_0 f(0) + \sum_{i=1}^{N} c_i(f(x_i) + f(-x_i)) + \sum_{i=1}^{N'} c'_i(f'(x'_i) - f'(-x'_i)), \tag{4.16}$$

where $x_i > 0$ and $x'_i > 0$ belong to the interval $[-a,a]$. Further, let L annihilate all functions $f \in P_n$ and have the additional properties that

$L(g_t)$ and $L(h_t)$ both exist, where g_t and h_t are defined by (4.6) and (4.14), respectively. Then the Peano kernels K_g and K_h, defined by (4.8) and (4.15), respectively, satisfy

$$K_g(-t) = (-1)^{n+1} K_h(t). \tag{4.17}$$

If $n \geq 2$ in (4.6) and (4.14), the kernels K_g and K_h are equal, belong to $C^{n-2}[-a, a]$, and are even or odd, depending on whether n is odd or even, respectively. If, with $n \geq 2$, there are no derivative terms present in (4.16), then we can improve the smoothness property, to say that

$$K_g = K_h \in C^{n-1}[-a, a].$$

Proof. From (4.8) we have

$$n! K_g(-t) = L\left((x+t)_+^n\right),$$

and we note that

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

Thus, for linear functionals L of the form (4.16), we obtain

$$\begin{aligned} n! K_g(-t) = \frac{(a+t)^{n+1}}{n+1} &+ c_0 (t)_+^n + \sum_{i=1}^{N} c_i \left((x_i+t)_+^n + (-x_i+t)_+^n\right) \\ &+ n \sum_{i=1}^{N'} c_i' \left((x_i+t)_+^{n-1} - (-x_i+t)_+^{n-1}\right), \end{aligned}$$

where we need to omit the terms involving the truncated power functions of order $n-1$ when $n = 0$. Let us now use (4.15) to evaluate the kernel $K_h(t)$. Since

$$\int_{-a}^{a} (t-x)_+^n dx = \int_{-a}^{t} (t-x)^n dx = \frac{(t+a)^{n+1}}{n+1},$$

we obtain

$$\begin{aligned} (-1)^{n+1} n! K_h(t) = \frac{(t+a)^{n+1}}{n+1} &+ c_0 (t)_+^n + \sum_{i=1}^{N} c_i \left((t-x_i)_+^n + (t+x_i)_+^n\right) \\ &+ n \sum_{i=1}^{N'} c_i' \left(-(t-x_i)_+^{n-1} + (t+x_i)_+^{n-1}\right), \end{aligned}$$

where again we need to omit the terms involving the truncated power functions of order $n-1$ when $n = 0$. A comparison of the latter expression

for $(-1)^{n+1}n!K_h(t)$ with that given above for $n!K_g(-t)$ justifies (4.17). It is also clear from the nature of the truncated power functions that for $n \geq 2$, both kernels are in $C^{n-2}[-a,a]$, and are in $C^{n-1}[-a,a]$ if there are no derivative terms in (4.16). This, together with the fact that K_g and K_h must be equal on $[-a,a]$, except possibly on a set of measure zero, shows that they are equal on the *whole* interval $[-a,a]$ when $n \geq 2$, and thus $K_g = K_h$ is even or odd, depending on whether n is odd or even, respectively. ■

Example 4.2.2 The trapezoidal rule with end correction (see Problem 3.1.9) on $[-1,1]$ is

$$\int_{-1}^{1} f(x)dx \approx (f(-1)+f(1)) - \frac{1}{3}(f'(1)-f'(-1)).$$

The related linear functional L annihilates all integrands f in P_3, and in Problem 4.1.5 we applied (4.8) with $n=2$. Here let us apply (4.8) with $n=3$ to give

$$3!K(t) = \int_{-1}^{1}(x-t)_+^3 dx - \left((-1-t)_+^3 + (1-t)_+^3\right) + \left((1-t)_+^2 - (-1-t)_+^2\right).$$

Now, on $-1 \leq t \leq 1$, since $(-1-t)_+^3$ and $(-1-t)_+^2$ are always zero, and $(1-t)_+^3$ and $(1-t)_+^2$ are always nonnegative, we may write

$$3!K(t) = \int_t^1 (x-t)^3 dx - (1-t)^3 + (1-t)^2 = \frac{1}{4}(1-t^2)^2.$$

Again, let us extend the definition of this kernel to the whole real line, by setting $K(t)=0$ outside the interval $[-1,1]$. It is easily verified that K belongs to $C^1(-\infty,\infty)$. Since the kernel K is nonnegative, we may apply Corollary 4.1.1, assuming continuity of $f^{(4)}$. We find that $L(x^4) = 16/15$ and (4.9) gives $L(f) = f^{(4)}(\xi)L(x^4)/4!$. Thus we have

$$\int_{-1}^{1} f(x)dx = (f(-1)+f(1)) - \frac{1}{3}(f'(1)-f'(-1)) + \frac{2}{45}f^{(4)}(\xi).$$

We will now make the linear change of variable $x = 2(u-x_0)/h - 1$, which maps $[-1,1]$ onto $[x_0,x_1]$. If we then replace u by x, we obtain

$$\int_{x_0}^{x_1} f(x)dx = \frac{h}{2}(f(x_0)+f(x_1)) - \frac{h^2}{12}(f'(x_1)-f'(x_0)) + \frac{h^5}{720}f^{(4)}(\eta),$$

where $x_0 < \eta < x_1 = x_0 + h$. By writing down the above rule plus error term over the intervals $[x_{j-1},x_j]$ of equal length, and summing from $j=1$ to N, we find that all but two of the derivative terms cancel, and we obtain the following expression for the composite trapezoidal rule with end correction,

$$\int_a^b f(x)dx = T_N(f) - \frac{h^2}{12}(f'(b)-f'(a)) + \frac{h^4}{720}(b-a)f^{(4)}(\zeta),$$

where $a < \zeta < b$ and T_N denotes the composite trapezoidal rule. ■

Example 4.2.3 As we saw, Simpson's rule is exact for all integrands in P_3. Thus, in seeking an error term, it is natural (see Problem 4.2.1) to evaluate the kernel K defined by (4.8), where g_t is given by (4.6) with $n = 3$. However (recall Remark 1 that follows Corollary 4.1.1), we will choose $n = 2$ in (4.6). We will work with the interval $[-1, 1]$, and define

$$L(f) = \int_{-1}^{1} f(x)dx - \frac{1}{3}(f(-1) + 4f(0) + f(1)).$$

Then the resulting kernel K satisfies

$$2!K(t) = \int_{-1}^{1} (x-t)_+^2 dx - \frac{1}{3}\left((-1-t)_+^2 + 4(-t)_+^2 + (1-t)_+^2\right),$$

and since $(-1-t)_+^2$ is always zero, and $(1-t)_+^2$ is always nonzero on $[-1, 1]$, we see that

$$2!K(t) = \int_{t}^{1} (x-t)^2 dx - \frac{1}{3}\left(4(-t)_+^2 + (1-t)^2\right).$$

Thus

$$2!K(t) = \frac{1}{3}(1-t)^3 - \frac{1}{3}\left(4(-t)^2 + (1-t)^2\right), \quad -1 \le t \le 0,$$

and

$$2!K(t) = \frac{1}{3}(1-t)^3 - \frac{1}{3}(1-t)^2, \quad 0 < t \le 1.$$

We simplify these expressions for $K(t)$ on $[-1, 0]$ and $[0, 1]$, to give

$$K(t) = \begin{cases} -\frac{1}{6}t(1+t)^2, & -1 \le t \le 0, \\ \\ -\frac{1}{6}t(1-t)^2, & 0 < t \le 1. \end{cases}$$

We see that K changes sign in $[-1, 1]$ and is an odd function. If we extend the definition of K to $(-\infty, \infty)$ by putting $K(t) = 0$ outside $[-1, 1]$, we may verify that $K \in C^1(-\infty, \infty)$. This error term for Simpson's rule is

$$\int_{-1}^{1} f(x)dx - \frac{1}{3}(f(-1) + 4f(0) + f(1)) = \int_{-1}^{1} f^{(3)}(t)K(t)dt,$$

and it holds for all $f \in C^3[-1, 1]$. ■

In Example 4.2.2 we applied the Peano kernel theory to derive an error term for the trapezoidal rule with end correction, which is a special case of

the Euler–Maclaurin formula. Let

$$L(f) = \int_0^1 f(x)dx - \frac{1}{2}(f(0)+f(1)) + \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}\left(f^{(2r-1)}(1) - f^{(2r-1)}(0)\right). \tag{4.18}$$

On the right of (4.18) we have the Euler–Maclaurin formula on a single interval with $h = 1$, that is, (3.40) with $N = 1$ and $h = 1$. Therefore, L annihilates all $f \in P_{2m+1}$, and we find that the kernel defined by (4.8) with $n = 2m+1$ satisfies

$$(2m+1)!\, K_m(t) = \frac{1}{2m+2}(1-t)^{2m+2} - \frac{1}{2}(1-t)^{2m+1} + \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}(2m+1)\cdots(2m-2r+3)(1-t)^{2m-2r+2}.$$

We have written the kernel as K_m to emphasize its dependence on m. With $m = 1$, this simplifies to give

$$3!\, K_1(t) = \frac{1}{4}t^2(1-t)^2 = \frac{1}{4}u^2,$$

where $u = t(1-t)$. All of the kernels K_m can be expressed as polynomials in u. The next few expressions for K_m simplify to give

$$\begin{aligned} 5!\, K_2(t) &= -\frac{1}{12}u^2(1+2u), \\ 7!\, K_3(t) &= \frac{1}{24}u^2(3u^2+4u+2), \\ 9!\, K_4(t) &= -\frac{1}{20}u^2(1+u)(2u^2+3u+3), \\ 11!\, K_5(t) &= \frac{1}{24}u^2(2u^4+8u^3+17u^2+20u+10), \end{aligned}$$

with $u = t(1-t)$, and we note that $0 \le t \le 1$ corresponds to $0 \le u \le \frac{1}{4}$. The above Peano kernels obviously do not change sign, and we will show that this holds for every $K_m(t)$. First, by extending Theorem 4.2.2 in an obvious way to deal with the linear operator defined in (4.18), we see that $K_m(t) = K_m(1-t)$. Let us replace $1-t$ by t on the right side of the above expression for $K_m(t)$, differentiate, and compare the expression for $K_m'(t)$ with that for $p_{2m+1}(t)$ given by (3.42). Recalling that $B_0 = 1$, $B_1 = -\frac{1}{2}$, and that $B_r = 0$ when $r > 1$ is odd, we see that

$$K_m'(t) = p_{2m+1}(t).$$

If we now integrate this equation, using (3.43) and (3.45), we obtain

$$K_m(t) = p_{2m+2}(t) - p_{2m+2}(0),$$

since $K_m(0) = 0$, and we see from Problem 3.2.6 that $K_m(t)$ indeed has constant sign. Thus we may apply Corollary 4.1.1, and find that

$$L(x^{2m+2}) = \frac{1}{2m+3} - \frac{1}{2} + \sum_{r=1}^{m} \frac{B_{2r}}{(2r)!}(2m+2)\cdots(2m-2r+4). \quad (4.19)$$

Now let us consider (3.35) and (3.36). On putting $s = m+1$ in (3.35), we see that the right side of (4.19) would be zero if we replaced m by $m+1$ in the upper limit of the sum. Thus (4.19) greatly simplifies to give

$$L\left(x^{2m+2}\right) = -B_{2m+2}, \quad (4.20)$$

and (4.9) then gives

$$L(f) = -\frac{B_{2m+2}}{(2m+2)!} f^{(2m+2)}(\xi), \quad 0 < \xi < 1, \quad (4.21)$$

where $f^{(2m+2)}$ is continuous on $[0,1]$. In (4.18) let us map $[0,1]$ onto $[x_j, x_{j+1}]$, with $x_{j+1} - x_j = h$, and sum from $j = 0$ to $N-1$. If $f^{(2m+2)}$ is continuous on $[x_0, x_N]$, we can combine the N error terms, as we did in deriving the error term for the composite trapezoidal rule, given in (3.18). Thus we obtain

$$\int_a^b f(x)dx = T_N(f) - \sum_{r=1}^{m} \frac{h^{2r} B_{2r}}{(2r)!} \left(f^{(2r-1)}(x_N) - f^{(2r-1)}(x_0)\right)$$
$$- h^{2m+2}(b-a)\frac{B_{2m+2}}{(2m+2)!} f^{(2m+2)}(\zeta), \quad (4.22)$$

where $a = x_0$, $b = x_N$, and $a < \zeta < b$.

Problem 4.2.1 Use the Peano kernel theory with $n = 3$ in (4.6) to find an error term for Simpson's rule based on the interval $[-1, 1]$. Show that the resulting kernel is the even function K defined by

$$K(t) = -\frac{1}{72}(1-t)^3(3t+1), \quad 0 \le t \le 1.$$

Set $K(t) = 0$ outside $[-1, 1]$, and verify that K belongs to $C^2(-\infty, \infty)$.

Problem 4.2.2 Use (4.6) with $n = 0$ to find an error term for Simpson's rule based on the interval $[-1, 1]$. Show that

$$\int_{-1}^{1} f(x)dx - \frac{1}{3}(f(-1) + 4f(0) + f(1)) = \int_{-1}^{1} f'(t)K(t),$$

for all $f \in C^1[-1,1]$, where

$$K(t) = \begin{cases} 0, & t = -1, \\ -\frac{2}{3} - t, & -1 < t \le 0, \\ \frac{2}{3} - t, & 0 < t \le 1. \end{cases}$$

Problem 4.2.3 Let us apply (4.8) and (4.15) in turn, with $n = 0$, to derive Peano kernels, say K_g and K_h, for the midpoint rule on $[-1,1]$. Show that

$$K_g(t) = \begin{cases} -1 - t, & -1 \le t \le 0, \\ 1 - t, & 0 < t \le 1, \end{cases}$$

and

$$K_h(t) = \begin{cases} -1 - t, & -1 \le t < 0, \\ 1 - t, & 0 \le t \le 1. \end{cases}$$

Note that $K_g(0) = -1$ and $K_h(0) = 1$, and that $K_g(t) = K_h(t)$ for all other values of t on $[-1,1]$.

Problem 4.2.4 Consider the integration rule with error term

$$\int_{-1}^{1} f(x)dx = f(-1) + f(1) - \frac{2}{5}(f'(1) - f'(-1)) + \frac{1}{15}(f''(-1) + f''(1)) - \frac{2}{1575}f^{(6)}(\xi),$$

where $a < \xi < b$. Verify that the rule is exact for all $f \in P_5$. Apply (4.8) with $n = 5$ to show that the resulting Peano kernel is

$$K(t) = -\frac{1}{720}(1 - t^2)^3, \quad -1 \le t \le 1,$$

and hence justify the above error term.

5
Multivariate Interpolation

5.1 Rectangular Regions

Multivariate interpolation is concerned with interpolation of a function of more than one variable, and we will find that it is by no means as simple and straightforward as interpolation of a function of one variable (univariate interpolation). As we saw in Section 1.1, given the values of a univariate function f at $n+1$ distinct abscissas $x_0, x_1, \ldots, x_n$, we can choose the $n+1$ monomials $1, x, x^2, \ldots, x^n$ as a *basis* for P_n, and we can always find a linear combination of these, a *polynomial* p_n, that provides a unique solution to the following interpolation problem: Find $p_n \in P_n$ such that $p_n(x_j) = f(x_j)$ for $0 \leq j \leq n$. We also saw how the choice of the fundamental polynomials $L_j(x)$ or Newton's polynomials $\pi_i(x)$, defined in (1.9) and (1.11) as alternative bases for P_n, led respectively to the Lagrange and Newton forms of the interpolating polynomial in one variable.

There is no *theoretical* difficulty in setting up a framework for discussing interpolation of a multivariate function f whose values are known at, say, N abscissas in real d-dimensional Euclidean space $\mathbb{R}^d$. Let $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ denote N distinct abscissas in $\mathbb{R}^d$, and let $\phi_1, \phi_2, \ldots, \phi_N$ denote N linearly independent functions in $C[\mathbb{R}^d]$, the linear space of all continuous mappings from $\mathbb{R}^d$ to $\mathbb{R}$, the real numbers. Thus none of the functions ϕ_j can be expressed as a linear combination of the others. Finally, let $S_\phi \subset C[\mathbb{R}^d]$ denote the *span* of $\phi_1, \phi_2, \ldots, \phi_N$, that is, the set of all linear combinations of the ϕ_j, and let $f \in C[\mathbb{R}^d]$ denote a function that is not in S_ϕ. Then we can obtain a unique solution of the interpolating problem of determining

$a_1, a_2, \ldots, a_N \in \mathbb{R}$ such that

$$a_1\phi_1(\mathbf{x}_j) + a_2\phi_2(\mathbf{x}_j) + \cdots + a_N\phi_N(\mathbf{x}_j) = f(\mathbf{x}_j), \tag{5.1}$$

for $1 \le j \le N$, if and only if the matrix

$$\mathbf{A} = \begin{bmatrix} \phi_1(\mathbf{x}_1) & \phi_2(\mathbf{x}_1) & \cdots & \phi_N(\mathbf{x}_1) \\ \phi_1(\mathbf{x}_2) & \phi_2(\mathbf{x}_2) & \cdots & \phi_N(\mathbf{x}_2) \\ \vdots & \vdots & \vdots & \vdots \\ \phi_1(\mathbf{x}_N) & \phi_2(\mathbf{x}_N) & \cdots & \phi_N(\mathbf{x}_N) \end{bmatrix} \tag{5.2}$$

is nonsingular. As a special case, let us take $d = 1$, choose the ϕ_j as the first N monomials, and then S_ϕ is simply P_{N-1}, the set of polynomials of degree at most $N-1$. Then the above matrix $\mathbf{A}$ is the $N \times N$ Vandermonde matrix, whose $(n+1)\times(n+1)$ form is given in (1.7). The following example warns us of possible pitfalls in multivariate interpolation.

Example 5.1.1 Suppose we have a mapping f from $\mathbb{R}^2$ to $\mathbb{R}$, and that the values of $f(x, y)$ are known at the points $(1, 0)$, $(-1, 0)$, $(0, 1)$, and $(0, -1)$. Let us construct an interpolating function of the form

$$p(x, y) = a_1 + a_2x + a_3y + a_4xy.$$

In this case, the system of equations (5.1) becomes

$$\begin{bmatrix} 1 & 1 & 0 & 0 \\ 1 & -1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 1 & 0 & -1 & 0 \end{bmatrix} \begin{bmatrix} a_1 \\ a_2 \\ a_3 \\ a_4 \end{bmatrix} = \begin{bmatrix} f(1,0) \\ f(-1,0) \\ f(0,1) \\ f(0,-1) \end{bmatrix}.$$

Let $\mathbf{r}_1^T$, $\mathbf{r}_2^T$, $\mathbf{r}_3^T$, and $\mathbf{r}_4^T$ denote the rows of the above matrix. Since the matrix is obviously singular, its rows must be linearly dependent, and we find that $\mathbf{r}_1 + \mathbf{r}_2 = \mathbf{r}_3 + \mathbf{r}_4$. The above system of linear equations has a solution if and only if

$$f(1, 0) + f(-1, 0) = f(0, 1) + f(0, -1).$$

If this condition holds, a_1, a_2, and a_3 are uniquely determined, and a_4 may be chosen arbitrarily. ■

In this book we confine our discussion of multivariate interpolation to two classes of abscissas in d-dimensional Euclidean space $\mathbb{R}^d$. For the remainder of this section we consider rectangular arrays of abscissas in $\mathbb{R}^2$, with an obvious extension to boxlike arrays in higher dimensions, and in Section 5.2 we consider abscissas arranged in triangular formations, with obvious extensions to higher dimensions. We will find that the methods of interpolation on both rectangular and triangular classes of abscissas,

and their extensions to higher dimensions, nicely generalize those used for interpolation of a function of one variable.

We begin with the equation $z = f(x, y)$, which corresponds to a *surface* in three-dimensional Euclidean space. The coordinates of each point (x, y, z) in the surface satisfy $z = f(x, y)$. For example, the equation $z = 1 + 2x + 3y$ corresponds to the unique *plane* that passes through the three points $(-\frac{1}{2}, 0, 0)$, $(0, -\frac{1}{3}, 0)$, and $(0, 0, 1)$, and the equation $z = (1 - x^2 - y^2)^{1/2}$ corresponds to that part of the sphere with centre at $(0, 0, 0)$ and radius 1 that lies above the xy-plane. This is the hemisphere of points such that $x^2 + y^2 + z^2 = 1$ and $z \geq 0$. If for a general function $f(x, y)$, we now *fix* the value of x, choosing $x = x_i$, then the equation $z = f(x_i, y)$ corresponds to the *curve* defined by the intersection of the plane $x = x_i$ and the surface $z = f(x, y)$. We can envisage, say, $m + 1$ such curves, corresponding to $x = x_0, x_1, \ldots, x_m$. Then, using our experience with univariate Lagrange interpolation, we construct an interpolating function

$$\xi(x, y) = \sum_{i=0}^{m} f(x_i, y) L_i(x), \tag{5.3}$$

where

$$L_i(x) = \prod_{j \neq i} \left(\frac{x - x_j}{x_i - x_j} \right),$$

as defined in (1.9), the above product being taken over all j from 0 to m, but excluding $j = i$. Note how (5.3) follows the form of (1.10), with $f(x_i, y)$ in place of $f(x_i)$. We call $\xi(x, y)$ a *blending function*. We will write $\xi(f; x, y)$ to denote this blending function when we want to emphasize its dependence on the function f. The blending function $\xi(f; x, y)$ agrees with $f(x, y)$ at all points where the plane $x = x_i$ intersects $z = f(x, y)$, for $0 \leq i \leq m$.

Example 5.1.2 Let us derive the blending function for 2^{x+y} that coincides with the given function for $x = -1, 0$, and 1. Following the model in (5.3), we obtain the blending function

$$\xi(x, y) = 2^{-1+y} \cdot \frac{1}{2} x(x - 1) + 2^y \cdot (1 - x^2) + 2^{1+y} \cdot \frac{1}{2} x(x + 1),$$

which simplifies to give

$$\xi(x, y) = 2^{y-2}(x^2 + 3x + 4). \qquad \blacksquare$$

We can interchange the roles of x and y in the construction of a blending function, and interpolate at $y = y_0, y_1, \ldots, y_n$. To avoid confusion, we will denote fundamental polynomials in the variable y by $M_j(y)$, where

$$M_j(y) = \prod_{k \neq j} \left(\frac{y - y_k}{y_j - y_k} \right)$$

and the product is taken over all k from 0 to n, but excluding $k = j$. Then we can construct the alternative blending function

$$\eta(x, y) = \eta(f; x, y) = \sum_{j=0}^{n} f(x, y_j) M_j(y), \tag{5.4}$$

which agrees with $f(x, y)$ along each of the $n + 1$ curves defined by the intersection of the plane $y = y_j$ with the surface $z = f(x, y)$, for $0 \le j \le n$. It is then of interest to apply this second blending process, defined by (5.4), not to $f(x, y)$, but to the first blending function $\xi(x, y)$, defined by (5.3). We obtain, say,

$$p(x, y) = \eta(\xi; x, y) = \sum_{j=0}^{n} \left(\sum_{i=0}^{m} f(x_i, y_j) L_i(x) \right) M_j(y), \tag{5.5}$$

and we note that $p(x, y)$ is a polynomial in x and y. On writing the repeated summation (5.5) as a double summation, we obtain

$$p(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} f(x_i, y_j) L_i(x) M_j(y). \tag{5.6}$$

Thus we have first derived $\xi(x, y)$, the x-blended function for $f(x, y)$, and then derived the polynomial $p(x, y)$, the y-blended function for $\xi(x, y)$, as approximations to $f(x, y)$. Denoting the two sets of abscissas defined above by

$$X = \{x_0, x_1, \ldots, x_m\} \quad \text{and} \quad Y = \{y_0, y_1, \ldots, y_n\},$$

we now construct a rectangular grid of $(m + 1) \times (n + 1)$ points, which we will write as

$$X \times Y = \{(x_i, y_j) \,|\, x_i \in X, y_j \in Y\}.$$

We read $X \times Y$ as "X cross Y" and call it the *Cartesian product* of the two sets X and Y. Then we see that $L_i(x)M_j(y)$ has the value 1 at the point (x_i, y_j), and the value zero at all other points in $X \times Y$. We call $L_i(x)$, $M_j(y)$, and their product, $L_i(x)M_j(y)$, the fundamental polynomials of Lagrange interpolation for the sets X, Y, and $X \times Y$, respectively. Since $L_i(x)M_j(y)$ is a fundamental polynomial for the set $X \times Y$, it is clear from (5.6) that $p(x, y)$ interpolates $f(x, y)$ on all $(m + 1) \times (n + 1)$ points of $X \times Y$. This interpolatory property of $p(x, y)$ should also be clear from the way we derived it via the double blending process. For $\xi(x, y)$ agrees with $f(x, y)$ along the $m + 1$ curves defined by the intersection of the planes $x = x_i$ with the surface $z = f(x, y)$, for $0 \le i \le m$, and $p(x, y)$ agrees with $\xi(x, y)$ along the $n+1$ curves defined by the intersection of the planes $y = y_j$ with the surface $z = \xi(x, y)$. Thus $p(x, y)$ agrees with $f(x, y)$ on all points in $X \times Y$.

Example 5.1.3 Let $X = Y = \{-1, 1\}$. Then the Cartesian product of X and Y is the set of points

$$\{(-1,-1),(-1,1),(1,-1),(1,1)\},$$

and the interpolating polynomial for a function f defined on these points is given by

$$\begin{aligned} p(x,y) = {} & f(-1,-1)\cdot\frac{1}{4}(x-1)(y-1) - f(-1,1)\cdot\frac{1}{4}(x-1)(y+1) \\ & -f(1,-1)\cdot\frac{1}{4}(x+1)(y-1) + f(1,1)\cdot\frac{1}{4}(x+1)(y+1), \end{aligned}$$

so that, for instance, $p(0,0)$ is just the arithmetic mean of the values of f on the four given points. If the four function values are all equal to some constant C, then we may easily verify that $p(x,y) = C$. Let us write $z = p(x,y)$. The coefficient of xy in the above expression for $p(x,y)$ is

$$\frac{1}{4}\big[f(-1,-1) - f(-1,1) - f(1,-1) + f(1,1)\big],$$

and it is not hard to see that this is zero if and only if the four points

$$\big(-1,-1,f(-1,-1)\big),\ \big(-1,1,f(-1,1)\big),\ \big(1,-1,f(1,-1)\big),\ \big(1,1,f(1,1)\big)$$

lie in a plane. In this special case, the surface $z = p(x,y)$ is a plane. Otherwise, we can divide $p(x,y)$ by the nonzero coefficient of xy (which corresponds to scaling the z-axis), giving an expression for $p(x,y)$ of the form

$$p(x,y) = xy + ax + by + c = (x+b)(y+a) + c - ab.$$

If we now change the origin from $(0,0,0)$ to $(-b,-a,c-ab)$, then in the new coordinate system, the surface $z = p(x,y)$ becomes $z = xy$. Shifting the origin and scaling (multiplying by a constant factor) does not change the shape of the original surface, which is a *hyperbolic paraboloid*. Although this is a curved surface, it has straight lines "embedded" in it, which are called *generators*. We can see this by looking at the above expression for $p(x,y)$. For if we replace y by a constant C, we see that $z = p(x,C)$ is the equation of a straight line. This shows that if we look at a "slice" of the surface $z = p(x,y)$, where the plane $y = C$ intersects the surface $z = p(x,y)$, we obtain the straight line $z = p(x,C)$. As we vary C we obtain an infinite system of generators parallel to the zx-plane. Similarly, by putting $x = C$, we obtain $z = p(C,y)$, revealing a second system of generators that are parallel to the yz-plane. ■

We will now obtain a divided difference form for the two-dimensional interpolating polynomial that is given in a Lagrangian form in (5.6). It is helpful to use the operator form of the divided differences, as in (1.20). We

need to use divided differences with respect to x, and divided differences with respect to y. Let us write

$$[x_0, \ldots, x_i]_x \, f$$

to denote the effect of the operator $[x_0, \ldots, x_i]$ acting on $f(x, y)$, regarded as a function of x, with the value of y being fixed. Thus, for example,

$$[x_0, x_1]_x f = \frac{f(x_1, y) - f(x_0, y)}{x_1 - x_0}.$$

Then, using Newton's form of the interpolating polynomial (1.19) applied to $f(x, y)$ as a function of x, with y fixed, we can rewrite (5.3) in the form

$$\xi(f; x, y) = \sum_{i=0}^{m} \pi_i(x) \, [x_0, \ldots, x_i]_x \, f, \tag{5.7}$$

where $\pi_i(x)$ is defined in (1.11). We can now apply Newton's form of the interpolating polynomial to $\xi(f; x, y)$, regarded as a function of y, with x fixed, giving

$$p(x, y) = \eta(\xi; x, y) = \sum_{j=0}^{n} \pi_j(y) \, [y_0, \ldots, y_j]_y \, \xi. \tag{5.8}$$

Henceforth, when there is no danger of ambiguity, we will write $[x_0, \ldots, x_i]_x$ more simply as $[x_0, \ldots, x_i]$, and similarly drop the suffix y from $[y_0, \ldots, y_j]_y$. We may combine (5.7) and (5.8) to give

$$p(x, y) = \sum_{j=0}^{n} \pi_j(y) \, [y_0, \ldots, y_j] \left(\sum_{i=0}^{m} \pi_i(x) \, [x_0, \ldots, x_i] f \right). \tag{5.9}$$

It is not difficult to see that the operators $[x_0, \ldots, x_i]$ and $[y_0, \ldots, y_j]$ *commute*; that is, they may be applied in either order to give the same result. (This is equivalent to the result that is the subject of Problem 5.1.2.) Thus the interpolating polynomial for f on the set $X \times Y$, which we wrote above in (5.6) in a Lagrange form, may now be expressed in the divided difference form

$$p(x, y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \pi_i(x) \pi_j(y) \, [x_0, \ldots, x_i] \, [y_0, \ldots, y_j] \, f. \tag{5.10}$$

Example 5.1.4 Let us write down the divided difference form (5.10) of the interpolating polynomial for the following data:

(x, y)	$(-1, -1)$	$(-1, 1)$	$(1, -1)$	$(1, 1)$
$f(x, y)$	1	5	-5	3

We compute

$$[-1,1]_x\, f(x,-1) = \frac{-5-1}{1+1} = -3,$$
$$[-1,1]_y\, f(-1,y) = \frac{5-1}{1+1} = 2,$$
$$[-1,1]_x\, f(x,1) = \frac{3-5}{1+1} = -1,$$

and so derive

$$[-1,1]_x[-1,1]_y\, f(x,y) = [-1,1]_x \left(\frac{f(x,1)-f(x,-1)}{1+1} \right) = \frac{-1+3}{1+1} = 1.$$

Hence we may write down the divided difference form of the interpolating polynomial,

$$p(x,y) = 1 - 3(x+1) + 2(y+1) + (x+1)(y+1),$$

which simplifies to give

$$p(x,y) = 1 - 2x + 3y + xy. \qquad \blacksquare$$

In the divided difference form (5.10), and in the Lagrange-type formula (5.6), the x_i are arbitrary distinct numbers that can be in any order, and the same holds for the y_j. Now let us consider the special case where both the x_i and the y_j are equally spaced, so that

$$x_i = x_0 + ih_x,\ 0 \le i \le m, \quad \text{and} \quad y_j = y_0 + jh_y,\ 0 \le j \le n,$$

where the values of h_x and h_y need not be the same. Following what we did in the one-dimensional case, we make the changes of variable

$$x = x_0 + sh_x \qquad \text{and} \qquad y = y_0 + th_y.$$

We define forward differences in the x-direction and forward differences in the y-direction:

$$\Delta_x f(x,y) = f(x+h_x) - f(x,y) \quad \text{and} \quad \Delta_y f(x,y) = f(x,y+h_y) - f(x,y).$$

We also define

$$\Delta_x \Delta_y f(x,y) = \Delta_x \left(\Delta_y f(x,y) \right).$$

We find (see Problem 5.1.5) that the two difference operators Δ_x and Δ_y commute, as we found above for the divided difference operators, so that

$$\Delta_x \Delta_y f(x,y) = \Delta_y \Delta_x f(x,y).$$

Then, for this "equally spaced" case, we can follow the method used for interpolation of a function of one variable (see Section 1.3) to transform the divided difference form (5.10) into the forward difference form

$$p(x_0 + sh_x, y_0 + th_y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \binom{s}{i} \binom{t}{j} \Delta_x^i \, \Delta_y^j f(x_0, y_0). \qquad (5.11)$$

In Section 1.5 we defined q-integers, and obtained a q-difference analogue of the univariate forward difference formula. We now derive a q-difference analogue of (5.11). Let us interpolate on the set

$$X \times Y, \quad \text{with} \quad X = \{x_0, \ldots, x_m\} \quad \text{and} \quad Y = \{y_0, \ldots, y_n\}, \qquad (5.12)$$

where

$$x_i = \frac{1 - {q_x}^i}{1 - q_x} = [i]_x \quad \text{and} \quad y_j = \frac{1 - {q_y}^j}{1 - q_y} = [j]_y. \qquad (5.13)$$

We also write

$$x = \frac{1 - {q_x}^s}{1 - q_x} = [s]_x \quad \text{and} \quad y = \frac{1 - {q_y}^t}{1 - q_y} = [t]_y,$$

for all real s and t. Note that we have chosen $q = q_x$ and $q = q_y$ as the bases of the q-integers for the x and y variables, respectively. We will denote q-differences with respect to x and y by Δ_x and Δ_y, respectively. (It should cause no confusion that we have used Δ_x and Δ_y above to denote forward differences, since these are just q-differences with $q = 1$.) Thus, as in (1.112), we have

$$\Delta_x^{k+1} f(x_i, y_j) = \Delta_x^k f(x_{i+1}, y_j) - q_x^k \, \Delta_x^k f(x_i, y_j) \qquad (5.14)$$

for $k \geq 0$, with

$$\Delta_x^0 f(x_i, y_j) = f(x_i, y_j) \quad \text{and} \quad \Delta_x^1 f(x_i, y_j) = \Delta_x f(x_i, y_j),$$

and

$$\Delta_y^{k+1} f(x_i, y_j) = \Delta_y^k f(x_i, y_{j+1}) - q_y^k \, \Delta_y^k f(x_i, y_j) \qquad (5.15)$$

for $k \geq 0$, with

$$\Delta_y^0 f(x_i, y_j) = f(x_i, y_j) \quad \text{and} \quad \Delta_y^1 f(x_i, y_j) = \Delta_y f(x_i, y_j).$$

The q-difference operators Δ_x and Δ_y commute, like their forward difference counterparts. As a consequence of (5.14), we have

$$\Delta_x^{k+1} \Delta_y^l f(x_i, y_j) = \Delta_x^k \Delta_y^l f(x_{i+1}, y_j) - q_x^k \Delta_x^k \Delta_y^l f(x_i, y_j),$$

and it follows from (5.15) that

$$\Delta_y^{l+1} \Delta_x^k f(x_i, y_j) = \Delta_y^l \Delta_x^k f(x_i, y_{j+1}) - q_y^l \Delta_y^l \Delta_x^k f(x_i, y_j).$$

Since the q-difference operators Δ_x and Δ_y commute, we have

$$\Delta_x^k \Delta_y^{l+1} f(x_i, y_j) = \Delta_x^k \Delta_y^l f(x_i, y_{j+1}) - q_y^l \Delta_y^l \Delta_y^l f(x_i, y_j).$$

Then, following the way we derived (1.115), we obtain

$$\pi_i([s]_x)\pi_j([t]_y)[x_0, \ldots, x_i]\,[y_0, \ldots, y_j]\,f = \left[\begin{array}{c} s \\ i \end{array} \right]_x \left[\begin{array}{c} t \\ j \end{array} \right]_y \Delta_x^i\, \Delta_y^j f(x_0, y_0),$$

where the suffixes x and y on the q-binomial coefficients indicate that they are based on q-integers with $q = q_x$ and $q = q_y$, respectively. By summing the last relation over i and j, we derive from (5.10) the interpolation formula

$$p([s]_x, [t]_y) = \sum_{i=0}^{m} \sum_{j=0}^{n} \left[\begin{array}{c} s \\ i \end{array} \right]_x \left[\begin{array}{c} t \\ j \end{array} \right]_y \Delta_x^i\, \Delta_y^j f(x_0, y_0). \tag{5.16}$$

Note that if we set $q_x = q_y = 1$ in (5.16), we obtain (5.11) with $x_0 = y_0 = 0$ and $h_x = h_y = 1$.

In this section we have shown how the Lagrange form, the Newton divided difference form, the forward difference formula, and the q-difference form of the interpolating polynomial can all be generalized from one dimension to rectangular arrays in two dimensions, and it is easy to see how all these ideas can be extended to boxlike regions in d-dimensions, that is, to interpolate on a set of points

$$X_1 \times X_2 \times \cdots \times X_d, \quad \text{where} \quad X_r = \left\{ x_1^{(r)}, x_2^{(r)}, \ldots, x_{j_r}^{(r)} \right\},\ 1 \le r \le d.$$

Example 5.1.5 Consider the circular cylinder defined by the set

$$C = \{(z, r, \theta) \mid 0 \le z \le 1,\ 0 \le r \le 1,\ 0 \le \theta < 2\pi\},$$

and let

$$Z = \{z_0, \ldots, z_l\}, \quad R = \{r_0, \ldots, r_m\}, \quad \Theta = \{\theta_0, \ldots, \theta_n\}.$$

Then define

$$p(z, r, \theta) = \sum_{i=0}^{l} \sum_{j=0}^{m} \sum_{k=0}^{n} L_i(z) M_j(r) N_k(\theta) f(z_i, r_j, \theta_k),$$

where $L_i(z)$, $M_j(r)$, and $N_k(\theta)$ are fundamental polynomials defined on Z, R, and Θ, respectively. We see that $p(z, r, \theta)$, a polynomial in each of the three variables z, r, and θ, interpolates $f(z, r, \theta)$ on the set defined by the Cartesian product $Z \times R \times \Theta$. ■

Because it is easy to extend one-dimensional interpolation processes to rectangular and boxlike sets of points, it is also easy to build on our knowledge of one-dimensional integration rules to construct integration rules over rectangular regions in two dimensions, and over boxlike regions in higher dimensions. Consider the integral

$$\int_c^d \left(\int_a^b f(x,y)dx \right) dy, \tag{5.17}$$

and let

$$X = \{x_0, x_1, \ldots, x_m\} \subset [a,b] \quad \text{and} \quad Y = \{y_0, y_1, \ldots, y_n\} \subset [c,d].$$

Then, we can replace the inner integral in (5.17) by an integration rule with weights $w_0, \ldots, w_m$, to give

$$\int_c^d \left(\int_a^b f(x,y)dx \right) dy \approx \int_c^d \left(\sum_{i=0}^m w_i f(x_i,y) \right) dy.$$

If we now apply an integration rule with weights $w'_0, \ldots, w'_n$ to the latter integral, we obtain

$$\int_c^d \left(\int_a^b f(x,y)dx \right) dy \approx \sum_{j=0}^n w'_j \left(\sum_{i=0}^m w_i f(x_i,y_j) \right),$$

giving the integration rule

$$\int_c^d \left(\int_a^b f(x,y)dx \right) dy \approx \sum_{i=0}^m \sum_{j=0}^n w_i w'_j f(x_i,y_j). \tag{5.18}$$

Thus the weight corresponding to the point (x_i, y_j) is just the product of the weights w_i and w'_j, and an integration rule of the form (5.18) is called a *product* rule. If we have error terms for the one-dimensional rules involving the w_i and the w'_j, it is not difficult to derive an error term for the rule given by (5.18). It is also easy to see that we can extend these ideas to higher dimensions.

Example 5.1.6 Let us estimate the integral

$$I = \int_{-1}^1 \int_{-1}^1 \frac{dx\,dy}{1+x^2+y^2},$$

using product rules. Let us choose $m = n$ in (5.18), and choose $w_i = w'_i$ as the weights of a Gaussian rule. The simplest is the one-point Gaussian

rule (3.61), also called the midpoint rule. For any integrand $f(x, y)$, the one-point Gaussian rule gives

$$\int_{-1}^{1} f(x, y)\, dx\, dy \approx 4f(0, 0),$$

and hence $I \approx 4$, which is not very accurate. On applying the two-point, three-point, and four-point Gaussian rules (see (3.62), (3.63), and Problem 3.3.4), which require 4, 9, and 16 evaluations of the integrand, respectively, we obtain 2.4, 2.586, and 2.554, respectively, as estimates of I. We could apply any of the above rules, or any other one-dimensional rule, in composite form. The midpoint rule is the easiest to apply in composite form, since the weights are all equal. With 4, 100, and 1600 evaluations of the integrand, the composite midpoint rule gives 2.667, 2.563, and 2.558, respectively, and the last estimate is correct to three decimal places. ■

As we have seen, we can use product rules to integrate over a region that is a Cartesian product of one-dimensional regions. Examples of such regions are the cylinder, which we encountered in Example 5.1.5, the square, the cube, and the hypercube in higher dimensions. Another approach, which works whether the region is a Cartesian product or not, is to design integration rules that are exact for certain monomials.

Example 5.1.7 Consider an integration rule with 13 abscissas, $(0, 0)$ with weight w_1, $(\pm r, 0)$ and $(0, \pm r)$ with weight w_2, and $(\pm s, \pm t)$, $(\pm t, \pm s)$ with weight w_3, for the square $S = \{(x, y) \mid -1 \le x, y \le 1\}$. We seek values of the parameters r, s, t, w_1, w_2, and w_3 such that the rule is exact for all monomials $x^i y^j$ with $i + j \le 7$, and so it is said to be of degree 7. Any rule with these abscissas is exact for all monomials $x^i y^j$ where at least one of i and j is odd. Thus the rule will be exact for all monomials $x^i y^j$ with $i + j \le 7$ if and only if it is exact for 1, x^2, x^4, $x^2 y^2$, and x^6. This gives the equations

$$\begin{aligned}
w_1 + 4w_2 + 8w_3 &= 4, \\
2r^2 w_2 + 4(s^2 + t^2)w_3 &= \frac{4}{3}, \\
2r^4 w_2 + 4(s^4 + t^4)w_3 &= \frac{4}{5}, \\
8s^2 t^2 w_3 &= \frac{4}{9}, \\
2r^6 w_2 + 4(s^6 + t^6)w_3 &= \frac{4}{7}.
\end{aligned}$$

It may be verified that these equations are satisfied by

$$r^2 = \frac{12}{35}, \quad s^2 = \frac{93+3\sqrt{186}}{155}, \quad t^2 = \frac{93-3\sqrt{186}}{155},$$

$$w_1 = \frac{4}{81}, \quad w_2 = \frac{49}{81}, \qquad w_3 = \frac{31}{162}.$$

Note that all of the abscissas lie within the square S, and the weights are all positive. This rule was published by the distinguished mathematician and physicist James Clerk Maxwell (1831–1879) in 1877. ■

Example 5.1.8 Let $C = \{(x, y, z) \,|\, -1 \le x, y, z \le 1\}$ denote the cube with side 2 and centre at the origin. Consider an integration rule for C with 14 abscissas, namely, the six abscissas $(\pm r, 0, 0)$, $(0, \pm r, 0)$, and $(0, 0, \pm r)$ with weight w_1, and the eight abscissas $(\pm s, \pm s, \pm s)$ with weight w_2. We seek values of the parameters r, s, w_1, and w_2 such that the rule is exact for all monomials $x^i y^j z^k$with $i + j + k \le 5$, and so it is said to be of degree 5. Any rule with these abscissas is exact for all monomials $x^i y^j z^k$ where at least one of i, j, and j is odd. Thus, because of the symmetry in x, y, and z, the rule will be exact for all monomials $x^i y^j z^k$ with $i+j+k \le 5$ if and only if it is exact for 1, x^2, x^4, and x^2y^2. This gives the equations

$$\begin{aligned} 6w_1 + 8w_2 &= 8, \\ 2r^2 w_1 + 8s^2 w_2 &= \frac{8}{3}, \\ 2r^4 w_1 + 8s^4 w_2 &= \frac{8}{5}, \\ 8s^4 w_2 &= \frac{8}{9}. \end{aligned}$$

We can verify that the above equations are satisfied by

$$r^2 = \frac{19}{30}, \quad s^2 = \frac{19}{33}, \quad w_1 = \frac{320}{361}, \quad w_2 = \frac{121}{361}.$$

Note that the abscissas all lie within the cube C, and the weights are positive. This rule was published by Hammer and Stroud in 1958. ■

A large number of integration rules for the square and the cube, including those given in Examples 5.1.7 and 5.1.8, and also for the general hypercube, the circle, sphere, and general hypersphere, the triangle, tetrahedron, and general simplex, and other regions, are to be found in the book by Stroud [53].

Problem 5.1.1 Show that if in Example 5.1.1 we choose the four interpolating abscissas as $(0, 0)$, $(1, 0)$, $(0, 1)$, and $(1, 1)$, we can always construct a unique interpolating function of the form

$$p(x, y) = a_1 + a_2 x + a_3 y + a_4 xy.$$

Problem 5.1.2 Verify that

$$p(x,y) = \xi(\eta; x, y) = \eta(\xi; x, y),$$

where $p(x,y)$ is the interpolating polynomial for $f(x,y)$ on the set $X \times Y$, and $\xi(f; x, y)$ and $\eta(f; x, y)$ are the blending functions defined in (5.3) and (5.4), respectively.

Problem 5.1.3 Extend the result obtained in Example 5.1.2 to find the polynomial $p(x,y)$ that interpolates the function 2^{x+y} on the set $X \times Y$, where $X = Y = \{-1, 0, 1\}$.

Problem 5.1.4 Verify that

$$\det \begin{bmatrix} 1 & x_0 & y_0 & x_0y_0 \\ 1 & x_1 & y_0 & x_1y_0 \\ 1 & x_0 & y_1 & x_0y_1 \\ 1 & x_1 & y_1 & x_1y_1 \end{bmatrix} = (x_1 - x_0)^2(y_1 - y_0)^2,$$

and thus prove that there is a unique interpolating function of the form

$$p(x,y) = a_1 + a_2x + a_3y + a_4xy$$

on the set of points $\{(x_0, y_0), (x_1, y_0), (x_0, y_1), (x_1, y_1)\}$, provided that x_0 and x_1 are distinct, and y_0 and y_1 are distinct.

Problem 5.1.5 Verify that if Δ_x and Δ_y are q-differences or forward differences, then

$$\begin{aligned} \Delta_x\Delta_y f(x_i, y_j) &= f(x_{i+1}, y_{j+1}) - f(x_i, y_{j+1}) - f(x_{i+1}, y_j) + f(x_i, y_j) \\ &= \Delta_y\Delta_x f(x_i, y_j), \end{aligned}$$

so that the operators Δ_x and Δ_y commute. Deduce that

$$\Delta_x^k\Delta_y^l f(x_i, y_j) = \Delta_y^l\Delta_x^k f(x_i, y_j),$$

for any integers $k, l \geq 0$.

Problem 5.1.6 Derive an integration rule with 8 abscissas, of the form $(\pm r, 0)$ and $(0, \pm r)$, with weight w_1, and $(\pm s, \pm s)$, with weight w_2, for the square $S = \{(x,y) \mid -1 \leq x, y \leq 1\}$, that is exact for all monomials x^iy^j with $i + j \leq 5$, and so is of degree 5. Argue that the rule must be exact for the monomials 1, x^2, x^4, and x^2y^2, giving the equations

$$\begin{aligned} 4w_1 + 4w_2 &= 4, \\ 2r^2w_1 + 4s^2w_2 &= \frac{4}{3}, \\ 2r^4w_1 + 4s^4w_2 &= \frac{4}{5}, \\ 4s^4w_2 &= \frac{4}{9}. \end{aligned}$$

Verify that these are satisfied by

$$r^2 = \frac{7}{15}, \quad s^2 = \frac{7}{9}, \qquad w_1 = \frac{40}{49}, \quad w_2 = \frac{9}{49}.$$

Problem 5.1.7 Let $C = \{(x, y, z) \mid -1 \le x, y, z \le 1\}$. Consider the 15-point integration rule for the cube C that has the abscissa $(0, 0, 0)$ with weight $\frac{16}{9}$, the six abscissas $(\pm 1, 0, 0)$, $(0, \pm 1, 0)$, and $(0, 0, \pm 1)$ with weight $\frac{8}{9}$, and the eight abscissas $(\pm 1, \pm 1, \pm 1)$ with weight $\frac{1}{9}$. Verify that this rule is exact for all monomials $x^i y^j z^k$ with $i + j + k \le 3$.

5.2 Triangular Regions

In this section we will explore interpolation methods on triangular sets of abscissas, and indicate generalizations to higher dimensions. The simplest case is to construct an interpolating polynomial in x and y of the form $a_1 + a_2 x + a_3 y$ for a function f on the three points (x_0, y_0), (x_1, y_0), and (x_1, y_1), where x_0 and x_1 are distinct, and y_0 and y_1 are distinct. Since

$$\det \begin{bmatrix} 1 & x_0 & y_0 \\ 1 & x_1 & y_0 \\ 1 & x_1 & y_1 \end{bmatrix} = (x_1 - x_0)(y_1 - y_0), \tag{5.19}$$

the above determinant is nonzero when x_0 and x_1 are distinct and y_0 and y_1 are distinct, and then the above interpolation problem has a unique solution. This follows from the geometrical property that there is a unique plane passing through three noncollinear points in three-dimensional Euclidean space. Having thus started with a polynomial of degree one in x and y that interpolates $f(x, y)$ at three noncollinear points, it is natural to consider next a polynomial of total degree two in x and y, of the form

$$a_1 + a_2 x + a_3 y + a_4 x^2 + a_5 xy + a_6 y^2.$$

As we will see, we can choose values of the six coefficients so that this polynomial interpolates $f(x, y)$ uniquely at six appropriately chosen points. In this way, we are led inevitably to triangular numbers of coefficients and interpolating points. The nth triangular number is the sum of the first n natural numbers, the first four being

$$1, \quad 1 + 2 = 3, \quad 1 + 2 + 3 = 6, \quad 1 + 2 + 3 + 4 = 10.$$

The monomials in x and y of total degree j are

$$x^j,\ x^{j-1}y,\ x^{j-2}y^2,\ \ldots,\ x^2 y^{j-2},\ xy^{j-1},\ y^j,$$

						y^3				
			y^2			y^2	xy^2			
	y		y	xy		y	xy	x^2y		
1	1	x	1	x	x^2	1	x	x^2	x^3	

TABLE 5.1. The monomials in x and y of degree 0, 1, 2, and 3. Each diagonal contains the monomials of the same total degree in x and y.

and there are $j+1$ of these. Thus the number of monomials of total degree not greater than n is the $(n+1)$th triangular number

$$1+2+\cdots+(n+1)=\frac{1}{2}(n+1)(n+2)=\left(\begin{array}{c} n+2 \\ 2 \end{array}\right).$$

The monomials in x and y of total degree not greater than n, for $0 \le n \le 3$, are depicted in Table 5.1, and the layout of this table shows why it is natural to seek an interpolating polynomial in x and y of total degree n to match a function on a triangular array of $1+2+\cdots+(n+1)$ abscissas.

Example 5.2.1 Let us consider the problem of constructing an interpolating polynomial for a given function f on the six points defined by

$$(x_i, y_j),\ \ i,j \ge 0,\ i+j \le 2,$$

where the x_i are distinct and the y_j are distinct. We will show that

$$\det\begin{bmatrix} 1 & x_0 & y_0 & x_0^2 & x_0y_0 & y_0^2 \\ 1 & x_1 & y_0 & x_1^2 & x_1y_0 & y_0^2 \\ 1 & x_2 & y_0 & x_2^2 & x_2y_0 & y_0^2 \\ 1 & x_0 & y_1 & x_0^2 & x_0y_1 & y_1^2 \\ 1 & x_1 & y_1 & x_1^2 & x_1y_1 & y_1^2 \\ 1 & x_0 & y_2 & x_0^2 & x_0y_2 & y_2^2 \end{bmatrix} = -\psi(x_0,x_1,x_2)\psi(y_0,y_1,y_2), \quad (5.20)$$

where

$$\psi(x_0,x_1,x_2)=(x_2-x_0)(x_2-x_1)(x_1-x_0)^2.$$

First, an argument similar to that used in Problem 1.1.1 shows that x_2-x_0 and x_2-x_1 are factors of this determinant. By putting $x_0=x_1$ in rows 1 and 2, and again in rows 4 and 5, we see that $(x_1-x_0)^2$ is a factor, and thus $\psi(x_0,x_1,x_2)$ is a factor of the determinant. (To see that the factor x_1-x_0 occurs twice, we could replace x_0 in row 4 by x_0', say, and replace x_1 in row 5 by x_1'. We then argue that $x_1'-x_0'$ must be a factor. Then, letting x_0' and x_1' tend to x_0 and x_1, respectively, we see that the factor x_1-x_0 occurs twice in the expansion of the determinant.) Similarly, we find that $\psi(y_0,y_1,y_2)$ is a factor. Since both sides of (5.20) are polynomials of the same degree in the same variables, (5.20) must be correct to within a multiplicative constant, and it will suffice to compare the coefficients of

$x_1^2x_2^2y_1^2y_2^2$, say, on both sides. The coefficient on the left is the same as the coefficient of $x_1^2x_2^2y_1^2$ in the 5×5 determinant obtained by deleting the sixth row and column of the above determinant. This, in turn, is the same as the coefficient of $-x_1^2y_1^2$ in the 4×4 determinant obtained by deleting the third and sixth rows and the fourth and sixth columns of the above determinant. This 4×4 determinant is that which appears in Problem 5.1.4, and we see that the coefficient of $-x_1^2y_1^2$ is indeed -1. This shows that there is a unique polynomial of total degree in x and y not greater than two that matches a given function f on the six points defined above. ■

The following theorem is concerned with a generalization of the result in Example 5.2.1.

Theorem 5.2.1 Given any positive integer n and a set of points

$$S_\Delta^n = \{(x_i, y_j) \mid i, j \geq 0,\ i + j \leq n\}, \tag{5.21}$$

where the x_i are distinct and the y_j are distinct, there is a unique polynomial of the form

$$p_n(x, y) = \sum_{k=0}^{n} \sum_{r=0}^{k} c_{r,k-r} x^r y^{k-r}$$

that takes the same values as a given function $f(x, y)$ on the set S_Δ^n.

Proof. Let $\mathbf{A}_n$ denote the square matrix each of whose $\frac{1}{2}(n+1)(n+2)$ rows consists of the $\frac{1}{2}(n+1)(n+2)$ elements

$$1,\ x,\ y,\ x^2,\ xy,\ y^2,\ \ldots,\ xy^{n-1},\ y^n$$

evaluated at the points (x_i, y_j) of the set S_Δ^n, taken in the order

$$\begin{array}{l}(x_0, y_0), (x_1, y_0), \cdots (x_{n-1}, y_0), (x_n, y_0),\\ (x_0, y_1), (x_1, y_1), \cdots (x_{n-1}, y_1),\\ \cdots\\ \cdots\\ (x_0, y_{n-1}), (x_1, y_{n-1}),\\ (x_0, y_n).\end{array}$$

The matrix $\mathbf{A}_1$ is the 3×3 matrix that appears in (5.19), and $\mathbf{A}_2$ is the 6×6 matrix in Example 5.2.1. We can extend the method used in Example 5.2.1 to show (see Problem 5.2.3) that the matrix $\mathbf{A}_n$ is nonsingular for a general value of n. Thus there is a unique interpolating polynomial of total degree n in x and y that interpolates a given function $f(x, y)$ on the set S_Δ^n. ■

The set of points S_Δ^n defined by (5.21) lie in a triangular formation when $x_j = y_j = j$. However, the set of points S_Δ^n may lie in a formation that bears no resemblance to a triangle, as shown by the formation in Figure 5.1, in which $n = 2$.

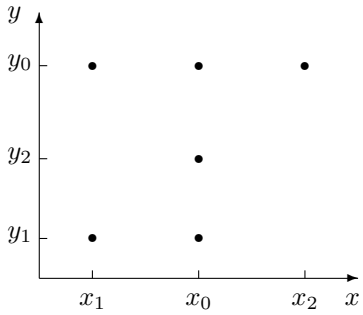

FIGURE 5.1. An example of the set S_Δ^2, defined by (5.21).

Having shown in Theorem 5.2.1 that there is a unique polynomial $p_n(x, y)$ of total degree n in x and y that interpolates a given function $f(x, y)$ on S_Δ^n, we now show how to construct this interpolating polynomial and obtain an error term. We achieve this by generalizing the one-dimensional relation (1.32),

$$f(x) = p_n(x) + (x - x_0) \cdots (x - x_n) f[x, x_0, x_1, \ldots, x_n], \tag{5.22}$$

which we obtained by repeatedly applying (1.29),

$$f[x, x_0, \ldots, x_{n-1}] = f[x_0, \ldots, x_n] + (x - x_n) f[x, x_0, \ldots, x_n]. \tag{5.23}$$

When we derived (5.22) we already knew from our study of Newton's divided difference formula that the polynomial $p_n(x)$ interpolates $f(x)$ on the set of points $\{x_0, x_1, \ldots, x_n\}$. We can deduce directly from (5.22) that $p_n(x)$ is the interpolating polynomial for $f(x)$. For we see from (5.23) that

$$(x - x_n) f[x, x_0, \ldots, x_n] = f[x, x_0, \ldots, x_{n-1}] - f[x_0, \ldots, x_n],$$

and this is zero when $x = x_n$, since a divided difference is unaltered when we change the order of its parameters. If we interchange x_n and x_j, we also have

$$(x - x_j) f[x, x_0, \ldots, x_n] = f[x, x_0, \ldots, x_{j-1}, x_{j+1}, \ldots, x_n] - f[x_0, \ldots, x_n] = 0$$

when $x = x_j$, for $0 < j < n$, and we have a similar result when $j = 0$. It follows that

$$(x - x_0) \cdots (x - x_n) f[x, x_0, x_1, \ldots, x_n] = 0$$

for $x = x_j,\ 0 \le j \le n$, and thus the polynomial $p_n(x)$ that is defined by (5.22) interpolates $f(x)$ on the set of points $\{x_0, x_1, \ldots, x_n\}$.

We now give a generalization of the Newton divided difference formula plus error term (5.22) from the one-dimensional set $\{x_0, x_1, \ldots, x_n\}$ to the two-dimensional set S_Δ^n.

Theorem 5.2.2 Given any function f defined on some subset of $\mathbb{R}^2$ that includes S^n_Δ, let

$$f(x,y) = p_m(x,y) + r_m(x,y), \qquad 0 \le m \le n, \tag{5.24}$$

where the sequence of polynomials (p_m) is defined recursively by

$$p_m(x,y) = p_{m-1}(x,y) + q_m(x,y), \qquad m \ge 1, \tag{5.25}$$

with

$$q_m(x,y) = \sum_{k=0}^{m} \pi_k(x)\pi_{m-k}(y)\,[x_0,\ldots,x_k]\,[y_0,\ldots,y_{m-k}]\,f, \tag{5.26}$$

for $m \ge 1$, beginning with

$$p_0(x,y) = [x_0]\,[y_0]\,f = f(x_0,y_0). \tag{5.27}$$

Then, for $m \ge 0$, the error term $r_m(x,y)$ satisfies

$$\begin{aligned} r_m(x,y) = &\sum_{k=0}^{m} \pi_{k+1}(x)\pi_{m-k}(y)\,[x,x_0,\ldots,x_k]\,[y_0,\ldots,y_{m-k}]\,f \\ &+ \pi_{m+1}(y)\,[x]\,[y,y_0,\ldots,y_m]\,f, \end{aligned} \tag{5.28}$$

and the polynomial $p_m(x,y)$ interpolates $f(x,y)$ on the set S^m_Δ.

Proof. We will verify by induction that (5.24) and (5.28) hold for all m, and then justify that $p_n(x,y)$ interpolates $f(x,y)$ on the set S^n_Δ. It is easily verified, by simplifying the right side of the equation, that

$$f(x,y) = [x_0]\,[y_0]\,f + (x-x_0)\,[x,x_0]\,[y_0]\,f + (y-y_0)\,[x]\,[y,y_0]\,f, \tag{5.29}$$

and so (5.24) and (5.28) both hold for $m = 0$. Let us assume that (5.24) and (5.28) both hold for some $m \ge 0$. Now, using (5.23), we split the summation in (5.28) into two sums, S_1 and S_2, and then express the final single term on the right side of (5.28) as the sum of the three terms T_1, T_2, and T_3, defined below. The sums S_1 and S_2, which are obtained by writing

$$[x,x_0,\ldots,x_k]f = [x_0,\ldots,x_{k+1}]f + (x-x_{k+1})[x,x_0,\ldots,x_{k+1}]f,$$

are

$$S_1 = \sum_{k=0}^{m} \pi_{k+1}(x)\pi_{m-k}(y)\,[x_0,\ldots,x_{k+1}]\,[y_0,\ldots,y_{m-k}]\,f$$

and

$$S_2 = \sum_{k=0}^{m} \pi_{k+2}(x)\pi_{m-k}(y)\,[x,x_0,\ldots,x_{k+1}]\,[y_0,\ldots,y_{m-k}]\,f,$$

and, similarly, we can write

$$T_1 = \pi_{m+1}(y)\,[x_0]\,[y_0, \ldots, y_{m+1}]\,f,$$
$$T_2 = \pi_1(x)\pi_{m+1}(y)\,[x, x_0]\,[y_0, \ldots, y_{m+1}]\,f,$$
$$T_3 = \pi_{m+2}(y)\,[x]\,[y, y_0, \ldots, y_{m+1}]\,f.$$

It follows immediately from (5.26) and (5.25) that

$$S_1 + T_1 = q_{m+1}(x, y) = p_{m+1}(x, y) - p_m(x, y). \tag{5.30}$$

If we now replace k by $k-1$ in the above expression for S_2, we have

$$S_2 = \sum_{k=1}^{m+1} \pi_{k+1}(x)\pi_{m+1-k}(y)\,[x, x_0, \ldots, x_k]\,[y_0, \ldots, y_{m+1-k}]\,f,$$

and we see that

$$S_2 + T_2 + T_3 = r_{m+1}(x, y). \tag{5.31}$$

It then follows from (5.30) and (5.31) that

$$r_m(x, y) = S_1 + S_2 + T_1 + T_2 + T_3 = r_{m+1}(x, y) + q_{m+1}(x, y). \tag{5.32}$$

Thus

$$p_m(x, y) + r_m(x, y) = (p_m(x, y) + q_{m+1}(x, y)) + r_{m+1}(x, y),$$

so that

$$f(x, y) = p_m(x, y) + r_m(x, y) = p_{m+1}(x, y) + r_{m+1}(x, y),$$

and by induction, this completes the main part of the proof. The interpolation property of $p_n(x, y)$ is easily verified by showing that with $m = n$ in (5.28), each of the $n + 2$ terms in (5.28) is zero for every $(x, y) \in S_\Delta^n$, so that the error term $r_n(x, y)$ is zero on S_Δ^n. ■

Note that in view of the relation between divided differences and derivatives (see (1.33)) we can construct, from (5.28) with $m = n$, an error term for interpolating the function f on the set S_Δ^n that involves the $n+2$ partial derivatives of f of total order $n + 1$; that is,

$$\frac{\partial^{n+1} f}{\partial x^{n+1}},\ \frac{\partial^{n+1} f}{\partial x^n \partial y},\ \ldots,\ \frac{\partial^{n+1} f}{\partial x \partial y^n},\ \frac{\partial^{n+1} f}{\partial y^{n+1}},$$

provided that these derivatives all exist.

We see from (5.25), (5.26), and (5.27) that the interpolating polynomial on S_Δ^n can be expressed explicitly in the form

$$p_n(x, y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \pi_k(x)\pi_{m-k}(y)[x_0, \ldots, x_k][y_0, \ldots, y_{m-k}]f. \tag{5.33}$$

The divided difference form in (5.33) for interpolation on the triangular set S_Δ^n complements the similar formula obtained in (5.10) for interpolation on a rectangular grid. Let us now examine separately the special case of the set S_Δ^n where the x_i and y_j are equally spaced, so that

$$x_i = x_0 + ih_x \quad \text{and} \quad y_j = y_0 + jh_y, \tag{5.34}$$

for $0 \le i, j \le n$, where the values of h_x and h_y need not be the same. By making a linear change of variable in x and a linear change of variable in y we can put $x_0 = y_0 = 0$ and $h_x = h_y = 1$, so that the set S_Δ^n defined by (5.21) becomes, say, S^n, where

$$S^n = \{(i, j) \mid i, j \ge 0,\ i + j \le n\}. \tag{5.35}$$

Beginning with the divided difference form for $p_n(x, y)$ on S_Δ^n, given by (5.33), we can deduce a forward difference form for the interpolating polynomial for $f(x, y)$ on the triangular grid S^n. We follow the same method as we adopted in the one-variable case, in Section 1.3, where we showed (see (1.73)) that

$$f[j, j + 1, \ldots, j + k] = \frac{1}{k!}\,\Delta^k f(j).$$

For interpolation on S^n, we need to evaluate

$$\pi_k(x) = x(x - 1) \cdots (x - k + 1)$$

for $k > 0$, with $\pi_0(x) = 1$, and

$$[0, 1, \ldots, k]_x\,[0, 1, \ldots, m - k]_y\, f = \frac{\Delta_x^k \Delta_y^{m-k} f(0, 0)}{k!(m - k)!}.$$

Then it follows from (5.33) that

$$p_n(x, y) = \sum_{m=0}^{n} \sum_{k=0}^{m} \binom{x}{k} \binom{y}{m-k} \Delta_x^k \Delta_y^{m-k} f(0, 0). \tag{5.36}$$

If we put $y = 0$ in (5.36), we see that $p_n(x, 0)$ is expressed as the one-dimensional forward difference formula (1.74) for the function $f(x, 0)$ on the point set $\{0, 1, \ldots, n\}$. Likewise, if we put $x = 0$ in (5.36), $p_n(0, y)$ is given by the one-dimensional forward difference formula for $f(0, y)$ on the point set $\{0, 1, \ldots, n\}$.

We will now obtain a Lagrange form of the polynomial $p_n(x, y)$ that interpolates a given function f on S^n. To achieve this, we need to find a fundamental polynomial $L_{i,j}(x, y)$ of degree at most n in x and y that takes the value 1 at $(x, y) = (i, j)$ and the value zero at all the other points in the set S^n. For example, let us seek the fundamental polynomial $L_{4,0}(x, y)$

for the set of interpolating points depicted in Figure 5.2, which is the set S^4. We see that the polynomial

$$x(x-1)(x-2)(x-3)$$

is zero at all the interpolating points except $(4,0)$, since all of the other points lie on one of the four lines whose equations are

$$x = 0, \quad x-1 = 0, \quad x-2 = 0, \quad x-3 = 0. \tag{5.37}$$

Then we can scale the above polynomial to give

$$L_{4,0}(x,y) = \frac{1}{24}x(x-1)(x-2)(x-3),$$

which indeed takes the value 1 when $(x,y) = (4,0)$ and the value zero on all the other points in S^4. We now make use of the fact that the point (i,j) lies on three lines, one from each of the three systems of parallel lines that appear in Figure 5.2. One system is that parallel to the y-axis, already

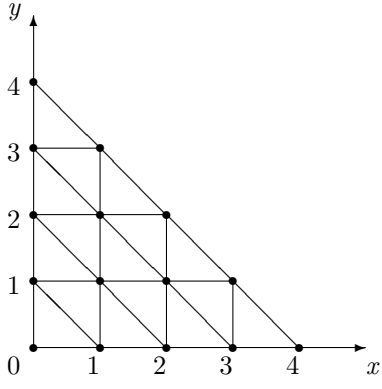

FIGURE 5.2. A triangular interpolation grid.

given in (5.37); there is also a system of lines parallel to the x-axis,

$$y = 0, \quad y-1 = 0, \quad y-2 = 0, \quad y-3 = 0, \tag{5.38}$$

and a system parallel to the third side of the triangle,

$$x+y-1 = 0, \quad x+y-2 = 0, \quad x+y-3 = 0, \quad x+y-4 = 0. \tag{5.39}$$

The point $(1,2)$, for example, has the line $x = 0$ to the left of it, that is, moving toward the y-axis. (See Figure 5.3.) It also has the lines $y = 0$ and $y-1 = 0$ below it, moving toward the x-axis, and has the line $x+y-4$ in the direction of the third side of the triangle. Thus the polynomial that is

the product of the left sides of these four equations, $xy(y-1)(x+y-4)$, is zero on all points in S^4 except for $(1,2)$. On scaling this polynomial, we find that

$$L_{1,2}(x,y) = -\frac{1}{2}xy(y-1)(x+y-4)$$

is the fundamental polynomial for $(1,2)$, since it has the value 1 at $(1,2)$ and is zero on all the other points.

This gives us sufficient insight to derive the fundamental polynomials for all points in the set S^n. These points are all contained in the triangle defined by the x-axis, the y-axis, and the line whose equation is $x+y-n=0$. Given any (i,j) in S^n, consider the following three sets of lines associated with the point (i,j).

1. Lines of the form $x-k=0$ that lie between (i,j) and the side of the triangle formed by the y-axis.

2. Lines of the form $y-k=0$ that lie between (i,j) and the side of the triangle formed by the x-axis.

3. Lines of the form $x+y-k=0$ that lie between (i,j) and the third side of the triangle, defined by the equation $x+y-n=0$.

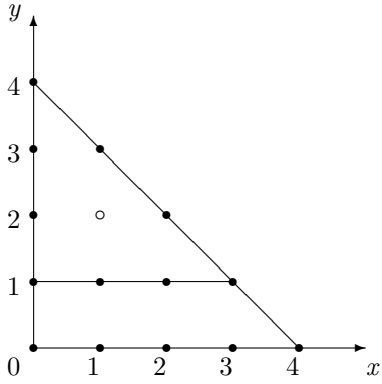

FIGURE 5.3. How the fundamental polynomial for $(1,2)$ is constructed.

There are no lines in the first of the three sets enumerated above if $i=0$, and if $i>0$, we have the lines

$$x=0,\ x-1=0,\ \ldots,\ x-i+1=0.$$

If $j=0$, there are no lines in the second set, and if $j>1$, we have the lines

$$y=0,\ y-1=0,\ \ldots,\ y-j+1=0.$$

If $i + j = n$, the point (i, j) is *on* the line $x + y - n = 0$ and there are no lines in the third set; otherwise, we have, working toward the third side of the triangle, the lines

$$x + y - i - j - 1 = 0, \quad x + y - i - j - 2 = 0, \ \ldots\ , \ x + y - n = 0.$$

Note that the total number of lines in the three sets enumerated above is

$$i + j + (n - i - j) = n.$$

Now if we draw all of these n lines that are associated with the given point (i, j) on a grid like that of Figure 5.2, as we did in Figure 5.3 for the point $(1, 2)$, we see that between them they cover all the points on the triangular grid except for the point (i, j). Thus, taking the product of the left sides of all these n equations, we see that

$$\prod_{s=0}^{i-1}(x - s)\prod_{s=0}^{j-1}(y - s)\prod_{s=i+j+1}^{n}(x + y - s) \tag{5.40}$$

is zero at all points on the triangular grid except for the point (i, j). If $i = 0$ or $j = 0$ or $i + j = n$, the corresponding product in (5.40) is said to be empty, and its value is defined to be 1. We then just need to scale the polynomial defined by this triple product to give

$$L_{i,j}(x, y) = \prod_{s=0}^{i-1}\left(\frac{x - s}{i - s}\right)\prod_{s=0}^{j-1}\left(\frac{y - s}{j - s}\right)\prod_{s=i+j+1}^{n}\left(\frac{x + y - s}{i + j - s}\right), \tag{5.41}$$

which simplifies to give

$$L_{i,j}(x, y) = \binom{x}{i}\binom{y}{j}\binom{n - x - y}{n - i - j}, \tag{5.42}$$

the fundamental polynomial corresponding to the point (i, j), where it takes the value 1. Thus the interpolating polynomial for a function $f(x, y)$ on the triangular grid defined by (5.35) is given by

$$p_n(x, y) = \sum_{i,j} f(i, j)L_{i,j}(x, y), \tag{5.43}$$

where the above summation is over all nonnegative integers i and j such that $i + j \le n$. Note from (5.41) that the numerator of each fundamental polynomial is a product of n factors, and so the interpolating polynomial $p_n(x, y)$ is a polynomial of total degree at most n in x and y.

Let us write λ_k, μ_k, and ν_k to denote the lines $x - k = 0$, $y - k = 0$, and $x + y - (n - k) = 0$, respectively. Then the point (i, j) lies on each of the lines λ_i, μ_j, and ν_{n-i-j}. We could use the notation $\{\lambda_i, \mu_j, \nu_{n-i-j}\}$ to denote the point (i, j), to emphasize that it lies on these three lines.

Example 5.2.2 When $n = 2$ in (5.43) we have six interpolating points, and the interpolating polynomial is

$$\begin{aligned} p_2(x,y) = \frac{1}{2}(2-x-y)(1-x-y)\,f(0,0) \;+\; x(2-x-y)\,f(1,0) \\ + y(2-x-y)\,f(0,1) \;+\; \frac{1}{2}x(x-1)\,f(2,0) \\ + xy\,f(1,1) \;+\; \frac{1}{2}y(y-1)\,f(0,2). \end{aligned}$$

If we evaluate $p_2(x,y)$ at the centroid of the triangle with vertices $(0,0)$, $(2,0)$, and $(0,2)$, we find that

$$p_2\left(\frac{2}{3},\frac{2}{3}\right) = \frac{1}{3}(4\alpha - \beta),$$

where β is the mean of the values of f on the vertices of the triangle, and α is the mean of the values of f on the other three points. ■

In Section 1.1 we saw that the interpolating polynomial in one variable can be evaluated iteratively by the Neville–Aitken algorithm. We can also derive an iterative process of Neville–Aitken type for evaluating the interpolating polynomial for $f(x,y)$ on the triangular set of points defined above in (5.35). Let us define $p_k^{[i,j]}(x,y)$ as the interpolating polynomial for $f(x,y)$ on the triangular set of points

$$S_k^{[i,j]} = \{(i+r, j+s) \mid r,s \geq 0,\ r+s \leq k\}. \tag{5.44}$$

The set $S_k^{[i,j]}$ contains $1 + 2 + \cdots + (k+1) = \frac{1}{2}(k+1)(k+2)$ points arranged in a right-angled triangle formation, with (i,j) as the bottom left-hand point. Figure 5.2 illustrates the set $S_4^{[0,0]}$. Thus $p_0^{[i,j]}(x,y)$ has the constant value $f(i,j)$. We can compute the interpolating polynomials $p_k^{[i,j]}(x,y)$ recursively in a Neville–Aitken style, as we will now see.

Theorem 5.2.3 For $k \geq 0$ and $i,j \geq 0$,

$$\begin{aligned} p_{k+1}^{[i,j]}(x,y) = \left(\frac{i+j+k+1-x-y}{k+1}\right) p_k^{[i,j]}(x,y) \\ + \left(\frac{x-i}{k+1}\right) p_k^{[i+1,j]}(x,y) + \left(\frac{y-j}{k+1}\right) p_k^{[i,j+1]}(x,y). \end{aligned} \tag{5.45}$$

Proof. From its definition above, each $p_0^{[i,j]}(x,y)$ interpolates $f(x,y)$ at the point (i,j). We now use induction. Let us assume that for some $k \geq 0$ and all i and j, the polynomial $p_k^{[i,j]}(x,y)$ interpolates $f(x,y)$ on the set $S_k^{[i,j]}$. Then we see that if all three polynomials $p_k^{[i,j]}(x,y)$, $p_k^{[i+1,j]}(x,y)$,

and $p_k^{[i,j+1]}(x,y)$ on the right of (5.45) have the same value C for some choice of x and y, then the right side of (5.45) has the value

$$\frac{C}{k+1}\Big[(i+j+k+1-x-y)+(x-i)+(y-j)\Big]=C,$$

and thus $p_{k+1}^{[i,j]}(x,y)=C$. Next we observe that these three polynomials *all* interpolate $f(x,y)$ on all points $(i+r,j+s)$ for which $r>0$, $s>0$, and $r+s<k+1$, and so $p_{k+1}^{[i,j]}(x,y)$ also interpolates $f(x,y)$ on all these points. We further show from (5.45) that $p_{k+1}^{[i,j]}(x,y)$ interpolates $f(x,y)$ also on the three "lines" of points, these being subsets of the set $S_{k+1}^{[i,j]}$ corresponding to taking $r=0$, $s=0$, and $r+s=k+1$ in turn. This completes the proof by induction. ■

Further references on bivariate interpolation may be found in Stancu [52].

Problem 5.2.1 Show that

$$\mathbf{A}=\begin{bmatrix}1&x_0&y_0\\1&x_1&y_0\\1&x_1&y_1\end{bmatrix}=\begin{bmatrix}1&0&x_0\\1&0&x_1\\0&1&x_1\end{bmatrix}\begin{bmatrix}1&0&y_0\\1&0&y_1\\0&1&0\end{bmatrix}$$

and hence verify that $\det\mathbf{A}=(x_1-x_0)(y_1-y_0)$.

Problem 5.2.2 Let $\mathbf{A}$ denote the 10×10 matrix concerned with the construction of the interpolating polynomial in x and y of total degree 3 on the 10 abscissas defined by

$$(x_i,y_j),\ \ i,j\geq0,\ i+j\leq3,$$

where the x_i are distinct and the y_j are distinct, and the points (x_i,y_j) are taken in the order specified in the proof of Theorem 5.2.1. By following the argument that we used in Example 5.2.1, show that

$$\det\mathbf{A}=C\,\psi(x_0,x_1,x_2,x_3)\,\psi(y_0,y_1,y_2,y_3),$$

where C is a nonzero constant and

$$\psi(x_0,x_1,x_2,x_3)=\prod_{i>j}(x_i-x_j)^{4-i}.$$

Deduce that this interpolation problem has a unique solution.

Problem 5.2.3 Generalize the result in Problem 5.2.2 from 3 to n, as follows. Let $\mathbf{A}_n$ denote the $\frac{1}{2}(n+1)(n+2)\times\frac{1}{2}(n+1)(n+2)$ matrix defined in the proof of Theorem 5.2.1, and let

$$\psi(x_0,\ldots,x_n)=\prod_{i>j}(x_i-x_j)^{n+1-i}.$$

Verify that $\psi(x_0, \ldots, x_n)$ is a polynomial of degree $\frac{1}{6}n(n+1)(n+2)$, and that $\det \mathbf{A}_n$ is a polynomial of degree $\frac{1}{3}n(n+1)(n+2)$, and deduce that

$$\det \mathbf{A}_n = C\,\psi(x_0, \ldots, x_n)\,\psi(y_0, \ldots, y_n),$$

where C is a nonzero constant.

Problem 5.2.4 Verify (5.29), and show also that

$$f(x,y) = [x_0]\,[y_0]\,f + (x - x_0)\,[x, x_0]\,[y]\,f + (y - y_0)\,[x_0]\,[y, y_0]\,f.$$

Problem 5.2.5 Extend the result in the last problem from two to three variables, showing that

$$\begin{aligned} f(x,y,z) = {} & [x_0]\,[y_0]\,[z_0]\,f + (x - x_0)\,[x, x_0]\,[y]\,[z]\,f \\ & + (y - y_0)\,[x_0]\,[y, y_0]\,[z]\,f + (z - z_0)\,[x_0]\,[y_0]\,[z, z_0]\,f. \end{aligned}$$

Problem 5.2.6 Let S denote the set of points in $\mathbb{R}^3$ defined by

$$S = \{(x_i, y_j, z_k) \mid i, j, k \geq 0,\ i + j + k \leq n\},$$

where the x_i are distinct, the y_j are distinct, and the z_j are distinct, and let $p_n(x,y,z)$ denote a polynomial that interpolates a given function f on the set S. Using the result in the last problem, determine $p_n(x,y,z)$ when $n = 0$ and $n = 1$. Also determine $p_n(x,y,z)$ when $n = 2$.

5.3 Integration on the Triangle

We now discuss integration rules over the triangle. If we integrate the interpolating polynomial $p_n(x,y)$, defined by (5.43), over the triangle T_n with vertices $(0,0)$, $(n,0)$, and $(0,n)$, we obtain an integration rule

$$R_n(f) = \sum_{i,j} w_{i,j}^{(n)}\, f(i,j), \tag{5.46}$$

where the summation is over all nonnegative integers i and j such that $i + j \leq n$. Thus the weight $w_{i,j}^{(n)}$ is obtained by integrating the fundamental polynomial $L_{i,j}(x,y)$, defined in (5.42), over the triangle T_n, and we see that

$$w_{i,j}^{(n)} = \int_0^n \left(\int_0^{n-y} L_{i,j}(x,y)\,dx \right) dy. \tag{5.47}$$

We say that R_n is an interpolatory integration rule on the triangle T_n. From the uniqueness of the interpolating polynomial it follows that when $f(x,y)$ is a polynomial of total degree at most n in x and y,

$$R_n(f) = \int_0^n \left(\int_0^{n-y} f(x,y)\,dx \right) dy.$$

We say that the rule is *exact* for such functions. In particular, we have

$$R_n(x^r y^s) = \int_0^n \left(\int_0^{n-y} x^r y^s dx \right) dy \tag{5.48}$$

for $r, s \geq 0$ and $r + s \leq n$. On evaluating the inner integral in (5.48), we obtain

$$R_n(x^r y^s) = \frac{1}{r+1} \int_0^n (n-y)^{r+1} y^s dy,$$

and on making the substitution $y = nt$, we find that

$$R_n(x^r y^s) = \frac{n^{r+s+2}}{r+1} \int_0^1 (1-t)^{r+1} t^s dt.$$

It is then not difficult to show, using integration by parts (see Problem 5.3.1), that

$$R_n(x^r y^s) = n^{r+s+2} \frac{r!\, s!}{(r+s+2)!}. \tag{5.49}$$

If we now replace $f(x, y)$ in (5.46) by $x^r y^s$, and use (5.49), for $r, s \geq 0$ and $r + s \leq n$, we obtain a system of linear equations to determine the weights $w_{i,j}^{(n)}$. This is preferable to evaluating the weights directly by integrating the fundamental polynomials, using (5.47). There are some symmetries in the weights that we can use to simplify the solution of the linear equations. For we have

$$w_{i,j}^{(n)} = w_{j,i}^{(n)} = w_{i,n-i-j}^{(n)} = w_{n-i-j,i}^{(n)} = w_{j,n-i-j}^{(n)} = w_{n-i-j,j}^{(n)}. \tag{5.50}$$

Let us consider $w_{i,j}^{(n)}$ and $w_{j,i}^{(n)}$. If we interchange i and j in (5.47), then from (5.42) this is equivalent to interchanging x and y, which leaves the integral unchanged, since the domain of integration is symmetric in x and y. This establishes the relation $w_{i,j}^{(n)} = w_{j,i}^{(n)}$. It now remains only to show that, say,

$$w_{i,j}^{(n)} = w_{i,n-i-j}^{(n)},$$

and the whole chain of equalities in (5.50) will follow. From (5.47) and (5.42) we can express $w_{i,n-i-j}^{(n)}$ as the double integral

$$w_{i,n-i-j}^{(n)} = \int_0^n \int_0^{n-y} \binom{x}{i} \binom{y}{n-i-j} \binom{n-x-y}{j} dx\, dy.$$

Let us now make the change of variables

$$\begin{aligned} x &= \xi, \\ y &= n - \xi - \eta, \end{aligned}$$

and we note that the domain of integration of the latter double integral, the triangular region $x, y \geq 0$, $x+y \leq n$, is transformed into the triangular region $\xi, \eta \geq 0$, $\xi + \eta \leq n$. Thus we obtain

$$w_{i,n-i-j}^{(n)} = \int_0^n \left(\int_0^{n-\eta} \begin{pmatrix} \xi \\ i \end{pmatrix} \begin{pmatrix} n-\xi-\eta \\ n-i-j \end{pmatrix} \begin{pmatrix} \eta \\ j \end{pmatrix} |J| \, d\xi \right) d\eta, \quad (5.51)$$

where $|J|$ denotes the modulus of the Jacobian,

$$J = \det \begin{bmatrix} \frac{\partial x}{\partial \xi} & \frac{\partial x}{\partial \eta} \\ \frac{\partial y}{\partial \xi} & \frac{\partial y}{\partial \eta} \end{bmatrix} = \det \begin{bmatrix} 1 & 0 \\ -1 & -1 \end{bmatrix} = -1.$$

On replacing $|J|$ by unity in (5.51), we see that $w_{i,n-i-j}^{(n)} = w_{i,j}^{(n)}$.

Example 5.3.1 With $n = 1$, we see from (5.49) with $r = s = 0$ and (5.46) with $f(x, y) = 1$ that

$$w_{0,0}^{(1)} + w_{1,0}^{(1)} + w_{0,1}^{(1)} = \frac{1}{2}.$$

It follows from (5.50) that all three weights are equal, and thus

$$w_{0,0}^{(1)} = w_{1,0}^{(1)} = w_{0,1}^{(1)} = \frac{1}{6}.$$

With $n = 2$, we note from (5.50) that there are at most two distinct weights, since

$$w_{0,0}^{(2)} = w_{2,0}^{(2)} = w_{0,2}^{(2)}$$

and

$$w_{1,0}^{(2)} = w_{0,1}^{(2)} = w_{1,1}^{(2)}.$$

Then, using (5.46) with $f(x, y) = 1$ and $f(x, y) = xy$, say, together with (5.49), we obtain the equations

$$3w_{0,0}^{(2)} + 3w_{1,1}^{(2)} = 2,$$
$$w_{1,1}^{(2)} = \frac{2}{3},$$

and hence find that the weights are

$$w_{0,0}^{(2)} = w_{2,0}^{(2)} = w_{0,2}^{(2)} = 0$$

and

$$w_{1,0}^{(2)} = w_{0,1}^{(2)} = w_{1,1}^{(2)} = \frac{2}{3}.$$

For $n = 3$ we see from (5.50) that there are at most three distinct weights. We have the weights at the vertices of the triangle T_3,

$$w_{0,0}^{(3)} = w_{3,0}^{(3)} = w_{0,3}^{(3)};$$

the weight $w_{1,1}^{(3)}$ at $(1,1)$, the centroid of the triangle T_3; and the remaining weights

$$w_{1,0}^{(3)} = w_{0,1}^{(3)} = w_{1,2}^{(3)} = w_{2,1}^{(3)} = w_{0,2}^{(3)} = w_{2,0}^{(3)}.$$

Using (5.46) with $f(x,y) = 1$, x^2, and x^3, say, together with (5.49), we obtain the equations

$$\begin{aligned} 3w_{0,0}^{(3)} + w_{1,1}^{(3)} + 6w_{1,0}^{(3)} &= \frac{9}{2}, \\ 9w_{0,0}^{(3)} + w_{1,1}^{(3)} + 10w_{1,0}^{(3)} &= \frac{27}{4}, \\ 27w_{0,0}^{(3)} + w_{1,1}^{(3)} + 18w_{1,0}^{(3)} &= \frac{243}{20}, \end{aligned}$$

whose solution is

$$w_{0,0}^{(3)} = \frac{3}{20}, \quad w_{1,1}^{(3)} = \frac{81}{40}, \quad w_{1,0}^{(3)} = \frac{27}{80}. \qquad \blacksquare$$

It is instructive to generalize our above account of interpolatory integration rules on the triangle. Instead of the triangle T_n, with vertices at $(0,0)$, $(n,0)$, and $(0,n)$, let us consider any triangle T, with vertices at (x_1,y_1), (x_2,y_2), and (x_3,y_3), and area $\Delta > 0$. We now write

$$x = x_1u_1 + x_2u_2 + x_3u_3, \tag{5.52}$$

$$y = y_1u_1 + y_2u_2 + y_3u_3, \tag{5.53}$$

where

$$u_1 + u_2 + u_3 = 1. \tag{5.54}$$

We call u_1, u_2, and u_3 the *barycentric coordinates* of the point (x,y). We can express (5.52), (5.53), and (5.54) in the matrix form

$$\begin{bmatrix} x \\ y \\ 1 \end{bmatrix} = \mathbf{A} \begin{bmatrix} u_1 \\ u_2 \\ u_3 \end{bmatrix}, \tag{5.55}$$

where

$$\mathbf{A} = \begin{bmatrix} x_1 & x_2 & x_3 \\ y_1 & y_2 & y_3 \\ 1 & 1 & 1 \end{bmatrix}. \tag{5.56}$$

Thus

$$\mathbf{A}^{-1} = \frac{1}{\det \mathbf{A}} \begin{bmatrix} \eta_1 & -\xi_1 & \tau_1 \\ \eta_2 & -\xi_2 & \tau_2 \\ \eta_3 & -\xi_3 & \tau_3 \end{bmatrix}, \tag{5.57}$$

where

$$\xi_1 = x_2 - x_3, \quad \eta_1 = y_2 - y_3, \quad \tau_1 = x_2y_3 - x_3y_2,$$

and ξ_i, η_i, and τ_i are defined cyclically for $i = 2$ and 3. It follows that

$$u_i = u_i(x, y) = \frac{\eta_i x - \xi_i y + \tau_i}{\eta_i x_i - \xi_i y_i + \tau_i}, \quad 1 \le i \le 3. \tag{5.58}$$

It is easily verified that the denominator in (5.58) is independent of i and is nonzero, since

$$\eta_i x_i - \xi_i y_i + \tau_i = \det \mathbf{A} = \pm 2\Delta, \tag{5.59}$$

for $i = 1, 2$, and 3. Each of the linear functions $u_i(x, y)$ defined in (5.58) has the value 1 at the vertex (x_i, y_i) and is zero at the other two vertices.

In place of S^n, defined by (5.35), we have the set of interpolation points

$$S_T = \left\{ \left(\sum_{i=1}^{3} \frac{\lambda_i x_i}{n}, \sum_{i=1}^{3} \frac{\lambda_i y_i}{n} \right) \;\middle|\; \lambda_1, \lambda_2, \lambda_3 \ge 0, \sum_{i=1}^{3} \lambda_i = n \right\}, \tag{5.60}$$

where the λ_i are integers. Observe that S_T reduces to S^n when we replace the vertices of T with the vertices of T_n. Let us now write f_λ to denote the value of $f(x, y)$ at the point

$$(x, y) = \left(\sum_{i=1}^{3} x_i u_i, \sum_{i=1}^{3} y_i u_i \right) = \left(\sum_{i=1}^{3} \frac{\lambda_i x_i}{n}, \sum_{i=1}^{3} \frac{\lambda_i y_i}{n} \right), \tag{5.61}$$

where λ_1, λ_2, and λ_3 are nonnegative integers such that $\lambda_1 + \lambda_2 + \lambda_3 = n$, and let us define

$$L_\lambda(x, y) = \prod_{i=1}^{3} \frac{1}{\lambda_i!} \prod_{j=0}^{\lambda_i - 1} (n u_i - j),$$

where, as usual, an empty product is taken to have the value 1. We can see that $L_\lambda(x, y)$ takes the value 1 at the point (x, y) defined by (5.61) and is zero at all the other points in the set S_T, defined by (5.60). Thus $L_\lambda(x, y)$ is a fundamental polynomial for the set S_T. Let us now define the polynomial

$$p_n(x, y) = \sum_{\lambda} f_\lambda L_\lambda(x, y),$$

where the summation is over all the points, enumerated by λ_1, λ_2, and λ_3, in the set S_T. It is clear that $p_n(x, y)$ interpolates $f(x, y)$ on the set S_T. We now integrate this interpolating polynomial over the triangle T, giving

$$\iint_T p_n(x, y)\, dx\, dy = \sum_{\lambda} w_\lambda f_\lambda = R_T(f), \tag{5.62}$$

say, where the weight w_λ is given by

$$w_\lambda = \iint_T L_\lambda(x, y)\, dx\, dy. \tag{5.63}$$

We see from (5.52), (5.53), and (5.54) that the Jacobian of this transformation is

$$J = \det \begin{bmatrix} \frac{\partial x}{\partial u_1} & \frac{\partial x}{\partial u_2} \\ \\ \frac{\partial y}{\partial u_1} & \frac{\partial y}{\partial u_2} \end{bmatrix} = \det \begin{bmatrix} x_1 - x_3 & x_2 - x_3 \\ \\ y_1 - y_3 & y_2 - y_3 \end{bmatrix} = \det \mathbf{A}. \tag{5.64}$$

We also note that under this transformation, the set of all points (x, y) in the triangle T corresponds to the set of points (u_1, u_2) in the triangle T', say, defined by

$$T' = \{(u_1, u_2) \mid u_1, u_2 \geq 0,\ u_1 + u_2 \leq 1\}.$$

It follows from (5.52), (5.53), and (5.54) that x and y are linear in u_1 and u_2, and so any polynomial in x and y of total degree at most n may be expressed as a polynomial in u_1 and u_2 of total degree at most n. Let us evaluate integrals of the monomials in the variables u_1 and u_2. We have

$$\iint_T u_1^r u_2^s \, dx \, dy = \iint_{T'} u_1^r u_2^s \, |J| \, du_1 \, du_2,$$

where the Jacobian J is given by (5.64), so that

$$\iint_T u_1^r u_2^s \, dx \, dy = 2\Delta \int_0^1 \left(\int_0^{1-u_2} u_1^r u_2^s \, du_1 \right) du_2.$$

Thus, following the method used in evaluating $R_n(x^r y^s)$ in (5.49), we obtain

$$\iint_T u_1^r u_2^s \, dx \, dy = 2\Delta \, \frac{r! \, s!}{(r+s+2)!}. \tag{5.65}$$

To determine the weights w_λ for the interpolatory integration rule over the triangle T, we follow the method used earlier to find the weights for the rule over the triangle T_n. We therefore set up and solve a system of linear equations, where each equation is of the form

$$\sum_\lambda w_\lambda f_\lambda = R_T(f), \tag{5.66}$$

and f is chosen as each of the monomials $u_1^r u_2^s$ in turn, so that

$$R_T(u_1^r u_2^s) = 2\Delta \, \frac{r! \, s!}{(r+s+2)!}. \tag{5.67}$$

It follows from (5.66) and (5.67) that, apart from a multiplicative constant that depends on the area of the triangle, the weights w_λ for these interpolatory rules on S_T, with n fixed, are otherwise independent of the triangle. Let us suppose we have a set of numbers w'_λ that have been obtained by

$n=1$	$n=2$	$n=3$
1	0	4
1 1	1 1	9 9
	0 1 0	9 54 9
		4 9 9 4

$n=4$	$n=5$
0	11
4 4	25 25
−1 8 −1	25 200 25
4 8 8 4	25 25 25 25
0 4 −1 4 0	25 200 25 200 25
	11 25 25 25 25 11

TABLE 5.2. Relative weights for interpolatory integration rules of order n on the triangle, for $1 \leq n \leq 5$.

multiplying all the weights w_λ by some positive constant. We call the numbers w'_λ a set of *relative weights*. An interpolatory integration rule of order n interpolates all polynomials of total degree n or less exactly, and so, in particular, integrates the constant function 1 exactly. Thus we see from (5.66) and (5.67) that

$$\sum_\lambda w_\lambda = \Delta,$$

and it follows from this last equation that if w'_λ are a set of relative weights, the true weights are given by

$$w_\lambda = \frac{\Delta}{\sum_\lambda w'_\lambda} \cdot w'_\lambda,$$

where the summation is over all the points, enumerated by λ_1, λ_2, and λ_3, in the set S_T. It is clear that the weights w_λ are rational numbers, since they are derived from a system of linear equations with rational coefficients. It is convenient to produce relative weights from these, consisting of a set of integers with no common factor. Relative weights for interpolatory rules on all triangular sets defined by (5.60), for $1 \leq n \leq 5$, are given in Table 5.2. The weights for the cases $n = 1$, 2, and 3 were derived in Example 5.3.1, and those for $n = 4$ and 5 may be obtained similarly.

Many other integration rules for the triangle and the general simplex are given in Stroud [53].

Problem 5.3.1 Let

$$I_{r,s} = \int_0^1 (1-t)^r t^s dt.$$

Use integration by parts to show that

$$I_{r,s} = \frac{r}{s+1} I_{r-1,s+1},$$

and deduce that

$$I_{r,s} = \frac{r!\, s!}{(r+s+1)!}.$$

5.4 Interpolation on the q-Integers

Consider the following set of points, illustrated in Figure 5.4 for the case where $n = 4$, defined in terms of q-integers (see Section 1.5) by

$$S_q^n = \{([i], [j]'),\ |\ i, j \geq 0,\ i + j \leq n\}, \tag{5.68}$$

where, with $q > 0$,

$$[i] = 1 + q + q^2 + \cdots + q^{i-1} \quad \text{and} \quad [j]' = 1 + q^{-1} + q^{-2} + \cdots + q^{-j+1},$$

for $i, j > 0$, and where $[0] = [0]' = 0$. When $q = 1$ we have $[i] = i$ and $[j]' = j$, and the grid S_q^n reduces to the simple triangular grid S^n defined in (5.35). The grid S_q^n has the property, shared with the grid S^n, that it is created by points of intersection of three systems of straight lines. As we saw in Figure 5.2, the set of points in S^n consists of three systems of parallel lines, one parallel to each axis and the third parallel to $x+y-n = 0$. The new set of points S_q^n is created by points of intersection of the three systems

$$\begin{aligned} x - [k] &= 0, \quad 0 \leq k \leq n-1, \\ y - [k]' &= 0, \quad 0 \leq k \leq n-1, \\ x + q^k y - [k+1] &= 0, \quad 0 \leq k \leq n-1. \end{aligned} \tag{5.69}$$

Note that the straight line with equation $x + q^k y - [k+1] = 0$ connects the two points $([k+1], 0)$ and $(0, [k+1]'$, and contains all points of the form $([i], [k+1-i]'),\ 0 \leq i \leq k+1$.

We define a *pencil of lines* as a set of lines that are all parallel, or all pass through a common point, which is called the *vertex* of the pencil. In the case where the lines are all parallel, we can think of the vertex as being at infinity, in the direction of the set of parallel lines. Thus, in (5.69) we have three pencils of lines. The first two pencils are systems of lines parallel to the axes. The third system is obviously *not* a parallel system except when $q = 1$. On substituting the values $x = 1/(1-q)$ and $y = -q/(1-q)$ into (5.69), with $q \neq 1$, we can see that every line in the third system passes through the vertex $(1/(1-q), -q/(1-q))$. Thus the x-coordinate of this vertex is negative for $q > 1$, as in Figure 5.4. We can say that this grid is

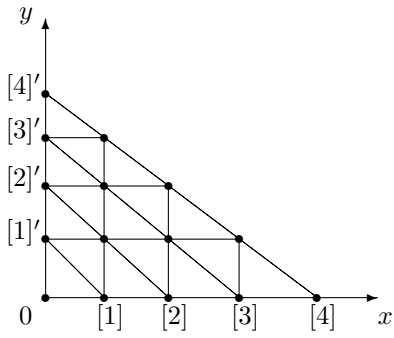

FIGURE 5.4. A triangular interpolation grid based on q-integers.

created by two pencils of lines with vertices at infinity, and a third pencil of lines that meet at a finite vertex. We can now write down the Lagrange form of an interpolating polynomial for a function $f(x, y)$ on this triangular grid, as we did for the special case of $q = 1$. The fundamental polynomial for the point $([i], [j]')$ in this new grid is given by

$$L_{i,j}(x, y) = a_{i,j}(x, y)\, b_{i,j}(x, y)\, c_{i,j}(x, y), \tag{5.70}$$

where

$$a_{i,j}(x, y) = \prod_{s=0}^{i-1} \left(\frac{x - [s]}{[i] - [s]} \right), \quad b_{i,j}(x, y) = \prod_{s=0}^{j-1} \left(\frac{y - [s]'}{[j]' - [s]'} \right),$$

$$c_{i,j}(x, y) = \prod_{s=i+j+1}^{n} \left(\frac{x + q^{s-1} y - [s]}{[i] + q^{s-1}[j]' - [s]} \right),$$

and, as usual, an empty product denotes 1. With $q = 1$, this reduces to the expression (5.42) for the fundamental polynomial corresponding to the point (i, j).

The mesh S_q^n and its special case S^n illustrate the following result due to Chung and Yao [8].

Theorem 5.4.1 Let S denote a set of points in Euclidean space $\mathbb{R}^2$, and suppose that to each point $p_i \in S$, there corresponds a set of n lines $l_{i,1}$, $l_{i,2}, \ldots, l_{i,n}$, such that $p \in S$ lies in the union of $l_{i,1}$, $l_{i,2}, \ldots, l_{i,n}$ if and only if $p \neq p_i$. Then there is a unique polynomial of the form

$$p_n(x, y) = \sum_{k=0}^{n} \sum_{r=0}^{k} c_{r,k-r} x^r y^{k-r}$$

that interpolates a given function $f(x, y)$ on the set S. ■

Later in this section we will give other meshes that fulfil the conditions of this theorem. We continue by giving a Neville–Aitken algorithm for evaluating the interpolating polynomial for $f(x,y)$ on the set of points S_q^n defined in terms of q-integers in (5.68). This generalizes the algorithm given in Theorem 5.2.3 for computing the interpolating polynomial for $f(x,y)$ on the set S^n, which is the special case of S_q^n with $q=1$. Let us define $p_k^{[i,j]}(x,y)$ as the interpolating polynomial for $f(x,y)$ on the triangular set of points

$$S_k^{[i,j]} = \{([i+r],[j+s]') \mid r,s \geq 0,\ r+s \leq k\}\,. \tag{5.71}$$

These interpolating polynomials can be computed recursively, as stated in the following theorem.

Theorem 5.4.2 For $k \geq 0$ and $i,j \geq 0$,

$$\begin{aligned} p_{k+1}^{[i,j]}(x,y) = {} & \frac{([1+i+j+k]-x-q^{i+j+k}y)}{q^i[k+1]}\, p_k^{[i,j]}(x,y) \\ & + \frac{(x-[i])}{q^i[k+1]}\, p_k^{[i+1,j]}(x,y) + q^{j+k}\frac{(y-[j]')}{[k+1]}\, p_k^{[i,j+1]}(x,y). \end{aligned}$$

Proof. The recurrence relation reduces to (5.45) when we put $q=1$, and is justified in the same way as the special case in Theorem 5.2.3. ■

Recall the grid of points

$$S_\Delta^n = \{(x_i,y_j) \mid i,j \geq 0,\ i+j \leq n\},$$

defined in (5.21), where the x_i are distinct, and the y_j are distinct. We showed in Theorem 5.2.1 that there is a unique polynomial of total degree n in x and y that interpolates a given function $f(x,y)$ on the set S_Δ^n, and in (5.33) showed that this polynomial can be expressed in the divided difference form

$$p_n(x,y) = \sum_{m=0}^{n}\sum_{k=0}^{m} \pi_k(x)\pi_{m-k}(y)[x_0,\ldots,x_k][y_0,\ldots,y_{m-k}]f, \tag{5.72}$$

where the polynomials π_k are defined in (1.11). If in (5.72) we now let

$$x_i = [i] = \frac{1-q^i}{1-q} \quad \text{and} \quad y_j = [j]' = \frac{1-q^{-j}}{1-q^{-1}},$$

for some choice of $q>0$, and write $x=[u]$ and $y=[v]'$, we can write the polynomial $p_n(x,y)$ in (5.72) in the form

$$p_n([u],[v]') = \sum_{m=0}^{n}\sum_{k=0}^{m} \begin{bmatrix} u \\ k \end{bmatrix} \begin{bmatrix} v \\ m-k \end{bmatrix}' \Delta_x^k \Delta_y^{m-k} f(0,0), \tag{5.73}$$

where Δ_x and Δ_y are q-difference operators with respect to q and q^{-1}, respectively, and the two factors that multiply the differences are q-binomial coefficients, involving q and q^{-1}, respectively, as defined in (1.116). We can justify (5.73) in the same way as we verified (5.36), the special case of (5.73) when $q = 1$.

Note that the points of the grid S_q^n, defined by (5.68), all lie on the triangle whose vertices are the three grid points $(0,0)$, $([n],0)$, and $(0,[n]')$. Let us scale the grid S_q^n, dividing the x-coordinates by $[n]$ and the y-coordinates by $[n]'$, to give a grid of points lying in the "unit" triangle, with vertices $(0,0)$, $(1,0)$, and $(0,1)$. Since

$$\frac{[j]'}{[n]'} = \frac{1-q^{-j}}{1-q^{-n}} = 1 - \frac{[n-j]}{[n]},$$

the scaled grid is the set of points

$$\left(\frac{[i]}{[n]}, 1 - \frac{[n-j]}{[n]}\right), \quad i,j \geq 0,\ i+j \leq n,$$

which is more conveniently described as the set of points

$$\left(\frac{[i]}{[n]}, 1 - \frac{[j]}{[n]}\right), \quad 0 \leq i \leq j \leq n. \tag{5.74}$$

We saw that S_q^n has a finite vertex $(1/(1-q), -q/(1-q))$, the point where the n lines defined by (5.69) intersect. After scaling, this point becomes $(1/(1-q^n), -q^n/(1-q^n))$, the finite vertex for the mesh defined by (5.74). As $q \to 1$, this vertex tends to infinity along the line $x + y = 1$, and its pencil is a system of lines parallel to $x + y = 1$. This limiting form of the grid defined by (5.74) is just a scaled version of S^n, defined in (5.35). The grid S_q^n or, equivalently, its scaled version given by (5.74), belongs to a family of grids based on q-integers derived by Lee and Phillips [32]. This family also contains grids created by one pencil of parallel lines and two pencils with finite vertices, and grids created by three pencils each of which has a finite vertex. All of these may be derived by a geometrical construction that relies on Pappus's theorem, which we state and prove below. We begin by defining *homogeneous coordinates*, in which we use a triple of numbers (x,y,z), not all zero, to denote a point. If $\lambda \neq 0$, then (x,y,z) and $(\lambda x, \lambda y, \lambda z)$ denote the same point, and if $z \neq 0$, the point (x,y,z) coincides with the point $(x/z, y/z)$ in $\mathbb{R}^2$, two-dimensional Euclidean space. The straight line denoted by $ax + by + c = 0$ in $\mathbb{R}^2$ is written in the form $ax + by + cz = 0$ in homogeneous coordinates. The three rather special points $(1,0,0)$, $(0,1,0)$, and $(0,0,1)$ are denoted by X, Y, and Z, respectively. We refer to XYZ as the triangle of reference. We can see that the line YZ is given by

$$\det \begin{bmatrix} x & y & z \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix} = 0,$$

or $x = 0$, since the determinant is zero (two rows equal) when we replace the coordinates (x, y, z) by those of Y or Z. Similarly, ZX and XY have the equations $y = 0$ and $z = 0$, respectively. We note also that any straight line through X has an equation of the form $by + cz = 0$, and any point on YZ can be expressed in the form $(0, y_1, z_1)$.

There is another "special" point, $(1, 1, 1)$, called the unit point. Its usefulness lies in the simplifications that we gain from the following property: We can find a coordinate system in which *any* point not on a side of the triangle of reference XYZ has coordinates $(1, 1, 1)$. To justify this, suppose that in a given coordinate system, a point U has coordinates (α, β, γ), and that U does not lie on a side of triangle XYZ. This implies that the three coordinates α, β, and γ are all nonzero. We now carry out a transformation that maps (x, y, z) to (x', y', z'), where

$$x' = x/\alpha, \quad y' = y/\beta, \quad z' = z/\gamma.$$

Thus, in the new coordinate system, U has the coordinates of the unit point $(1, 1, 1)$, and because of the property that (x, y, z) and $(\lambda x, \lambda y, \lambda z)$ denote the same point when λ is nonzero, the triangle of reference XYZ is unchanged by this simple transformation. Under the more general transformation

$$\begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{5.75}$$

the special points X, Y, and Z, represented by column vectors, are mapped to the points, say, X', Y', and Z', represented by the columns of the matrix $\mathbf{A}$. The points X', Y', and Z' will be collinear if and only if the columns of $\mathbf{A}$ are linearly dependent, that is, if and only if the matrix $\mathbf{A}$ is singular. Likewise, beginning with any three noncollinear points X', Y', and Z', we can construct a (nonsingular) matrix $\mathbf{A}$ whose columns are obtained from their coordinates. Then the matrix $\mathbf{A}^{-1}$ will map X', Y', and Z' to the special points X, Y, and Z, respectively. It is easy to see that under such nonsingular transformations, straight lines are mapped to straight lines. We will now state and prove Pappus's theorem.

Theorem 5.4.3 Let A_1, A_2, A_3, B_1, B_2, and B_3 be six distinct points lying in a plane, where A_1, A_2, and A_3 lie on a straight line l_A, and B_1, B_2, and B_3 lie on a straight line l_B. We now construct the points C_1, C_2, and C_3, as follows:

C_1 is the point where A_2B_3 and A_3B_2 intersect;
C_2 is the point where A_3B_1 and A_1B_3 intersect;
C_3 is the point where A_1B_2 and A_2B_1 intersect.

Then the points C_1, C_2, and C_3 lie on a straight line l_C. See Figure 5.5.

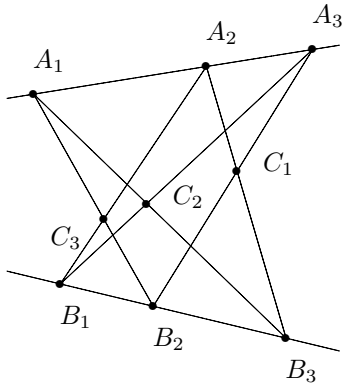

FIGURE 5.5. Pappus's theorem: The points C_1, C_2, and C_3 are collinear.

Proof. Before commencing the proof note that, having defined the point C_1 above, we obtain the definitions of C_2 and C_3 by permuting the subscripts 1, 2, and 3 cyclically. We remark also that this geometrical construction is fully symmetric in the A_j, B_j, and C_j. For not only do the A_j and B_j determine the C_j, but the B_j and C_j determine the A_j, and the C_j and A_j determine the B_j. For example, having constructed the points C_1, C_2, and C_3, suppose we were to remove the points A_1, A_2, and A_3 from our diagram. Then we could restore them by defining A_1 as the point where B_2C_3 and B_3C_2 intersect, and defining A_2 and A_3 similarly. The following proof is based on that of Maxwell [37]. Let l_A and l_B intersect at P. We need to allow the possibility that l_A and l_B are parallel, because Pappus's theorem holds whether they are parallel or not. If l_A and l_B are parallel, we can represent them by the equations

$$ax + by + cz = 0,$$
$$ax + by + c'z = 0,$$

where $c \neq c'$. In this case we can take the point P as $(-b, a, 0)$, since the coordinates of P satisfy the equations of both lines. Let Q denote any other point on l_B, and let R denote any other point on l_A. We then apply the unique linear transformation that maps P, Q, and R onto X, Y, and Z, respectively. In this coordinate system, l_A becomes XZ. Thus any point on the line l_A may be expressed in the form $(\alpha, 0, 1)$, and we recover the point Z on putting $\alpha = 0$. If we multiply by $1/\alpha$, the point with homogeneous coordinates $(\alpha, 0, 1)$ is the same as $(1, 0, 1/\alpha)$, and we recover X on letting $\alpha \to \infty$. Any point on the line l_B may be expressed in the form $(\beta, 1, 0)$, and we recover Y on putting $\beta = 0$, and X by dividing throughout by β and then letting $\beta \to \infty$. We may therefore write

$$A_i = (\alpha_i, 0, 1), \quad B_i = (\beta_i, 1, 0), \quad i = 1, 2, 3.$$

We can now determine the coordinates of the C_i. The line A_2B_3 has the equation $x - \beta_3 y - \alpha_2 z = 0$, and on interchanging the subscripts 2 and 3, we see that the equation of the line A_3B_2 is $x - \beta_2 y - \alpha_3 z = 0$. We then find the point of intersection of these two lines,

$$C_1 = (\alpha_2\beta_2 - \alpha_3\beta_3, \alpha_2 - \alpha_3, \beta_2 - \beta_3).$$

By permuting the subscripts 1, 2, and 3 in cyclic order, we see that the points C_2 and C_3 are

$$C_2 = (\alpha_3\beta_3 - \alpha_1\beta_1, \alpha_3 - \alpha_1, \beta_3 - \beta_1),$$
$$C_3 = (\alpha_1\beta_1 - \alpha_2\beta_2, \alpha_1 - \alpha_2, \beta_1 - \beta_2).$$

Now consider any three points x_i, y_i, z_i, $i = 1, 2, 3$. These will lie on a straight line $ax + by + cz = 0$ if and only if there exist a, b, and c, not all zero, such that

$$\begin{bmatrix} x_1 & y_1 & z_1 \\ x_2 & y_2 & z_2 \\ x_3 & y_3 & z_3 \end{bmatrix} \begin{bmatrix} a \\ b \\ c \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix},$$

and there exist such numbers a, b, and c if and only if the above matrix is singular. The corresponding matrix formed from the coordinates of C_1, C_2, and C_3 is

$$\begin{bmatrix} \alpha_2\beta_2 - \alpha_3\beta_3 & \alpha_2 - \alpha_3 & \beta_2 - \beta_3 \\ \alpha_3\beta_3 - \alpha_1\beta_1 & \alpha_3 - \alpha_1 & \beta_3 - \beta_1 \\ \alpha_1\beta_1 - \alpha_2\beta_2 & \alpha_1 - \alpha_2 & \beta_1 - \beta_2 \end{bmatrix}.$$

The rows of this matrix are evidently linearly dependent, since their sum is the zero row vector. Thus the matrix is singular, and the points C_1, C_2, and C_3 indeed lie on a straight line. This completes the proof. ■

We now derive the three-pencil mesh obtained by Lee and Phillips [32]. It is convenient to use X, Y, and Z to denote the vertices of these pencils. Each point P of the mesh lies on three lines, say, l_X^P, l_Y^P, and l_Z^P, where l_X^P is a member of a pencil of lines that has X as its vertex. Similarly, we write l_Y^P and l_Z^P to denote members of pencils of lines that have Y and Z, respectively, as their vertices. The mesh is generated by constructing one line in each of the three pencils at a time, as we now describe.

To construct the first line in each pencil, we draw arbitrary lines $l_X^{(1)}$ through X, and $l_Z^{(1)}$ through Z that are not sides of the triangle of reference. These intersect at an arbitrary point that is not on a side of triangle XYZ. Thus, without loss of generality, we can take this to be the unit point $U = (1, 1, 1)$, as we justified above. Since $l_X^{(1)}$ joins X and U, it has equation

$$\det \begin{bmatrix} x & y & z \\ 1 & 0 & 0 \\ 1 & 1 & 1 \end{bmatrix} = 0,$$

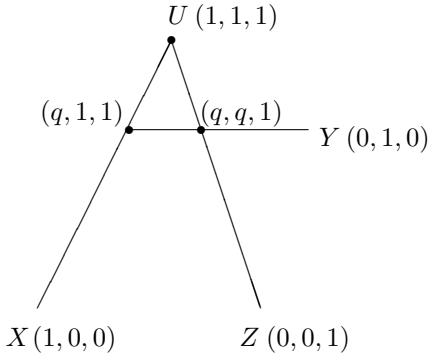

FIGURE 5.6. The first stage in constructing the three-pencil mesh.

which is $y = z$, and similarly, $l_Z^{(1)}$, which joins Z and U, has equation $x = y$. We now draw an arbitrary line $l_Y^{(1)}$ through Y that does not pass through X, Z, or U. This must be of the form $x = qz$, where $q \neq 0$ or 1. (If $q = 0$, the line would pass through Z, and if $q = 1$, it would pass through U.) The lines $l_X^{(1)}$ and $l_Y^{(1)}$ intersect where $y = z$ and $x = qz$, which gives the point $(q, 1, 1)$. Likewise, $l_Y^{(1)}$ and $l_Z^{(1)}$ intersect where $x = qz$ and $x = y$, which gives the point $(q, q, 1)$. At this stage we have the lines $l_X^{(1)}$, $l_Y^{(1)}$, and $l_Z^{(1)}$, and the first three points of the mesh, $(1, 1, 1)$, $(q, 1, 1)$, and $(q, q, 1)$, as depicted in Figure 5.6.

Next we construct the second lines of the X, Y, and Z pencils. We denote the line joining X and the mesh point $(q, q, 1)$ by $l_X^{(2)}$, with equation

$$\det \begin{bmatrix} x & y & z \\ 1 & 0 & 0 \\ q & q & 1 \end{bmatrix} = 0,$$

which is $y = qz$. The line $l_Z^{(2)}$ is defined as that joining Z and the mesh point $(q, 1, 1)$, which has equation $x = qy$. We continue by finding the point where $l_X^{(2)}$ and $l_Z^{(2)}$ intersect, which is where

$$y = qz \quad \text{and} \quad x = qy.$$

This gives our fourth mesh point, $(q^2, q, 1)$. We choose the line joining this latest point $(q^2, q, 1)$ to Y as $l_Y^{(2)}$, and find that it has equation $x = q^2 z$. We then find the fifth and sixth mesh points, where $l_Y^{(2)}$ intersects $l_X^{(1)}$ and $l_Z^{(1)}$. These are $(q^2, 1, 1)$ and $(q^2, q^2, 1)$, respectively. This completes the second stage of the construction of the three-pencil mesh, giving Figure 5.7.

It is worthwhile to pause and review the construction we have carried out so far, and to realize that it is more general than the emerging pattern

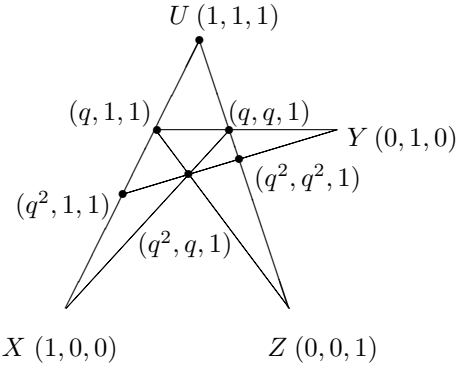

FIGURE 5.7. The second stage in constructing the three-pencil mesh.

in the coordinates might lead us to suppose. We began with essentially any lines $l_X^{(1)}$, $l_Y^{(1)}$, and $l_Z^{(1)}$, through X, Y, and Z, respectively. However, the second set of lines, $l_X^{(2)}$, $l_Y^{(2)}$, and $l_Z^{(2)}$, is then completely determined, as are all the remaining lines, whose construction we now describe.

We continue to criss-cross, as we did above in creating the fourth mesh point, with coordinates $(q^2,q,1)$. Let $l_X^{(3)}$ be the line joining X and the mesh point $(q^2,q^2,1)$, and let $l_Z^{(3)}$ be the line joining Z and the mesh point $(q^2,1,1)$. Further, let C_2 denote the point of intersection of $l_X^{(3)}$ and $l_Z^{(2)}$, and let C_3 denote the point of intersection of $l_Z^{(3)}$ and $l_X^{(2)}$. Let us, for the present, relabel Y as C_1. This is the moment of truth: We find that C_1, C_2, and C_3 are collinear! The reader will not be surprised that this is a consequence of Pappus's theorem. For if we choose

$$\begin{array}{lll} A_1=(1,0,0), & A_2=(q^2,1,1), & A_3=(q,1,1), \\ B_1=(0,0,1), & B_2=(q,q,1), & B_3=(q^2,q^2,1), \end{array}$$

and apply the Pappus construction, we obtain the collinear points $Y=C_1$, and C_2, C_3, as defined above. We find that $l_X^{(3)}$ has equation $y=q^2z$, and it intersects with $l_Z^{(2)}$, whose equation is $x=qy$, at the point $C_2=(q^3,q^2,1)$. Similarly, we find that $l_Z^{(3)}$ has equation $x=q^2y$ and that $C_3=(q^3,q,1)$. The algebraic power of the homogeneous coordinates makes it obvious that

$$C_1=Y=(0,1,0), \quad C_2=(q^3,q^2,1), \quad \text{and} \quad C_3=(q^3,q,1)$$

lie on the straight line with equation $x=q^3z$, which we denote by $l_Y^{(3)}$. We complete this third stage of our construction by finding also the points $(q^3,1,1)$ and $(q^3,q^3,1)$, where $l_Y^{(3)}$ intersects $l_X^{(1)}$ and $l_Z^{(1)}$, respectively. Thus, at this third stage, we have constructed three new lines, $l_X^{(3)}$, $l_Y^{(3)}$, and $l_Z^{(3)}$,

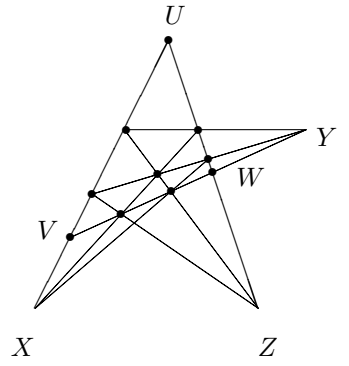

FIGURE 5.8. The third stage in constructing the three-pencil mesh.

and four new mesh points,

$$(q^3, 1, 1), \quad (q^3, q, 1), \quad (q^3, q^2, 1), \quad (q^3, q^3, 1).$$

After k stages, suppose we have the three pencils of lines

$$l_X^{(i)} \quad \text{with equation} \quad y = q^{i-1}z, \quad 1 \le i \le k,$$

$$l_Y^{(i)} \quad \text{with equation} \quad x = q^i z, \qquad 1 \le i \le k,$$

$$l_Z^{(i)} \quad \text{with equation} \quad x = q^{i-1}y, \quad 1 \le i \le k,$$

with vertices X, Y, and Z, respectively, and the $\frac{1}{2}(k+1)(k+2)$ mesh points

$$(q^i, q^j, 1), \quad \text{for} \quad 0 \le j \le i \le k.$$

At the $(k+1)$th stage, we define $l_X^{(k+1)}$ as the line that joins X and $(q^k, q^k, 1)$, and thus has equation $y = q^k z$. Similarly, we define $l_Z^{(k+1)}$ as the line joining Z and $(q^k, 1, 1)$, that has equation $x = q^k y$. We then find the point, a "new" C_2, where the lines $l_X^{(k+1)}$ and $l_Z^{(2)}$ intersect, and a "new" C_3, where $l_Z^{(k+1)}$ and $l_X^{(2)}$ intersect. We find that $C_2 = (q^{k+1}, q^k, 1)$ and $C_3 = (q^{k+1}, 1, 1)$. The points C_2, C_3, and $C_1 = Y$ are collinear, lying on the line $x = q^{k+1}z$, which we denote by $l_Y^{(k+1)}$. We obtain further mesh points by finding all the points where this new line $l_Y^{(k+1)}$ intersects $l_X^{(i)}$, $1 \le i \le k+1$, and $l_Z^{(i)}$, $1 \le i \le k+1$. This yields two sets, each with $k+1$ points,

$$(q^{k+1}, q^{i-1}, 1),\ 1 \le i \le k+1, \quad \text{and} \quad (q^{k+1}, q^{k+2-i}, 1),\ 1 \le i \le k+1.$$

We see that the two sets have k points in common, and at stage $k+1$, we have added the $k+2$ mesh points

$$(q^{k+1}, q^i, 1), \quad 0 \le i \le k+1,$$

and three lines, $l_X^{(k+1)}$, $l_Y^{(k+1)}$, and $l_Z^{(k+1)}$. Thus, by induction, our above assumption about the mesh points and lines at stage k is justified. Figure 5.8 shows the three-pencil mesh after the third stage of its construction.

If we terminate the construction of the three-pencil mesh after the nth stage, the mesh consists of the $\frac{1}{2}(n+1)(n+2)$ points

$$(q^i, q^j, 1), \quad \text{for} \quad 0 \le j \le i \le n. \tag{5.76}$$

These are contained within the triangle UVW, say, where U is the unit point $(1,1,1)$, $V = (q^n, 1, 1)$, and $W = (q^n, q^n, 1)$. We have constructed n lines in each of the three pencils, and every mesh point except U, V, and W lies on one line of each pencil. If we add one further line to each pencil, to give the three pencils

$$x = q^i y, \quad y = q^i z, \quad x = q^i z, \quad 0 \le i \le n, \tag{5.77}$$

then every mesh point will lie on one line of each pencil. Now recall how we constructed fundamental polynomials for points in the set

$$S^n = \{(i,j) \mid i,j \ge 0,\ i+j \le n\},$$

which we introduced in (5.35). We found the fundamental polynomial for the point (i,j) by considering all lines in the three-pencil mesh associated with S^n that lie between (i,j) and the sides of the triangle containing the mesh points. We can adapt this process to find the fundamental polynomial for the point $(q^i, q^j, 1)$ in the mesh contained in the triangle UVW. We begin by writing down the product of the linear forms of all lines of the form $l_X^{(r)}$, $l_Y^{(r)}$, and $l_Z^{(r)}$ lying between $(q^i, q^j, 1)$ and the sides of the triangle UVW. This gives

$$\Lambda_{i,j}(x,y,z) = \prod_{r=0}^{j-1}(y - q^r z) \prod_{r=i+1}^{n}(x - q^r z) \prod_{r=0}^{i-j-1}(x - q^r y),$$

a homogeneous polynomial in x, y, and z that is zero at all $\frac{1}{2}(n+1)(n+2)$ points of the three-pencil mesh except at $(q^i, q^j, 1)$. Thus, if we write

$$L_{i,j}(x,y,z) = \Lambda_{i,j}(x,y,z)/\Lambda_{i,j}(q^i, q^j, 1),$$

the homogeneous polynomial

$$\sum_{i=0}^{n}\sum_{j=0}^{i} f(q^i, q^j, 1) L_{i,j}(x,y,z)$$

interpolates f at all points of the mesh.

Now that we have a three-pencil mesh on the triangle UVW, we can apply a linear transformation to it. Let $\mathbf{A}$ denote the transformation that maps U to $(0,0,a)$, V to $(b,0,b)$, and W to $(0,c,c)$, where a, b, and c are all positive. The images of U, V, and W under this transformation correspond to the points $(0,0)$, $(1,0)$, and $(0,1)$ in the Euclidean space $\mathbb{R}^2$. Then, representing these points by column vectors, we have (see (5.75))

$$\begin{bmatrix} 0 & b & 0 \\ 0 & 0 & c \\ a & b & c \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1 & q^n & q^n \\ 1 & 1 & q^n \\ 1 & 1 & 1 \end{bmatrix}, \tag{5.78}$$

and since

$$\begin{bmatrix} 1 & q^n & q^n \\ 1 & 1 & q^n \\ 1 & 1 & 1 \end{bmatrix}^{-1} = \frac{1}{1-q^n} \begin{bmatrix} 1 & 0 & -q^n \\ -1 & 1 & 0 \\ 0 & -1 & 1 \end{bmatrix} \tag{5.79}$$

if the positive number q is not equal to 1, we find that

$$\mathbf{A} = \frac{1}{1-q^n} \begin{bmatrix} -b & b & 0 \\ 0 & -c & c \\ a-b & b-c & c-q^n a \end{bmatrix}. \tag{5.80}$$

Under this transformation, the point $(q^i, q^j, 1)$, with $0 \le j \le i \le n$, is mapped onto the point $(x_{i,j}, y_{i,j}, z_{i,j})$, where

$$x_{i,j} = bq^j \frac{[i-j]}{[n]}, \tag{5.81}$$

$$y_{i,j} = c\frac{[j]}{[n]}, \tag{5.82}$$

$$z_{i,j} = x_{i,j} + y_{i,j} + aq^i \frac{[n-i]}{[n]}. \tag{5.83}$$

If we write

$$\begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ z \end{bmatrix}, \tag{5.84}$$

where $\mathbf{A}$ is given in (5.80), then

$$\begin{bmatrix} x \\ y \\ z \end{bmatrix} = \mathbf{A}^{-1} \begin{bmatrix} \xi \\ \eta \\ \zeta \end{bmatrix}, \tag{5.85}$$

and we find that

$$\mathbf{A}^{-1} = \begin{bmatrix} q^n b^{-1} - a^{-1} & q^n c^{-1} - a^{-1} & a^{-1} \\ b^{-1} - a^{-1} & q^n c^{-1} - a^{-1} & a^{-1} \\ b^{-1} - a^{-1} & c^{-1} - a^{-1} & a^{-1} \end{bmatrix}. \tag{5.86}$$

Then, under the transformation $\mathbf{A}$, we see from (5.85) that the lines

$$x = q^i y, \quad y = q^i z, \quad x = q^i z$$

become

$$r_1 = q^i r_2, \quad r_2 = q^i r_3, \quad r_1 = q^i r_3,$$

respectively, where $r_j = a_{j1}^{-1}\xi + a_{j2}^{-1}\eta + a_{j3}^{-1}\zeta$, and a_{jk}^{-1} denotes the (j,k)th element of $\mathbf{A}^{-1}$. If $z_{i,j} \neq 0$, the point whose homogeneous coordinates are given by (5.81), (5.82), and(5.83) corresponds to the point

$$P_{i,j} = (x_{i,j}/z_{i,j}, y_{i,j}/z_{i,j}) \tag{5.87}$$

in the Euclidean space $\mathbb{R}^2$.

We also see from (5.84) that the vertices X, Y, and Z are mapped onto the points whose homogeneous coordinates are given by the columns of the transformation matrix $\mathbf{A}$, and if $a \neq b$, $b \neq c$, and $c \neq q^n a$, these correspond to the points

$$\left(\frac{-b}{a-b}, 0\right), \quad \left(\frac{b}{b-c}, \frac{-c}{b-c}\right), \quad \left(0, \frac{c}{c-q^n a}\right), \tag{5.88}$$

respectively, in $\mathbb{R}^2$. Let us now write

$$\alpha = \frac{a}{b-a}, \quad \beta = \frac{b}{c-b}, \quad \gamma = \frac{c}{q^n a - c}. \tag{5.89}$$

Then, from (5.88), the vertices of the three pencils may be expressed in the form

$$(1+\alpha, 0), \quad (-\beta, 1+\beta), \quad (0, -\gamma). \tag{5.90}$$

If $\alpha\beta\gamma \neq 0$ and

$$\left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\beta}\right)\left(1+\frac{1}{\gamma}\right) > 0, \tag{5.91}$$

we see from (5.89) that

$$\frac{b}{a} = 1 + \frac{1}{\alpha}, \quad \frac{c}{b} = 1 + \frac{1}{\beta}, \tag{5.92}$$

and

$$q^n = \left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\beta}\right)\left(1+\frac{1}{\gamma}\right) > 0. \tag{5.93}$$

Thus any nonzero values of α, β, and γ that satisfy the inequality(5.91) determine unique positions of the three vertices, a unique value of $q > 0$, unique values of the ratios b/a and c/b, and hence a unique three-pencil mesh, determined by equations (5.81) to (5.87).

Example 5.4.1 Let us consider the special case where $a = b = c \neq 0$. It is easily verified that the points $P_{i,j}$, whose homogeneous coordinates are defined by equations (5.81) to (5.83), correspond to points in $\mathbb{R}^2$ given by

$$P_{i,j} = \left(q^j \frac{[i-j]}{[n]}, \frac{[j]}{[n]} \right), \quad 0 \leq j \leq i \leq n. \tag{5.94}$$

We note from (5.89) that when we put $a = b = c \neq 0$, we have $\alpha \to \pm\infty$ and $\beta \to \pm\infty$. Then we see from (5.90) that the vertex $(1+\alpha, 0)$ tends to infinity along the x-axis, and the vertex $(-\beta, 1+\beta)$ tends to infinity in the direction of the line $x+y=1$. Also, since $\gamma = -1/(1-q^n)$, where $q > 0$, we see from (5.90) that the third vertex, $(0, 1/(1-q^n))$, is finite unless $q = 1$, when $\gamma \to \pm\infty$ and the vertex $(0, -\gamma)$ tends to infinity along the y-axis.

The set of points defined by (5.94) all lie in the standard triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$. If we carry out the linear transformation that maps the point (x, y) to $(y, 1-x-y)$, the triangle with vertices $(0,0)$, $(1,0)$, and $(0,1)$ is mapped onto itself, and the grid of points given in (5.94) is mapped to that defined by

$$P_{i,j} = \left(\frac{[j]}{[n]}, 1 - \frac{[i]}{[n]} \right), \quad 0 \leq j \leq i \leq n. \tag{5.95}$$

This is the mesh we have already encountered in (5.74), which is just a scaled version of the mesh S_q^n, defined in (5.68). As we saw, the three-pencil mesh defined by (5.74) has one pencil parallel to the x-axis, one parallel to the y-axis, and a third pencil with vertex $(1/(1-q^n), -q^n/(1-q^n))$. On putting $q = 1$, we have the mesh consisting of the points

$$P_{i,j} = \left(\frac{i}{n}, \frac{j}{n} \right), \quad i, j \geq 0, \; i + j \leq n,$$

a scaled version of the mesh S^n given in (5.35), which consists of three pencils of parallel lines. ■

Example 5.4.2 If we choose $a = b \neq 0$, $b \neq c$, and $c \neq q^n a$, the vertex $(1+\alpha, 0)$ is sent off to infinity and the corresponding pencil of lines is parallel to the x-axis. The other two vertices are finite, since β and γ are finite. In particular, let us choose $\beta = \frac{1}{4}$ and $\gamma = \frac{1}{3}$, and let $n = 4$. It then follows from (5.92) and (5.93) that $a = b = 1$, $c = 5$, and $q^4 = 20$. The resulting mesh is illustrated in Figure 5.9. The two finite vertices are at $(-\frac{1}{4}, \frac{5}{4})$ and $(0, -\frac{1}{3})$. ■

Example 5.4.3 To obtain a three-pencil mesh with all three vertices finite, we need only choose finite values of α, β, and γ. Let us choose $\alpha = \beta = \gamma = \frac{1}{3}$ in (5.90), and let $n = 4$. This determines the ratios

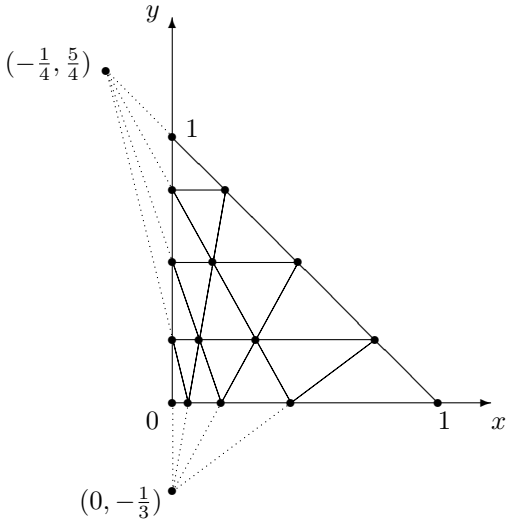

FIGURE 5.9. A three-pencil mesh with two finite vertices.

b/a and c/b. Without loss of generality we may choose $a = 1$, and it follows from (5.92) that $b = 4$ and $c = 16$. Also, we obtain from (5.93) that $q = 2\sqrt{2}$. To display the full symmetry of this mesh, let us map the point (x, y) to $(x + \frac{1}{2}y, \frac{1}{2}\sqrt{3}y)$, so that the standard right-angled triangle with vertices $(0, 0)$, $(1, 0)$, and $(0, 1)$ is mapped to the equilateral triangle with vertices $U' = (0, 0)$, $V' = (1, 0)$, and $W' = (\frac{1}{2}, \frac{1}{2}\sqrt{3})$. See Figure 5.10. The points U', V', and W' correspond to the points $U = (1, 1, 1)$, $V = (q^n, 1, 1)$, and $W = (q^n, q^n, 1)$, respectively, in the original mesh of points given in homogeneous coordinates by (5.76). The vertices X', Y', and Z' in Figure 5.10 correspond, respectively, to the vertices $X = (1, 0, 0)$, $Y = (0, 1, 0)$, and $Z = (0, 0, 1)$ of the original mesh. ■

It is not difficult to generalize the above account of three-pencil meshes from $\mathbb{R}^2$ to $\mathbb{R}^d$, for any integer $d > 2$. For notational simplicity, we will discuss the extension to $\mathbb{R}^3$ only, and the extension of these ideas for $d > 3$ is obvious. Again, it is convenient to work in homogeneous coordinates. We use a quadruple of numbers (x, y, z, t), not all zero, to denote a point. If $\lambda \neq 0$, then (x, y, z, t) and $(\lambda x, \lambda y, \lambda z, \lambda t)$ denote the same point, and if $t \neq 0$, the point (x, y, z, t) coincides with the point $(x/t, y/t, z/t)$ in the Euclidean space $\mathbb{R}^3$. The plane denoted by $ax + by + cz + d = 0$ in $\mathbb{R}^3$ is written in the form $ax + by + cz + dt = 0$ in homogeneous coordinates. The four special points $(1, 0, 0, 0)$, $(0, 1, 0, 0)$, $(0, 0, 1, 0)$, and $(0, 0, 0, 1)$ are denoted by X, Y, Z, and T, respectively. We refer to $XYZT$ as the tetrahedron of reference. (In higher dimensions, the generalizations of the plane and the tetrahedron are called the hyperplane and the simplex, respectively.) We also have the unit point, $U = (1, 1, 1, 1)$.

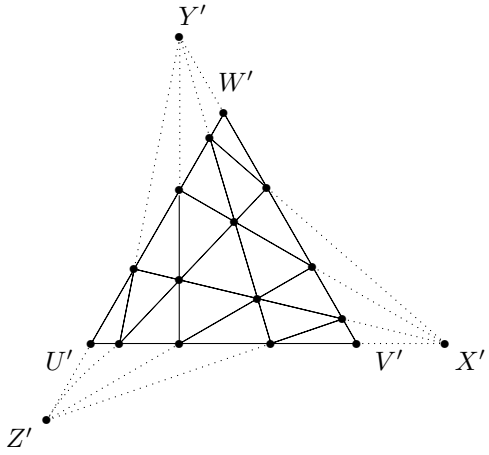

FIGURE 5.10. A symmetric three-pencil mesh with three finite vertices.

We can see that the plane YZT is given by

$$\det \begin{bmatrix} x & y & z & t \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 1 \end{bmatrix} = 0,$$

or $x = 0$, since the determinant is zero when we replace the coordinates (x, y, z, t) by those of Y, Z, or T. Similarly, ZTX, TXY, and XYZ have equations $y = 0$, $z = 0$, and $t = 0$, respectively. Note also that any plane through X has an equation of the form $by + cz + dt = 0$, and any point on ZT can be expressed in the form $(0, 0, z_1, t_1)$. We can always find a coordinate system in which *any* point not on a face of the tetrahedron of reference $XYZT$ has the coordinates $(1, 1, 1, 1)$ of the unit point. This is proved in the same way as we verified the analogous result concerning the unit point $(1, 1, 1)$ with three homogeneous coordinates.

We define a mesh of points

$$(q^i, q^j, q^k, 1), \quad 0 \le k \le j \le i \le n, \tag{5.96}$$

and the four pencils of planes

$$x = q^i y, \quad y = q^i z, \quad z = q^i t, \quad x = q^i t, \quad 0 \le i \le n, \tag{5.97}$$

which have *common lines* ZT, TX, XY, and YZ, respectively. Then, corresponding to (5.78), we have

$$\begin{bmatrix} 0 & b & 0 & 0 \\ 0 & 0 & c & 0 \\ 0 & 0 & 0 & d \\ a & b & c & d \end{bmatrix} = \mathbf{A} \begin{bmatrix} 1 & q^n & q^n & q^n \\ 1 & 1 & q^n & q^n \\ 1 & 1 & 1 & q^n \\ 1 & 1 & 1 & 1 \end{bmatrix}, \tag{5.98}$$

where a, b, c, and d are all positive. If $q \neq 1$, we find that

$$\begin{bmatrix} 1 & q^n & q^n & q^n \\ 1 & 1 & q^n & q^n \\ 1 & 1 & 1 & q^n \\ 1 & 1 & 1 & 1 \end{bmatrix}^{-1} = \frac{1}{1-q^n} \begin{bmatrix} 1 & 0 & 0 & -q^n \\ -1 & 1 & 0 & 0 \\ 0 & -1 & 1 & 0 \\ 0 & 0 & -1 & 1 \end{bmatrix}, \tag{5.99}$$

and thus the transformation matrix $\mathbf{A}$ is given by

$$\mathbf{A} = \frac{1}{1-q^n} \begin{bmatrix} -b & b & 0 & 0 \\ 0 & -c & c & 0 \\ 0 & 0 & -d & d \\ a-b & b-c & c-d & d-q^n a \end{bmatrix}, \tag{5.100}$$

and its inverse is

$$\mathbf{A}^{-1} = \begin{bmatrix} q^n b^{-1} - a^{-1} & q^n c^{-1} - a^{-1} & q^n d^{-1} - a^{-1} & a^{-1} \\ b^{-1} - a^{-1} & q^n c^{-1} - a^{-1} & q^n d^{-1} - a^{-1} & a^{-1} \\ b^{-1} - a^{-1} & c^{-1} - a^{-1} & q^n d^{-1} - a^{-1} & a^{-1} \\ b^{-1} - a^{-1} & c^{-1} - a^{-1} & d^{-1} - a^{-1} & a^{-1} \end{bmatrix}. \tag{5.101}$$

It should be clear from (5.100) and (5.101) how to write down the counterparts of $\mathbf{A}$ and $\mathbf{A}^{-1}$ in higher dimensions.

Let us introduce new variables ξ, η, ζ, and τ, defined by

$$\begin{bmatrix} \xi \\ \eta \\ \zeta \\ \tau \end{bmatrix} = \mathbf{A} \begin{bmatrix} x \\ y \\ z \\ t \end{bmatrix}. \tag{5.102}$$

Then, under the transformation $\mathbf{A}$, the planes

$$x = q^i y, \quad y = q^i z, \quad z = q^i t, \quad x = q^i t$$

become

$$r_1 = q^i r_2, \quad r_2 = q^i r_3, \quad r_3 = q^i r_4, \quad r_1 = q^i r_4, \tag{5.103}$$

respectively, where $r_j = a_{j1}^{-1}\xi + a_{j2}^{-1}\eta + a_{j3}^{-1}\zeta + a_{j4}^{-1}\tau$, and a_{jk}^{-1} denotes the (j,k)th element of $\mathbf{A}^{-1}$.

Let us write

$$\alpha = \frac{a}{b-a}, \quad \beta = \frac{b}{c-b}, \quad \gamma = \frac{c}{d-c}, \quad \delta = \frac{d}{q^n a - d}. \tag{5.104}$$

Then, from (5.100), the vertices of the tetrahedron of reference, X, Y, Z, T, are mapped to points X', Y', Z', and T', say, whose homogeneous

coordinates are given by the columns of the matrix $\mathbf{A}$ in (5.100). Thus their coordinates in $\mathbb{R}^3$ are

$$X' = (1+\alpha, 0, 0), \qquad Y' = (-\beta, 1+\beta, 0),$$
$$Z' = (0, -\gamma, 1+\gamma), \qquad T' = (0, 0, -\delta). \tag{5.105}$$

If $\alpha\beta\gamma\delta \neq 0$ and

$$\left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\beta}\right)\left(1+\frac{1}{\gamma}\right)\left(1+\frac{1}{\delta}\right) > 0, \tag{5.106}$$

we see from (5.104) that

$$\frac{b}{a} = 1+\frac{1}{\alpha}, \quad \frac{c}{b} = 1+\frac{1}{\beta}, \quad \frac{d}{c} = 1+\frac{1}{\gamma}, \tag{5.107}$$

and

$$q^n = \left(1+\frac{1}{\alpha}\right)\left(1+\frac{1}{\beta}\right)\left(1+\frac{1}{\gamma}\right)\left(1+\frac{1}{\delta}\right). \tag{5.108}$$

Thus any nonzero values of α, β, γ, and δ that satisfy the inequality (5.106) determine unique positions of the four points X', Y', Z', and T', a unique value of $q > 0$, unique values of the ratios b/a, c/b, and d/c, and hence a unique four-pencil mesh.

Under the transformation $\mathbf{A}$, defined by (5.100), the mesh point whose homogeneous coordinates are $(q^i, q^j, q^k, 1)$ is mapped to the point with coordinates given by

$$x_{i,j,k} = bq^j \frac{[i-j]}{[n]}, \tag{5.109}$$

$$y_{i,j,k} = cq^k \frac{[j-k]}{[n]}, \tag{5.110}$$

$$z_{i,j,k} = d\frac{[k]}{[n]}, \tag{5.111}$$

$$t_{i,j,k} = x_{i,j,k} + y_{i,j,k} + z_{i,j,k} + aq^i \frac{[n-i]}{[n]}. \tag{5.112}$$

Example 5.4.4 When $a = b = c = d \neq 0$, we see from equations (5.109) to (5.112) that $t_{i,j,k} = a$ for all i and j, and we obtain the grid in $\mathbb{R}^3$ whose mesh points are

$$P_{i,j,k} = \left(q^j \frac{[i-j]}{[n]}, q^k \frac{[j-k]}{[n]}, \frac{[k]}{[n]}\right), \quad 0 \leq k \leq j \leq i \leq n. \tag{5.113}$$

These points lie on four pencils of planes whose equations in homogeneous coordinates are

$$(q^n-1)\xi + (q^n-1)\eta + (q^n-1)\zeta + \tau = q^i[(q^n-1)\eta + (q^n-1)\zeta + \tau],$$

$$
\begin{aligned}
(q^n-1)\eta+(q^n-1)\zeta+\tau &= q^i[(q^n-1)\zeta+\tau],\\
(q^n-1)\zeta+\tau &= q^i\tau,\\
(q^n-1)\xi+(q^n-1)\eta+(q^n-1)\zeta+\tau &= q^i\tau,
\end{aligned}
$$

where the coefficients of ξ, η, ζ, and τ are taken from the rows of $\mathbf{A}^{-1}$, as noted in (5.103). Then, in view of our observation following (5.97), we see that the above pencils of planes have common lines $Z'T'$, $T'X'$, $X'Y'$, and $Y'Z'$, respectively. From (5.104) and (5.105), the conditions $a=b=c=d\neq 0$ imply that the points X', Y', and Z' go off to infinity and, unless $q=1$, the point T' remains finite. It follows that the first two pencils have finite common lines and the last two pencils are systems of parallel planes. Indeed, if we write $x=\xi/\tau$, $y=\eta/\tau$, and $z=\zeta/\tau$, we see that the first pencil has common line given by the intersection of the planes $x=0$ and $y+z=1/(1-q^n)$, and the second pencil has common line given by the intersection of the planes $y=0$ and $z=1/(1-q^n)$. The third and fourth pencils are the parallel systems

$$
z=\frac{[i]}{[n]},\quad 0\le i\le n,\quad \text{and}\quad x+y+z=\frac{[i]}{[n]},\quad 0\le i\le n,
$$

respectively. ■

Example 5.4.5 If we choose $a=b=c$ and $d=q^n a$, we see from equations (5.109) to (5.112) that $t_{i,j,k}=aq^k$, and we obtain the grid in $\mathbb{R}^3$ whose mesh points are

$$
P_{i,j,k}=\left(q^{j-k}\frac{[i-j]}{[n]},\frac{[j-k]}{[n]},q^{n-k}\frac{[k]}{[n]}\right),\quad 0\le k\le j\le i\le n. \tag{5.114}
$$

It is easily verified that each mesh point lies on one of each of the four pencils of planes

$$
\begin{aligned}
x+y &= \frac{[i]}{[n]},\quad 0\le i\le n,\\
y &= \frac{[i]}{[n]},\quad 0\le i\le n,\\
z &= 1-\frac{[i]}{[n]},\quad 0\le i\le n,\\
x+y+q^{i-n}z &= \frac{[i]}{[n]},\quad 0\le i\le n.
\end{aligned}
$$

The first three pencils are systems of parallel planes, and the fourth is the pencil of planes with common line

$$
x+y=\frac{1}{1-q^n},\quad z=\frac{-q^n}{1-q^n}.
$$

■

Because the application of homogeneous coordinates proved to be so helpful in our treatment of three-pencil meshes in $\mathbb{R}^2$ earlier in this section, it was natural to use homogeneous coordinates right from the start in our discussion of four-pencil meshes in $\mathbb{R}^3$. This turned out to be equally fruitful, as we have seen, and it is clear that we can use homogeneous coordinates to discuss $(N+1)$-pencil meshes in N dimensions, for any $N \geq 2$. We showed how the construction of a three-pencil mesh in $\mathbb{R}^2$ can be carried out geometrically, and justified the geometrical construction by applying Pappus's theorem. Lee and Phillips [33] discuss a theorem that justifies an analogous construction for obtaining four-pencil meshes in $\mathbb{R}^3$.

Problem 5.4.1 Verify that the n lines defined by

$$x + q^k y - [k+1] = 0, \qquad 0 \leq k \leq n-1,$$

form a pencil with vertex $(1/(1-q), 1/(1-q'))$, where $q' = 1/q$.

Problem 5.4.2 Derive the q-difference form (5.73) from the divided difference form (5.72).

Problem 5.4.3 Let $C_1 = (\alpha_2\beta_2 - \alpha_3\beta_3, \alpha_2 - \alpha_3, \beta_2 - \beta_3)$, with C_2 and C_3 defined cyclically, as in the proof of Theorem 5.4.3. Show that C_1, C_2, and C_3 lie on the line $ax + by + cz = 0$, where

$$\begin{aligned} a &= (\alpha_2\beta_3 - \alpha_3\beta_2) + (\alpha_3\beta_1 - \alpha_1\beta_3) + (\alpha_1\beta_2 - \alpha_2\beta_1), \\ b &= \alpha_1\beta_1(\beta_3 - \beta_2) + \alpha_2\beta_2(\beta_1 - \beta_3) + \alpha_3\beta_3(\beta_2 - \beta_1), \\ c &= \alpha_1\beta_1(\alpha_2 - \alpha_3) + \alpha_2\beta_2(\alpha_3 - \alpha_1) + \alpha_3\beta_3(\alpha_1 - \alpha_2). \end{aligned}$$

Problem 5.4.4 Construct a three-pencil mesh for which, in (5.90), both α and β are finite, γ is infinite, and $n = 4$.

Problem 5.4.5 Construct a three-pencil mesh for which, in (5.90), α is finite, both β and γ are infinite, and $n = 4$.

6
Splines

6.1 Introduction

In our study of numerical integration in Chapter 3 we discussed interpolatory rules, in which the integrand is replaced by an interpolating polynomial. When such a rule is applied in composite form the interval of integration is split into subintervals and the integrand is approximated by an interpolating polynomial on each subinterval. An approximation of this kind is called a piecewise polynomial. In general, a piecewise polynomial can be a discontinuous function, with discontinuities at the points where the constituent polynomials meet, or the constituent polynomials can join together smoothly to form a function that is continuous. A piecewise polynomial can be even smoother, possessing a first or higher-order derivative that is continuous. This leads us to the concept of a *spline*, which we define as follows.

Definition 6.1.1 Given $a = t_0 < t_1 < \cdots < t_N = b$, a function S is called a spline of degree $n \geq 1$ with respect to the t_i, which are called the *knots*, if the following two conditions hold:

(i) S restricted to $[t_j, t_{j+1}]$ is a polynomial of degree at most n,

(ii) $S \in C^{n-1}[a, b]$.

We refer to the latter property as the *smoothness* condition. A spline may also be defined on an infinite interval. ■

We can define splines where the smoothness condition is chosen differently or, more radically, we can define splines that are constructed from functions other than polynomials. However, in this account, we will restrict our attention to the particular splines defined above. Many of the early and important results concerning splines are due to I. J. Schoenberg (1903–1990), who is often referred to as the father of splines.

From the above definition, we see from the first condition that a linear spline S (with $n = 1$) is a sequence of straight line segments, and the second condition tells us that S is continuous. Thus a linear spline is just a polygonal arc. Two conditions are required to determine the first line segment of a polygonal arc, and then one further condition per line is required to determine each of the remaining line segments forming the spline, which is thus determined by $N + 1$ conditions.

Example 6.1.1 Let S denote the polygonal arc that connects the $N + 1$ points (t_i, y_i), $0 \le i \le N$, where

$$t_i = \left(\frac{i(i+1)}{N(N+1)}\right)^2, \quad y_i = t_i^{1/2} + \frac{1}{4N(N+1)}, \tag{6.1}$$

so that $t_0 = 0$ and $t_N = 1$. The function S is continuous on $[0, 1]$, and so is a spline approximation of degree one (a linear spline). We can verify (see Problem 6.1.3) that S restricted to the interval $[t_{i-1}, t_i]$ is the linear minimax approximation for $x^{1/2}$ on $[t_{i-1}, t_i]$, for $1 \le i \le N$. ■

Consider a spline of degree n defined on $N + 1$ knots. We require $n + 1$ conditions to determine each of the N polynomials that make up the spline, *less* n conditions at each of the $N - 1$ interior knots to satisfy the property that $S \in C^{n-1}[a, b]$. Thus, to determine a spline of degree n defined on $N + 1$ knots, we require

$$N(n + 1) - (N - 1)n = N + n$$

conditions. We can look at this in a more constructive way, using the truncated power function, defined by (4.3). Let the spline S be represented on $[t_0, t_1]$ by the polynomial

$$p_1(x) = a_0 + a_1(x - t_0) + a_2(x - t_0)^2 + \cdots + a_n(x - t_0)^n.$$

Suppose that S is represented by the polynomial p_2 on the interval $[t_1, t_2]$. It follows that p_2 must have the form

$$p_2(x) = p_1(x) + a_{n+1}(x - t_1)^n,$$

for some choice of a_{n+1}, since p_1 and p_2 must be equal, and their first $n - 1$ derivatives must be equal, at the knot t_1. Thus, using the truncated power function, we may express S on the *double* interval $[t_0, t_2]$ in the form

$$a_0 + a_1(x - t_0) + a_2(x - t_0)^2 + \cdots + a_n(x - t_0)^n + a_{n+1}(x - t_1)_+^n.$$

Note that this last expression is valid because $(x-t_1)_+^n$ and its first $n-1$ derivatives are zero for $x < t_1$. Clearly, we can use the above argument repeatedly to build up an expression for S that is valid on the whole of $[t_0, t_N]$, and we obtain

$$S(x) = \sum_{i=0}^{n} a_i (x-t_0)^i + \sum_{j=1}^{N-1} a_{n+j} (x-t_j)_+^n, \quad t_0 \le x \le t_N. \tag{6.2}$$

This last expression contains $N+n$ parameters, as we predicted before carrying out this construction. Thus any spline of degree n on the interval $[t_0, t_N]$, with intermediate knots at $t_1, \ldots, t_{N-1}$, may be written as a sum of multiples of the $N+n$ functions

$$1, x, x^2, \ldots, x^n, (x-t_1)_+^n, (x-t_2)_+^n, \ldots, (x-t_{N-1})_+^n.$$

Since these functions are linearly dependent, they are a basis for this set of splines, which is a linear space.

Problem 6.1.1 What can be said about the function S if we were to amend Definition 6.1.1 and define S as a polynomial of degree at most n on any interval of $[a, b]$ not containing a knot, but ask that $S \in C^n[a, b]$ rather than $S \in C^{n-1}[a, b]$?

Problem 6.1.2 Let us amend the expression for the spline S in (6.2) by choosing $a_0 = a_1 = \cdots = a_{n-1} = 0$ and replacing $(x-t_0)^n$ by $(x-t_0)_+^n$ in the first summation, to give the function

$$S^*(x) = \sum_{j=0}^{N-1} a_{n+j} (x-t_j)_+^n, \quad -\infty \le x < \infty.$$

Verify that $S^*(x)$ and its first $n-1$ derivatives are zero for $-\infty < x \le t_0$, so that $S^*(x)$ is a spline of degree n on the interval $-\infty < x < \infty$.

Problem 6.1.3 Consider the spline S with knots $t_0, \ldots, t_N$, as defined in Example 6.1.1. Verify that S restricted to the interval $[t_{i-1}, t_i]$ is the linear minimax approximation for $x^{1/2}$ on $[t_{i-1}, t_i]$, by showing that

$$\max_{t_{i-1} \le x \le t_i} \left| x^{1/2} - S(x) \right| = \frac{1}{4N(N+1)} = e_N$$

is attained at both endpoints t_{i-1} and t_i, and also at

$$\tau_i = \frac{1}{4} \left(t_{i-1}^{1/2} + t_i^{1/2} \right)^2 = \frac{i^4}{N^2 (N+1)^2},$$

where $t_{i-1} < \tau_i < t_i$. Show that S approximates $x^{1/2}$ on $[0, 1]$ with an error of maximum modulus e_N, which is attained at each of the $N+1$ knots t_i and also at each of the N points τ_i.

6.2 B-Splines

We saw in the last section that the $n+1$ monomials $1, x, \ldots, x^n$ together with $N-1$ truncated power functions of degree n are a basis for the linear space of splines of degree n on an interval with $N-1$ interior knots. It is convenient to extend the sequence of $N+1$ knots $t_0, \ldots, t_N$ so that they become a subset of the infinite sequence of knots

$$\cdots < t_{-2} < t_{-1} < t_0 < t_1 < t_2 < \cdots,$$

where $t_{-i} \to -\infty$ as $i \to \infty$ and $t_i \to \infty$ as $i \to \infty$, with $i > 0$. In this section we will show that an alternative basis for this linear space is the set of B-splines of order n, which we will now define recursively.

Definition 6.2.1 The B-splines of degree zero are piecewise constants defined by

$$B_i^0(x) = \begin{cases} 1, & t_i < x \le t_{i+1}, \\ 0, & \text{otherwise}, \end{cases} \tag{6.3}$$

and those of degree $n > 0$ are defined recursively in terms of those of degree $n-1$ by

$$B_i^n(x) = \left(\frac{x - t_i}{t_{i+n} - t_i}\right) B_i^{n-1}(x) + \left(\frac{t_{i+n+1} - x}{t_{i+n+1} - t_{i+1}}\right) B_{i+1}^{n-1}(x). \qquad \blacksquare \tag{6.4}$$

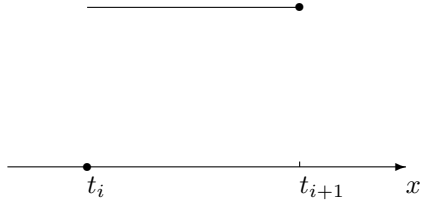

FIGURE 6.1. Graph of $B_i^0(x)$. Note that $B_i^0(t_i) = 0$ and $B_i^0(t_{i+1}) = 1$.

Definition 6.2.2 Let S denote a spline defined on the whole real line. The *interval of support* of the spline S is the smallest closed interval outside which S is zero. $\blacksquare$

Theorem 6.2.1 The interval of support of the B-spline B_i^n is $[t_i, t_{i+n+1}]$, and B_i^n is positive in the interior of this interval.

Proof. Since the interval of support of B_i^0 is $[t_i, t_{i+1}]$ and B_i^0 is positive in the interior of this interval, the above statement holds for $n = 0$ and all i.

We complete the proof by induction on n. Let us assume that the above result is true for $n-1 \geq 0$ and all i. We then deduce from (6.4) that it is true for n and all i. Thus, by induction, the theorem holds for all $n \geq 0$ and all i. ■

As we will prove later in this section, for $n > 0$, the B-spline $B_i^n(x)$ is indeed a spline of degree n, as its name and its notation suggest. It is remarkable that splines of increasing smoothness are generated by the simple recurrence relation (6.4), beginning with functions of such utter simplicity as the B-splines of degree zero, which are not even continuous. We can easily deduce from Definition 6.2.1 that

$$B_i^1(x) = \begin{cases} \dfrac{x - t_i}{t_{i+1} - t_i}, & t_i < x \leq t_{i+1}, \\ \dfrac{t_{i+2} - x}{t_{i+2} - t_{i+1}}, & t_{i+1} < x \leq t_{i+2}, \\ 0, & \text{otherwise.} \end{cases} \tag{6.5}$$

The graphs of B_i^0 and B_i^1 are shown in Figures 6.1 and 6.2, respectively. (In the graph of B_i^0, which has discontinuities at t_i and t_{i+1}, we have put a dot at each of the points $(t_i, 0)$ and $(t_{i+1}, 1)$ to emphasize that these points belong to the graph.)

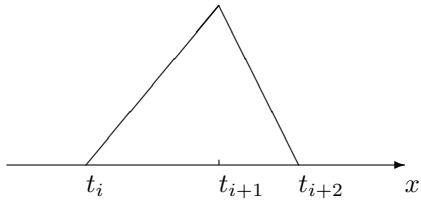

FIGURE 6.2. Graph of the B-spline $B_i^1(x)$.

Example 6.2.1 Consider the function

$$S(x) = \sum_{i=0}^{N} f(t_i) B_{i-1}^1(x), \quad t_0 \leq x \leq t_N, \tag{6.6}$$

where f is any function defined on $[t_0, t_N]$. Since the only B-splines of degree one that make a nonzero contribution to S on the subinterval $[t_j, t_{j+1}]$ are

B^1_{j-1} and B^1_j, we see from (6.5) that

$$S(x) = f(t_j)\left(\frac{t_{j+1}-x}{t_{j+1}-t_j}\right) + f(t_{j+1})\left(\frac{x-t_j}{t_{j+1}-t_j}\right), \quad t_j \le x \le t_{j+1}. \tag{6.7}$$

We observe that $S(t_j) = f(t_j)$, $S(t_{j+1}) = f(t_{j+1})$, and S is linear on $[t_j, t_{j+1}]$. Thus the expression on the right of (6.7) is the linear interpolating polynomial for f on $[t_j, t_{j+1}]$, and the function S defined on $[t_0, t_N]$ by (6.6) is the polygonal arc that connects the $N+1$ points $(t_i, f((t_i))$, $0 \le i \le N$. We call S an interpolating first-degree spline. ■

It follows from Definition 6.1.1 that the derivative of a spline of degree $n \ge 2$ is a spline of degree $n-1$. As we will see later, in Theorem 6.2.8, any spline of degree $n-1$ can be expressed as a sum of multiples of B-splines of degree $n-1$. In particular, since the interval of support of the B-spline B^n_i is $[t_i, t_{i+n+1}]$, its derivative must be expressible as a sum of multiples of those B-splines of degree $n-1$ whose intervals of support overlap that of B^n_i. The following theorem shows that in fact, the two B-splines B^{n-1}_i and B^{n-1}_{i+1} suffice, and we will deduce the smoothness properties of the B-splines from this theorem

Theorem 6.2.2 For $n \ge 2$, we have

$$\frac{d}{dx}B^n_i(x) = \left(\frac{n}{t_{i+n}-t_i}\right)B^{n-1}_i(x) - \left(\frac{n}{t_{i+n+1}-t_{i+1}}\right)B^{n-1}_{i+1}(x) \tag{6.8}$$

for all real x. For $n = 1$, (6.8) holds for all x except at the three knots t_i, t_{i+1}, and t_{i+2}, where the derivative of B^1_i is not defined.

Proof. We will first show that the equation in (6.8) holds for all real x excluding the knots t_j. It is easily verified from (6.5) and (6.3) that

$$\frac{d}{dx}B^1_i(x) = \frac{B^0_i(x)}{t_{i+1}-t_i} - \frac{B^0_{i+1}(x)}{t_{i+2}-t_{i+1}},$$

except at the knots t_i, t_{i+1}, and t_{i+2}. Thus (6.8) holds for $n = 1$ and all i, except at some of the knots. We will assume that (6.8) holds for some $n \ge 1$ and all i, except at the knots. Now let us write down (6.4) with n replaced by $n+1$, and differentiate it, using (6.8), to give

$$\frac{d}{dx}B^{n+1}_i(x) = \frac{B^n_i(x)}{t_{i+n+1}-t_i} - \frac{B^n_{i+1}(x)}{t_{i+n+2}-t_{i+1}} + nC(x), \tag{6.9}$$

say, where

$$C(x) = \left(\frac{x-t_i}{t_{i+n+1}-t_i}\right)\left(\frac{B^{n-1}_i(x)}{t_{i+n}-t_i} - \frac{B^{n-1}_{i+1}(x)}{t_{i+n+1}-t_{i+1}}\right)$$

$$+ \left(\frac{t_{i+n+2}-x}{t_{i+n+2}-t_{i+1}}\right)\left(\frac{B^{n-1}_{i+1}(x)}{t_{i+n+1}-t_{i+1}} - \frac{B^{n-1}_{i+2}(x)}{t_{i+n+2}-t_{i+2}}\right).$$

Guided by the terms $B_i^{n-1}(x)$ and $B_{i+2}^{n-1}(x)$ in the latter equation, and having the recurrence relation (6.4) in mind, we find that we can rearrange the terms involving $B_{i+1}^{n-1}(x)$ in the expression for $C(x)$ to give

$$C(x) = \left(\frac{1}{t_{i+n+1} - t_i}\right)\left(\frac{(x - t_i)B_i^{n-1}(x)}{t_{i+n} - t_i} + \frac{(t_{i+n+1} - x)B_{i+1}^{n-1}(x)}{t_{i+n+1} - t_{i+1}}\right)$$

$$- \left(\frac{1}{t_{i+n+2} - t_{i+1}}\right)\left(\frac{(x - t_{i+1})B_{i+1}^{n-1}(x)}{t_{i+n+1} - t_{i+1}} + \frac{(t_{i+n+2} - x)B_{i+2}^{n-1}(x)}{t_{i+n+2} - t_{i+2}}\right).$$

We now see from the recurrence relation (6.4) that

$$C(x) = \frac{B_i^n(x)}{t_{i+n+1} - t_i} - \frac{B_{i+1}^n(x)}{t_{i+n+2} - t_{i+1}},$$

and (6.9) shows that (6.8) holds when n is replaced by $n+1$. This completes the proof of the theorem for all real values of x, except at the knots. By an induction argument using the recurrence relation (6.4) we see that for $n \geq 1$, B_i^n is continuous for all real x. It follows that for $n \geq 2$, the right side of (6.8) is continuous for all x. Since we have just proved that (6.8) holds for all x except at the knots, this continuity argument shows that for $n \geq 2$, the relation (6.8) is valid for all x. ■

Note that if we choose the knots as $t_i = i$, for all integers i, then (6.8) simplifies to give

$$\frac{d}{dx}B_i^n(x) = B_i^{n-1}(x) - B_{i+1}^{n-1}(x). \tag{6.10}$$

Let us now replace n by $n + 1$ in (6.8), divide throughout by $n + 1$, and integrate over $[t_i, t_{i+n+2}]$. Noting that $B_i^{n+1}(x)$ is zero at the endpoints of this interval, which is its interval of support, we find that

$$\frac{1}{t_{i+n+1} - t_i}\int_{t_i}^{t_{i+n+2}} B_i^n(x)dx = \frac{1}{t_{i+n+2} - t_{i+1}}\int_{t_i}^{t_{i+n+2}} B_{i+1}^n(x)dx. \tag{6.11}$$

Since $B_i^n(x)$ is zero on $[t_{i+n+1}, t_{i+n+2}]$ and $B_{i+1}^n(x)$ is zero on $[t_i, t_{i+1}]$, we deduce from (6.11) that the average value of a B-spline B_i^n over its interval of support is independent of i and so is independent of the choice of knots. We can show (see Problem 6.2.2) that

$$\frac{1}{t_{i+n+1} - t_i}\int_{t_i}^{t_{i+n+1}} B_i^n(x)dx = \frac{1}{n+1} \tag{6.12}$$

for all integers i.

Theorem 6.2.3 For $n \geq 1$ the B-spline B_i^n is a spline of degree n on $(-\infty, \infty)$, with interval of support $[t_i, t_{i+n+1}]$.

Proof. We have already shown in Theorem 6.2.1 that B_i^n has interval of support $[t_i, t_{i+n+1}]$. It remains only to show that $B_i^n \in C^{n-1}(-\infty, \infty)$. As we have seen, this holds when $n = 1$, since $B_i^1 \in C(-\infty, \infty)$, and we will complete the proof by induction on n. Let us assume that for some $n \geq 2$, $B_i^{n-1} \in C^{n-2}(-\infty, \infty)$. Then it follows from (6.8) that the derivative of B_i^n belongs to $C^{n-2}(-\infty, \infty)$, and hence $B_i^n \in C^{n-1}(-\infty, \infty)$. ■

It is obvious from (6.3) that the B-splines of degree zero form a partition of unity, that is,

$$\sum_{i=-\infty}^{\infty} B_i^0(x) = 1, \tag{6.13}$$

and it is easy to verify that the B-splines of degree one have this property. We now state and prove a most helpful identity named after M. J. Marsden (born 1937), and deduce from it that the B-splines of any degree form a partition of unity.

Theorem 6.2.4 (Marsden's identity) For any fixed value of $n \geq 0$,

$$(t - x)^n = \sum_{i=-\infty}^{\infty} (t - t_{i+1}) \cdots (t - t_{i+n}) B_i^n(x). \tag{6.14}$$

When $n = 0$ the empty product $(t - t_{i+1}) \cdots (t - t_{i+n})$ is taken to be 1.

Proof. We have already seen from (6.13) that (6.14) holds for $n = 0$. For our next step in the proof we require the identity

$$\left(\frac{t - t_{i+n+1}}{t_i - t_{i+n+1}}\right)(t_i - x) + \left(\frac{t - t_i}{t_{i+n+1} - t_i}\right)(t_{i+n+1} - x) = t - x. \tag{6.15}$$

This is just the linear interpolating polynomial for $t - x$, regarded as a function of t, that interpolates at $t = t_i$ and $t = t_{i+n+1}$. Let us assume that (6.14) holds for some $n \geq 0$. We then multiply both sides of (6.14) by $t - x$. In the ith term on the right of (6.14) we replace $t - x$ by its linear interpolant, given in (6.15), and so obtain

$$(t - x)^{n+1} = \sum_{i=-\infty}^{\infty} (t - t_{i+1}) \cdots (t - t_{i+n+1}) \left(\frac{t_i - x}{t_i - t_{i+n+1}}\right) B_i^n(x)$$

$$+ \sum_{i=-\infty}^{\infty} (t - t_i) \cdots (t - t_{i+n}) \left(\frac{t_{i+n+1} - x}{t_{i+n+1} - t_i}\right) B_i^n(x).$$

In the second summation above, we replace i by $i + 1$, which leaves this sum unaltered, and we also make a trivial change in the first summation,

to give

$$(t-x)^{n+1} = \sum_{i=-\infty}^{\infty} (t-t_{i+1})\cdots(t-t_{i+n+1}) \left(\frac{x-t_i}{t_{i+n+1}-t_i}\right) B_i^n(x)$$

$$+ \sum_{i=-\infty}^{\infty} (t-t_{i+1})\cdots(t-t_{i+n+1}) \left(\frac{t_{i+n+2}-x}{t_{i+n+2}-t_{i+1}}\right) B_{i+1}^n(x).$$

Finally, we combine the ith terms in the above two summations, using the recurrence relation (6.4), giving

$$(t-x)^{n+1} = \sum_{i=-\infty}^{\infty} (t-t_{i+1})\cdots(t-t_{i+n+1}) B_i^{n+1}(x), \tag{6.16}$$

which completes the proof by induction. ■

Theorem 6.2.5 Given any integer $r \geq 0$, we can express the monomial x^r as a linear combination of the B-splines B_i^n, for any fixed $n \geq r$, in the form

$$\binom{n}{r} x^r = \sum_{i=-\infty}^{\infty} \sigma_r(t_{i+1},\ldots,t_{i+n}) B_i^n(x), \tag{6.17}$$

where $\sigma_r(t_{i+1},\ldots,t_{i+n})$ is the elementary symmetric function of order r in the variables $t_{i+1},\ldots,t_{i+n}$ (see Definition 1.2.1), with generating function given in (1.44). In particular, with $r=0$ in (6.17), we obtain

$$\sum_{i=-\infty}^{\infty} B_i^n(x) = 1, \tag{6.18}$$

and thus the B-splines of degree n form a partition of unity.

Proof. We see from (1.44) that

$$(1+t_{i+1}x)\cdots(1+t_{i+n}x) = \sum_{r=0}^{n} \sigma_r(t_{i+1},\ldots,t_{i+n}) x^r.$$

On replacing x by $-1/t$, and multiplying throughout by t^n, we find that

$$(t-t_{i+1})\cdots(t-t_{i+n}) = t^n \sum_{r=0}^{n} \sigma_r(t_{i+1},\ldots,t_{i+n})(-t)^{-r},$$

and we obtain (6.17), multiplied throughout by the factor $(-1)^r$, on equating coefficients of t^{n-r} on both sides of Marsden's identity (6.14). ■

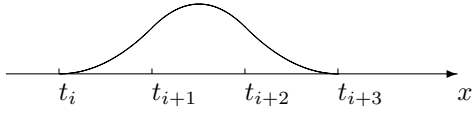

FIGURE 6.3. Graph of the B-spline $B_i^2(x)$.

Example 6.2.2 Let us derive an explicit expression for the quadratic B-spline $B_i^2(x)$, using the recurrence relation (6.4) and the expressions for $B_i^1(x)$ given in (6.5). The B-spline $B_i^2(x)$ is zero outside its interval of support, $[t_i, t_{i+3}]$, and we need to evaluate B_i^2 separately on the three subintervals $[t_i, t_{i+1}]$, $[t_{i+1}, t_{i+2}]$, and $[t_{i+2}, t_{i+3}]$. On the first of these intervals, we find that

$$B_i^2(x) = \frac{(x - t_i)^2}{(t_{i+2} - t_i)(t_{i+1} - t_i)}, \quad t_i < x \le t_{i+1}, \tag{6.19}$$

and on the third subinterval, we similarly obtain

$$B_i^2(x) = \frac{(t_{i+3} - x)^2}{(t_{i+3} - t_{i+2})(t_{i+3} - t_{i+1})}, \quad t_{i+2} < x \le t_{i+3}. \tag{6.20}$$

Since, as we know, B_i^2 and its derivative are continuous on the whole real line, we can use (6.19) and (6.20) to find the values of B_i^2 and its derivative on the *closed* intervals $[t_i, t_{i+1}]$ and $[t_{i+2}, t_{i+3}]$. From (6.19) it is clear that on $[t_i, t_{i+1}]$, B_i^2 increases monotonically from zero at $x = t_i$, where B_i^2 has a zero derivative, to $x = t_{i+1}$. Similarly, B_i^2 decreases monotonically from $x = t_{i+2}$ to $x = t_{i+3}$, where B_i^2 and its derivative are both zero. We also note that

$$0 < B_i^2(t_{i+1}) = \frac{t_{i+1} - t_i}{t_{i+2} - t_i} < 1, \tag{6.21}$$

$$0 < B_i^2(t_{i+2}) = \frac{t_{i+3} - t_{i+2}}{t_{i+3} - t_{i+1}} < 1. \tag{6.22}$$

In the middle interval, $[t_{i+1}, t_{i+2}]$, we have

$$B_i^2(x) = \frac{(x - t_i)(t_{i+2} - x)}{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})} + \frac{(t_{i+3} - x)(x - t_{i+1})}{(t_{i+3} - t_{i+1})(t_{i+2} - t_{i+1})}. \tag{6.23}$$

If we use (6.23) to evaluate $B_i^2(x)$ at the endpoints of $[t_{i+1}, t_{i+2}]$, we obtain values that agree with those given by (6.21) and (6.22), as we should expect from the continuity of B_i^2. We also know that the derivative of B_i^2 is continuous at these points. On differentiating the quadratic polynomial

in (6.23), we obtain a linear expression that is zero for $x = x^*$, say, where

$$x^* = \frac{t_{i+3}t_{i+2} - t_{i+1}t_i}{t_{i+3} + t_{i+2} - t_{i+1} - t_i}. \tag{6.24}$$

We now determine the values of $x^* - t_{i+1}$ and $t_{i+2} - x^*$, and find that

$$x^* - t_{i+1} = \frac{(t_{i+3} - t_{i+1})(t_{i+2} - t_{i+1})}{t_{i+3} + t_{i+2} - t_{i+1} - t_i} > 0$$

and

$$t_{i+2} - x^* = \frac{(t_{i+2} - t_i)(t_{i+2} - t_{i+1})}{t_{i+3} + t_{i+2} - t_{i+1} - t_i} > 0,$$

which shows that $t_{i+1} < x^* < t_{i+2}$. Thus the quadratic B-spline B_i^2 has a unique turning value at the point x^*, given by by (6.24), in the interior of $[t_{i+1}, t_{i+2}]$. Since the derivative of B_i^2 is positive at $x = t_{i+1}$ and is negative at $x = t_{i+1}$, B_i^2 has a local *maximum* at $x = x^*$. We now see that the function B_i^2 is *unimodal*: It increases monotonically from zero at t_i to its maximum value at x^* and then decreases monotonically to zero at t_{i+3}. Let us now substitute $x^* \in [t_{i+1}, t_{i+2}]$, given by by (6.24), into the right side of (6.23). After a little manipulation, we obtain

$$0 < \max_{-\infty<x<\infty} B_i^2(x) = B_i^2(x^*) = \frac{t_{i+3} - t_i}{t_{i+3} + t_{i+2} - t_{i+1} - t_i} < 1. \tag{6.25}$$

The right-hand inequality in (6.25) is consistent with the upper bound obtained for the general B-spline in Problem 6.2.3. Observe also from (6.25) that the maximum value of $B_i^2(x)$ can be made as close to 1 as we please, by taking t_{i+2} sufficiently close to t_{i+1}. ■

In Section 6.1 we used the monomials and truncated powers as a basis for splines. Our next theorem adds greatly to our understanding of the connection between splines and truncated powers, for it shows that a B-spline of degree n can be expressed as a linear combination of $n+2$ truncated powers of degree n. Not only that, but this linear combination of truncated powers is simply a multiple of a divided difference of a truncated power.

Theorem 6.2.6 For any $n \geq 0$ and all i,

$$B_i^n(x) = (t_{i+n+1} - t_i) \cdot [t_i, \ldots, t_{i+n+1}](t - x)_+^n, \tag{6.26}$$

where $[t_i, \ldots, t_{i+n+1}]$ denotes a divided difference operator of order $n + 1$ that is applied to the truncated power $(t - x)_+^n$, regarded as a function of the variable t.

Proof. The proof is by induction on n. We begin by showing that (6.26) holds for $n = 0$. (Because both $B_i^0(x)$ and $(t - x)_+^0$ are discontinuous, we need to check carefully what happens at $x = t_i$ and $x = t_{i+1}$.) Then we

assume that (6.26) holds for some $n \geq 0$, and use the recurrence relation (6.4) to write

$$B_i^{n+1}(x) = \left(\frac{x - t_i}{t_{i+n+1} - t_i}\right) B_i^n(x) + \left(\frac{t_{i+n+2} - x}{t_{i+n+2} - t_{i+1}}\right) B_{i+1}^n(x). \quad (6.27)$$

Let us now write

$$(t - x)_+^{n+1} = (t - x) \cdot (t - x)_+^n, \quad n \geq 0.$$

Thus we may express a divided difference of $(t - x)_+^{n+1}$ as a divided difference of the product of the two functions $t - x$ and $(t - x)_+^n$, and apply (1.86) to give

$$[t_i, \ldots, t_{i+n+1}](t - x)_+^{n+1} = (t_i - x) \cdot [t_i, \ldots, t_{i+n+1}](t - x)_+^n + [t_{i+1}, \ldots, t_{i+n+1}](t - x)_+^n. \quad (6.28)$$

From (6.26) the first term on the right of (6.27) is equivalent to

$$(x - t_i) \cdot [t_i, \ldots, t_{i+n+1}](t - x)_+^n = \beta_i,$$

say. Then, on using (6.28), we obtain

$$\beta_i = [t_{i+1}, \ldots, t_{i+n+1}](t - x)_+^n - [t_i, \ldots, t_{i+n+1}](t - x)_+^{n+1}. \quad (6.29)$$

Similarly, the second term on the right of (6.27) is equivalent to

$$(t_{i+n+2} - x) \cdot [t_{i+1}, \ldots, t_{i+n+2}](t - x)_+^n = \gamma_i,$$

say. We now write

$$\gamma_i = \{(t_{i+1} - x) + (t_{i+n+2} - t_{i+1})\} \cdot [t_{i+1}, \ldots, t_{i+n+2}](t - x)_+^n,$$

and apply (6.28) to give

$$\gamma_i = [t_{i+1}, \ldots, t_{i+n+2}](t - x)_+^{n+1} - [t_{i+2}, \ldots, t_{i+n+2}](t - x)_+^n + (t_{i+n+2} - t_{i+1}) \cdot [t_{i+1}, \ldots, t_{i+n+2}](t - x)_+^n. \quad (6.30)$$

If we now add β_i, given by (6.29), to γ_i, given by (6.30), we find that $\beta_i + \gamma_i$ is the sum of five terms, of which three cancel, leaving only the second term on the right of (6.29) and the first term on the right of (6.30). Thus we obtain

$$\beta_i + \gamma_i = [t_{i+1}, \ldots, t_{i+n+2}](t - x)_+^{n+1} - [t_i, \ldots, t_{i+n+1}](t - x)_+^{n+1},$$

so that from this last equation and (6.27),

$$B_i^{n+1}(x) = \beta_i + \gamma_i = (t_{i+n+2} - t_i) \cdot [t_i, \ldots, t_{i+n+2}](t - x)_+^{n+1},$$

which completes the proof. ■

Having shown in the last theorem that a B-spline of degree n can be expressed as a sum of multiples of truncated powers of degree n, we now obtain a converse result, that a truncated power can be written as a sum of multiples of B-splines.

Theorem 6.2.7 For any knot t_j and any integer $n \geq 0$,

$$(t_j - x)_+^n = \sum_{i=-\infty}^{j-n-1} (t_j - t_{i+1}) \cdots (t_j - t_{i+n}) B_i^n(x). \tag{6.31}$$

Proof. We begin by writing

$$(t_j - x)^n = \sum_{i=-\infty}^{\infty} (t_j - t_{i+1}) \cdots (t_j - t_{i+n}) B_i^n(x), \tag{6.32}$$

using Marsden's identity (6.14). Since in (6.32), $B_i^n(x)$ is multiplied by the product $(t_j - t_{i+1}) \cdots (t_j - t_{i+n})$, there is no contribution to $(t_j - x)^n$ from terms involving $B_i^n(x)$ for which $j - n \leq i \leq j - 1$. In addition, by considering the interval of support of $B_i^n(x)$, there is no contribution to the truncated power $(t_j - x)_+^n$ from terms involving $B_i^n(x)$ for which $i \geq j$, and this justifies (6.31). ■

Theorems 6.2.5 and 6.2.7 lead us to the following result, which shows the importance of the B-splines.

Theorem 6.2.8 For any integer $n \geq 0$, the B-splines of degree n are a basis for splines of degree n defined on the knots t_i.

Proof. This result is obviously true for $n = 0$. We saw earlier that the monomials $1, x, \ldots, x^n$, together with the truncated powers $(x - t_i)_+^n$, are a basis for splines of degree n defined on the knots t_i. Since (see Problem 6.2.6) we have

$$(t_i - x)^n = (t_i - x)_+^n + (-1)^n (x - t_i)_+^n,$$

for any integer $n \geq 1$, we can replace each function $(x - t_i)_+^n$ in the basis by $(t_i - x)_+^n$. It then follows from Theorems 6.2.5 and 6.2.7 that each spline of degree n can be expressed as a sum of multiples of the B_i^n, and since these B-splines are linearly independent, they are a basis. ■

Problem 6.2.1 When $t_i = i$ for all integers i, deduce from (6.10), using induction on k, that

$$\frac{d^k}{dx^k} B_i^n(x) = \sum_{r=0}^{k} (-1)^r \binom{k}{r} B_{i+r}^{n-k}(x), \quad 0 \leq k \leq n.$$

Problem 6.2.2 For $t_i \le t \le t_{i+n+1}$, write

$$\int_{t_i}^{t_{i+n+1}} (t-x)_+^n \, dx = \int_{t_i}^{t} (t-x)^n dx = \frac{(t-t_i)^{n+1}}{n+1},$$

deduce from (6.26) that

$$\frac{1}{t_{i+n+1}-t_i} \int_{t_i}^{t_{i+n+1}} B_i^n(x)dx = \frac{1}{n+1}\,[t_i,\ldots,t_{i+n+1}](t-t_i)^{n+1},$$

and hence, using (1.33), show that

$$\frac{1}{t_{i+n+1}-t_i} \int_{t_i}^{t_{i+n+1}} B_i^n(x)dx = \frac{1}{n+1}.$$

Problem 6.2.3 Deduce from Theorem 6.2.1 and (6.18) that

$$0 < \max_{-\infty<x<\infty} B_i^n(x) \le 1$$

for all $n \ge 0$ and all i.

Problem 6.2.4 Verify (see Example 6.2.2) that

$$B_i^2(x^*) - B_i^2(t_{i+1}) = \frac{x^* - t_{i+1}}{t_{i+2}-t_i} > 0$$

and

$$B_i^2(x^*) - B_i^2(t_{i+2}) = \frac{t_{i+2} - x^*}{t_{i+3}-t_{i+1}} > 0,$$

where x^*, given by (6.24), is the point where B_i^2 attains its maximum value.

Problem 6.2.5 Use the recurrence relation (6.4) and the formulas obtained for the general quadratic B-spline in Example 6.2.2 to derive the cubic B-spline, say $C(x)$, on the knots -1, $-t$, 0, t, and 1, where $0 < t < 1$. Show that $C(x)$ is the even function for which

$$C(x) = \begin{cases} 0, & x \le -1, \\ \dfrac{(1+x)^3}{1-t^2}, & -1 < x \le -t, \\ \dfrac{t^2 - 3tx^2 - (1+t)x^3}{t^2(1+t)}, & -t < x \le 0. \end{cases}$$

Verify that $C(\pm 1) = 0$, $C(\pm t) = (1-t)^2/(1+t)$, and that the maximum value of $C(x)$ is $1/(1+t)$, attained at $x = 0$. Note that this maximum value can be made as close to 1 as we please, by taking t sufficiently close to zero.

Problem 6.2.6 Show, by checking the two cases $x \le t_i$ and $x > t_i$, that

$$(t_i - x)^n = (t_i - x)_+^n + (-1)^n (x - t_i)_+^n,$$

for any positive integer n.

6.3 Equally Spaced Knots

The B-splines are greatly simplified when we choose the knots t_i to be equally spaced. It is easiest to choose $t_i = i$, for all integers i, and then any system of equally spaced knots is obtained by shifting the origin and scaling. When $t_i = i$, the recurrence relation (6.4) becomes

$$B_i^n(x) = \frac{1}{n}(x-i)B_i^{n-1}(x) + \frac{1}{n}(i+n+1-x)B_{i+1}^{n-1}(x), \tag{6.33}$$

and the B-spline $B_i^0(x)$ is given by

$$B_i^0(x) = \begin{cases} 1, & i < x \le i+1, \\ 0, & \text{otherwise.} \end{cases} \tag{6.34}$$

These are called *uniform* B-splines, and all such B-splines of the same degree are *translates* of one another, for we have

$$B_i^n(x) = B_{i+1}^n(x+1), \quad -\infty < x < \infty, \tag{6.35}$$

for all integers i. We also find that each B-spline B_i^n is symmetric about the centre of its interval of support $[i, i+n+1]$, so that

$$B_i^n(x) = B_i^n(2i+n+1-x), \quad -\infty < x < \infty, \tag{6.36}$$

for all integers i. Both (6.35) and (6.36) can be proved by induction on n.

Example 6.3.1 If we replace t_i by i in Example 6.2.2, we see from (6.19), (6.23), and (6.20) that

$$B_i^2(x) = \begin{cases} \frac{1}{2}(x-i)^2, & i < x \le i+1, \\ \frac{3}{4} - \left(x - (i+\frac{3}{2})\right)^2, & i+1 < x \le i+2, \\ \frac{1}{2}(i+3-x)^2, & i+2 < x \le i+3, \\ 0, & \text{otherwise.} \end{cases} \tag{6.37}$$

We see from (6.37) that the quadratic B-spline B_i^2 takes the value $\frac{1}{2}$ at each of the knots $i+1$ and $i+2$, and the form in which $B_i^2(x)$ is expressed on the interval $[i+1, i+2]$ makes it clear that $B_i^2(x)$ attains its maximum value of $\frac{3}{4}$ at $x = i + \frac{3}{2}$, the midpoint of the interval of support $[i, i+3]$. It is obvious from (6.37) that $B_i^2(x)$ is symmetric about the point $x = i + \frac{3}{2}$. These observations are consistent with our findings in Example 6.2.2. Finally, we note that the derivative of $B_i^2(x)$ has the value 1 at $x = i+1$, the value -1 at $x = i+2$, and is zero at all the other knots. ■

Example 6.3.2 Let us now derive the uniform cubic B-spline $B_i^3(x)$, by applying the recurrence relation (6.33) with $n = 3$, and using (6.37). Since this B-spline is symmetric about the midpoint of its interval of support $[i, i+4]$, it will suffice to compute the value of $B_i^3(x)$ on the two subintervals $[i, i+1]$ and $[i+1, i+2]$. Thus we find that $B_i^3(x)$ is the function that satisfies

$$B_i^3(x) = \begin{cases} 0, & x \le i, \\ \frac{1}{6}(x-i)^3, & i < x \le i+1, \\ \frac{2}{3} - \frac{1}{2}(x-i)(i+2-x)^2, & i+1 < x \le i+2, \end{cases} \tag{6.38}$$

and is symmetric about the knot $x = i + 2$, so that

$$B_i^3(x) = B_i^3(2i + 4 - x). \tag{6.39}$$

Thus (6.38) defines $B_i^3(x)$ on $(-\infty, i + 2)$, and the symmetry condition (6.39) extends the definition of $B_i^3(x)$ to the whole real line. We see that $B_i^3(x)$ has the value $\frac{1}{6}$ at the knots $i + 1$ and $i + 3$, attains its maximum value of $\frac{2}{3}$ at the knot $i+2$, and is zero on all the other knots. We also note that the derivative of $B_i^3(x)$ has the value $\frac{1}{2}$ at $x = i + 1$, the value $-\frac{1}{2}$ at $x = i + 3$, and is zero at all the other knots. ■

Having determined the uniform cubic B-spline explicitly in Example 6.3.2, we now evaluate all the uniform B-splines at the knots.

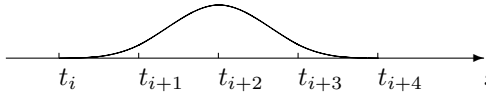

FIGURE 6.4. Graph of the B-spline $B_i^3(x)$.

Theorem 6.3.1 Consider the uniform B-splines whose knots are at the integers. Then we have

$$B_0^n(j) = \frac{1}{n!} \sum_{r=0}^{j-1} (-1)^r \binom{n+1}{r} (j-r)^n, \qquad 1 \le j \le n, \tag{6.40}$$

and $B_0^n(j) = 0$ otherwise.

Proof. This is a special case of Theorem 6.4.4, which we will prove later in this chapter. We obtain (6.40) by putting $q = 1$ in (6.85). ■

Suppose we wish to construct a spline of degree n to approximate a given function f on the interval $[0, N]$. We can write such a spline in the form

$$S(x) = \sum_{r=-n}^{N-1} a_r B_r^n(x), \tag{6.41}$$

a sum of multiples of all the uniform B-splines of degree n whose interval of support contains at least one of the N subintervals $[j-1, j]$, where $1 \le j \le N$.

If we put $n = 1$ in (6.41), we have a *linear spline*, which is simply a polygonal arc, as we saw in Example 6.2.1. On putting $n = 2$ in (6.41) we obtain the *quadratic spline*

$$S(x) = \sum_{r=-2}^{N-1} a_r B_r^2(x), \quad 0 \le x \le N. \tag{6.42}$$

We see from Example 6.3.1 that B_i^2 has the value $\frac{1}{2}$ at each of the knots $i+1$ and $i+2$ and is zero at all other knots. If we choose S so that it interpolates a given function f at the knots, it follows from (6.42) with $x = i$ that

$$\frac{1}{2}(a_{i-2} + a_{i-1}) = f(i), \tag{6.43}$$

and this holds for $0 \le i \le N$, giving $N+1$ equations. We require one further equation, since in (6.42) we need to determine $N+2$ coefficients a_r. For example, we may impose the condition $S'(0) = f'(0)$, and (see Example 6.3.1) this gives the further equation

$$-a_{-2} + a_{-1} = f'(0). \tag{6.44}$$

Then we find from (6.44) and equation (6.43) with $i = 0$ that

$$a_{-2} = f(0) - \frac{1}{2} f'(0) \tag{6.45}$$

and

$$a_{-1} = f(0) + \frac{1}{2} f'(0). \tag{6.46}$$

Thus, given a function f defined on $[0, N]$, and differentiable at $x = 0$, we can derive the quadratic spline in (6.42) as follows. We first compute a_{-2} from (6.45) and a_{-1} from (6.46). Then, using (6.43), we compute a_0, $a_1, \ldots, a_{N-1}$ recursively from

$$a_i = 2f(i+1) - a_{i-1}, \quad 0 \le i \le N-1. \tag{6.47}$$

Example 6.3.3 Let us compute a quadratic spline approximation to e^x on $[0, 1]$ with $N+1$ equally spaced knots. This is equivalent to finding

an approximation to $e^{x/N}$ on $[0, N]$. Thus we will compute the quadratic spline S on $[0, N]$, as defined by (6.42), for the function $e^{x/N}$. We obtain

$$a_{-2} = 1 - \frac{1}{2N}, \quad a_{-1} = 1 + \frac{1}{2N},$$

and

$$a_i = 2e^{(i+1)/N} - a_{i-1}, \quad 0 \le i \le N-1.$$

For example, with $N = 5$ the coefficients a_{-2}, $a_{-1}, \ldots, a_4$ are 0.9, 1.1, 1.3428, 1.6408, 2.0034, 2.4477, and 2.9889, where the last five coefficients are rounded to four decimal places. Note that when we evaluate the resulting spline $S(x)$ for any given value of x, there are at most three nonzero terms in the sum on the right of (6.42). In the table that follows we evaluate the spline and the exponential function at the midpoints between consecutive knots, to test the accuracy of the approximation. The values of the spline are obtained using the relation

$$S\left(i + \frac{1}{2}\right) = \frac{1}{8}a_{i-2} + \frac{3}{4}a_{i-1} + \frac{1}{8}a_i, \tag{6.48}$$

which follows from (6.42) and (6.37):

x	0.5	1.5	2.5	3.5	4.5
$S(x)$	1.1054	1.3497	1.6489	2.0136	2.4598
$e^{x/5}$	1.1052	1.3499	1.6487	2.0138	2.4596

Of course, by construction, $S(x)$ and $e^{x/5}$ are equal at the knots, $x = 0$, 1, 2, 3, 4, and 5. ■

The above uniform quadratic spline, which interpolates a given function f at the knots, is so easily computed that it may scarcely seem worthwhile to consider any other uniform quadratic interpolatory spline. However, it is aesthetically more pleasing to construct a quadratic spline that interpolates f at those points where the uniform quadratic B-splines have their maximum values. This requires us to interpolate at the points $x = i + \frac{1}{2}$, for $0 \le i \le N-1$. In view of (6.48), this gives the N conditions

$$\frac{1}{8}a_{i-2} + \frac{3}{4}a_{i-1} + \frac{1}{8}a_i = f\left(i + \frac{1}{2}\right), \quad 0 \le i \le N-1. \tag{6.49}$$

We require two further conditions to make up the $N+2$ conditions required to determine the $N+2$ coefficients a_i on the right of (6.42). It seems appropriate to ask that S interpolate f also at the endpoints $x = 0$ and $x = N$, which gives the conditions

$$\frac{1}{2}(a_{-2} + a_{-1}) = f(0) \tag{6.50}$$

and

$$\frac{1}{2}(a_{N-2} + a_{N-1}) = f(N). \tag{6.51}$$

We remark, in passing, that the conditions imposed on this quadratic spline are symmetric with respect to the interval $[0, N]$. This is not true of the first type of quadratic spline, as constructed in Example 6.3.3. Continuing with our construction of this second type of quadratic spline, we now eliminate a_{-2} between equation (6.49) with $i = 0$ and (6.50), to give

$$\frac{5}{8}a_{-1} + \frac{1}{8}a_0 = f\left(\frac{1}{2}\right) - \frac{1}{4}f(0), \tag{6.52}$$

and eliminate a_{N-1} between equation (6.49) with $i = N - 1$ and (6.51), to give

$$\frac{1}{8}a_{N-3} + \frac{5}{8}a_{N-2} = f\left(N - \frac{1}{2}\right) - \frac{1}{4}f(N). \tag{6.53}$$

Thus (6.52), (6.53), and (6.49) for $1 \leq i \leq N - 2$ only, give an $N \times N$ system of linear equations to determine $a_{-1}, a_0, \ldots, a_{N-2}$. It is convenient to multiply each equation throughout by the factor 8, and we obtain the tridiagonal system

$$\mathbf{Ma} = \mathbf{b}, \tag{6.54}$$

where

$$\mathbf{a}^T = [a_{-1}, a_0, \ldots, a_{N-2}],$$

$$\mathbf{b}^T = \left[8f\left(\frac{1}{2}\right) - 2f(0), 8f\left(\frac{3}{2}\right), \ldots, 8f\left(N - \frac{3}{2}\right), 8f\left(N - \frac{1}{2}\right) - 2f(N)\right],$$

and $\mathbf{M}$ is the $N \times N$ matrix

$$\mathbf{M} = \begin{bmatrix} 5 & 1 & & & & & & \\ 1 & 6 & 1 & & & & & \\ & 1 & 6 & 1 & & & & \\ & & . & . & . & & & \\ & & & . & . & . & & \\ & & & & 1 & 6 & 1 & \\ & & & & & 1 & 5 & \end{bmatrix}. \tag{6.55}$$

The elements of the matrix $\mathbf{M}$ are zero except on the main diagonal and the diagonals immediately above and below the main diagonal. Such a matrix is called *tridiagonal*. It is also a *strictly diagonally dominant* matrix, a square matrix in which the modulus of each element on the main diagonal is greater than the sum of the moduli of all the other elements in the same row. It is well known (see, for example, Phillips and Taylor [45]) that a strictly diagonally dominant matrix is nonsingular. Thus $\mathbf{M}$ is nonsingular, and the system of linear equations (6.54) has a unique solution. After solving the

linear equations to determine the values of $a_{-1}, a_0, \ldots, a_{N-2}$, we compute a_{-2} and a_{N-1} from

$$a_{-2} = 2f(0) - a_{-1} \quad \text{and} \quad a_{N-1} = 2f(N) - a_{N-2}, \tag{6.56}$$

respectively. A tridiagonal system is much easier to solve than a linear system with a full matrix, that is, one whose elements are mostly nonzero. In solving a tridiagonal system we first "remove" the elements in the diagonal immediately below the main diagonal, using the row operations of Gaussian elimination (see, for example, Phillips and Taylor [45]), and then solve the resulting system of linear equations by using back substitution.

Example 6.3.4 Let us again obtain an approximation to the exponential function, this time finding the uniform quadratic spline that interpolates the function $f(x) = e^{x/N}$ at the knots $x = 0$ and $x = N$, and at the midpoint of each interval $[i, i+1]$, for $0 \le i \le N-1$. We will choose $N = 5$ and solve the 5×5 tridiagonal system defined by (6.54). We use Gaussian elimination to remove the elements below the main diagonal of the tridiagonal matrix $\mathbf{M}$ and then use back substitution to find, in turn, a_3, $a_2, \ldots, a_{-1}$. Then we find a_{-2} and a_4 from (6.56). The coefficients $a_{-2}, a_{-1}, \ldots, a_4$ are 0.90035, 1.09965, 1.34312, 1, 64049, 2.00370, 2.44731, and 2.98925, to five decimal places. As a check on the accuracy, we will compare the values of the spline and the exponential function at the interior knots $x = 1$, 2, 3, and 4, remembering that the spline interpolates the exponential function at $x = 0$ and $x = 5$:

x	1	2	3	4
$S(x)$	1.22139	1.49181	1.82210	2.22551
$e^{x/5}$	1.22140	1.49182	1.82212	2.22554

Note that $S(i) = \frac{1}{2}(a_{i-2} + a_{i-1})$. ■

Let us now consider a cubic spline on $[0, N]$, of the form (6.41) with $n = 3$. We will write

$$C(x) = B^3_{-2}(x), \tag{6.57}$$

as an alternative notation for the uniform cubic B-spline whose interval of support is $[-2, 2]$. Then, in view of (6.35), we can express any uniform cubic B-spline on $[0, N]$ as

$$S(x) = \sum_{r=-3}^{N-1} a_r C(x - r - 2). \tag{6.58}$$

Now we will seek an *interpolating spline* S, of the form given in (6.58), that interpolates a given function f at the knots $0, 1, \ldots, N$. This gives $N + 1$

conditions, and since there are $N+3$ coefficients a_i in (6.58), we need to impose two further conditions on S to determine a unique spline. One possibility is to choose the interpolating spline so that $S''(0) = S''(N) = 0$. This is called a *natural spline*. Another possibility, which is the one we will pursue here, is to choose the interpolating spline S for which

$$S'(0) = f'(0) \quad \text{and} \quad S'(N) = f'(N). \tag{6.59}$$

It follows from (6.57) and the results derived in Example 6.3.2 that

$$C(0) = \frac{2}{3}, \quad C(\pm 1) = \frac{1}{6}, \tag{6.60}$$

and $C(x)$ is zero at all other knots. We also have

$$C'(-1) = \frac{1}{2}, \quad C'(1) = -\frac{1}{2}, \tag{6.61}$$

and $C'(x)$ is zero at all other knots. If we put $x = i$ in (6.58), we see from (6.60) that the only nonzero terms in the summation are those for which $r = i-3$, $i-2$, and $i-1$, and this yields

$$S(i) = \frac{1}{6}(a_{i-3} + 4a_{i-2} + a_{i-1}), \quad 0 \le i \le N. \tag{6.62}$$

On differentiating (6.58), we obtain

$$S'(x) = \sum_{r=-3}^{N-1} a_r C'(x - r - 2). \tag{6.63}$$

When $x = i$, the only nonzero terms in the latter summation are those corresponding to $r = i-3$ and $r = i-1$, which gives

$$S'(i) = \frac{1}{2}(a_{i-1} - a_{i-3}), \quad 0 \le i \le N. \tag{6.64}$$

We now determine a cubic spline S that interpolates a given function f at the $N+1$ knots $0, 1, \ldots, N$ and whose derivative matches that of f at $x = 0$ and $x = N$. It is clear from (6.62) and (6.64) that the $N+3$ coefficients a_i for such a spline must satisfy the $N+3$ linear equations

$$\frac{1}{2}(a_{-1} - a_{-3}) = f'(0), \tag{6.65}$$

$$\frac{1}{2}(a_{N-1} - a_{N-3}) = f'(N), \tag{6.66}$$

and

$$\frac{1}{6}(a_{i-3} + 4a_{i-2} + a_{i-1}) = f(i), \qquad 0 \le i \le N. \tag{6.67}$$

We can simplify this a little by eliminating the coefficient a_{-3} between equation (6.65) and (6.67) with $i = 0$, to give

$$\frac{1}{6}(4a_{-2} + 2a_{-1}) = f(0) + \frac{1}{3}f'(0), \tag{6.68}$$

and also eliminate a_{N-1} between equation (6.66) and (6.67) with $i = N$, to give

$$\frac{1}{6}(2a_{N-3} + 4a_{N-2}) = f(N) - \frac{1}{3}f'(N). \tag{6.69}$$

We therefore solve the system of $N+1$ linear equations consisting of (6.68) and (6.69), together with (6.67) for $1 \le i \le N-1$ only, to determine the $N+1$ unknowns $a_{-2}, a_{-1}, \ldots, a_{N-2}$. After solving this linear system, we need to compute a_{-3} and a_{N-1} from (6.65) and (6.66), respectively. Before solving the linear system it is helpful to multiply throughout by the factor 6, to give

$$\mathbf{Ma} = \mathbf{b}, \tag{6.70}$$

where

$$\mathbf{a}^T = [a_{-2}, a_{-1}, \ldots, a_{N-2}],$$

$$\mathbf{b}^T = [6f(0) + 2f'(0), 6f(1), \ldots, 6f(N-1), 6f(N) - 2f'(N)],$$

and $\mathbf{M}$ is the $(N+1) \times (N+1)$ matrix

$$\mathbf{M} = \begin{bmatrix} 4 & 2 & & & & & & \\ 1 & 4 & 1 & & & & & \\ & 1 & 4 & 1 & & & & \\ & & . & . & . & & & \\ & & & . & . & . & & \\ & & & & 1 & 4 & 1 & \\ & & & & & 2 & 4 & \end{bmatrix}. \tag{6.71}$$

Since this matrix $\mathbf{M}$ is diagonally dominant, the above system of linear equations has a unique solution. In fact (see Problem 6.3.4), for the above matrix $\mathbf{M}$,

$$\det \mathbf{M} = 2\sqrt{3}\left((2+\sqrt{3})^N - (2-\sqrt{3})^N\right), \tag{6.72}$$

for $N \ge 1$. When $N = 1$, the matrix $\mathbf{M}$ is simply

$$\mathbf{M} = \begin{bmatrix} 4 & 2 \\ 2 & 4 \end{bmatrix}.$$

If we wish to compute a cubic spline approximation to a given function f on an interval $[a, b]$, we can choose a positive integer N and map $[a, b]$ onto $[0, N]$ by making a linear change of variable. We then evaluate the spline as described above.

Example 6.3.5 As in Examples 6.3.3 and 6.3.4, let us obtain approximations to e^x on $[0,1]$, but this time we will use cubic splines. Let us choose $N = 5$ and solve the 6×6 tridiagonal system defined by (6.70). We use Gaussian elimination to remove the elements below the main diagonal of the tridiagonal matrix $\mathbf{M}$. This gives an upper triangular system that we solve by back substitution to find, in turn, $a_3, a_2, \ldots, a_{-2}$. Finally, we obtain a_{-3} and a_4 from (6.65) and (6.66), respectively. The coefficients $a_{-3}, a_{-2}, \ldots, a_4$ are 0.813287, 0.993357, 1.213287, 1.481912, 1.810012, 2.210754, 2.700217, and 3.298067, to six decimal places. As a check on the accuracy of the interpolating spline, let us compare the values of the spline and the exponential function at the midpoints of the intervals between consecutive knots, as shown in the following table:

x	0.5	1.5	2.5	3.5	4.5
$S(x)$	1.105167	1.349853	1.648714	2.013745	2.459592
$e^{x/5}$	1.105171	1.349859	1.648721	2.013753	2.459603

The values of $S(x)$ in the table are computed from

$$S\left(i + \frac{1}{2}\right) = \frac{1}{48}a_{i-3} + \frac{23}{48}a_{i-2} + \frac{23}{48}a_{i-1} + \frac{1}{48}a_i,$$

which follows from (6.58), (6.57), and (6.38). Note that $C(\pm\frac{1}{2}) = \frac{23}{48}$ and $C(\pm\frac{3}{2}) = \frac{1}{48}$. ■

Problem 6.3.1 Deduce from (6.26) and (1.73) that

$$B_i^n(x) = \frac{1}{n!}\Delta^{n+1}(i-x)_+^n,$$

where B_i^n is a uniform B-spline, and the forward difference operates on i.

Problem 6.3.2 Deduce from the result in Problem 6.2.2 that

$$\int_{-\infty}^{\infty} B_i^n(x)dx = 1,$$

for all $n \geq 0$ and all integers i, where B_i^n is a uniform B-spline.

Problem 6.3.3 Let $B_0^n(j) = a_{j,n}$, where B_0^n is a uniform B-spline. Deduce from (6.33) and (6.35) that for $n \geq 1$,

$$a_{j,n} = \left(1 + \frac{1}{n}\right)[\lambda_j a_{j,n-1} + (1-\lambda_j)a_{j-1,n-1}], \qquad 1 \leq j \leq n,$$

where $\lambda_j = j/(n+1)$, noting that $a_{0,n} = a_{n+1,n} = 0$ for all n. Beginning with $a_{0,1} = a_{2,1} = 0$ and $a_{1,1} = 1$, use the above recurrence relation

to compute the $a_{j,n}$ for $n = 2, 3$, and 4. Deduce that $a_{j,n}$ lies between $(1 + 1/n)a_{j-1,n-1}$ and $(1 + 1/n)a_{j,n-1}$. Finally, show by induction on n that the $a_{j,n}$ increase monotonically from zero to a maximum value, which is attained twice when n is even, and then decrease monotonically to zero.

Problem 6.3.4 Let us define a sequence of matrices $(\mathbf{A}_j)$, where $\mathbf{A}_j$ is the $j \times j$ matrix

$$\mathbf{A}_j = \begin{bmatrix} 4 & 1 & & & & & & \\ 1 & 4 & 1 & & & & & \\ & 1 & 4 & 1 & & & & \\ & & . & . & . & & & \\ & & & . & . & . & & \\ & & & & 1 & 4 & 1 & \\ & & & & & 2 & 4 & \end{bmatrix},$$

for $j \geq 3$, and

$$\mathbf{A}_1 = [4], \quad \mathbf{A}_2 = \begin{bmatrix} 4 & 1 \\ 2 & 4 \end{bmatrix}.$$

Expand $\det \mathbf{A}_j$ by its first row to show that

$$\det \mathbf{A}_j = 4 \det \mathbf{A}_{j-1} - \det \mathbf{A}_{j-2}, \quad j \geq 3,$$

and verify by induction on j that

$$\det \mathbf{A}_j = (2 + \sqrt{3})^j + (2 - \sqrt{3})^j, \quad j \geq 1.$$

Observe that we obtain the matrix $\mathbf{M}$, defined in (6.71), on replacing 1 by 2 in the first row of $\mathbf{A}_{N+1}$. On expanding $\det \mathbf{M}$ by its first row, show that

$$\det \mathbf{M} = 4 \det \mathbf{A}_N - 2 \det \mathbf{A}_{N-1}, \quad N \geq 2,$$

and thus justify (6.72).

Problem 6.3.5 Solve the linear system (6.70) with $N = 1$ and $f(x) = e^x$, and so find the coefficients a_{-2} and a_{-1}. Then determine the values of a_{-3} and a_0 from (6.65) and (6.66), respectively. Thus show that the resulting cubic spline S, defined by (6.58) with $N = 1$, is given explicitly by

$$S(x) = \frac{2}{3}(-5 + 2e)C(x + 1) + \frac{1}{3}(8 - 2e)C(x)$$

$$+ \frac{2}{3}(-2 + 2e)C(x - 1) + \frac{1}{3}(8 + 4e)C(x - 2).$$

Verify that the error of greatest modulus, attained at $x = \frac{1}{2}$, is approximately 4.4×10^{-3}.

6.4 Knots at the q-Integers

Having explored the topic of B-splines with knots at the integers, the uniform B-splines, we devote this section to an examination of B-splines with knots at the q-integers, which we introduced in Section 1.5. For any choice of $q > 0$, let us put $t_i = [i]$ in (6.3) and (6.4). We find that the B-splines on the q-integers are defined by

$$B_i^0(x) = \begin{cases} 1, & [i] < x \le [i+1], \\ 0, & \text{otherwise}, \end{cases} \tag{6.73}$$

and, recursively, beginning with $n = 1$,

$$B_i^n(x) = \frac{(x - [i])}{q^i[n]} B_i^{n-1}(x) + \frac{([i+n+1] - x)}{q^{i+1}[n]} B_{i+1}^{n-1}(x). \tag{6.74}$$

Although the B-splines of degree n on the q-integers are not translates of one another as we found in (6.35) for the uniform B-splines, they are very simply related to each other, as the following theorem shows.

Theorem 6.4.1 The B-splines with knots at the q-integers satisfy the relation

$$B_i^n(x) = B_{i+1}^n(qx + 1), \tag{6.75}$$

for all $n \ge 0$ and all integers i.

Proof. Let us express (6.75) in the form

$$B_i^n(x) = B_{i+1}^n(t), \quad \text{where} \quad t = qx + 1. \tag{6.76}$$

We observe that

$$\frac{x - [i]}{q^i} = \frac{t - [i+1]}{q^{i+1}}, \tag{6.77}$$

and see that $[i] < x \le [i+1]$ if and only if $[i+1] < t \le [i+2]$. It then follows immediately from (6.73) that (6.76) holds for $n = 0$ and all i. We complete the proof by induction on n. Suppose that (6.76) holds for some $n \ge 0$ and all i. Then we see from the recurrence relation (6.74) that

$$B_{i+1}^{n+1}(t) = \frac{(t - [i+1])}{q^{i+1}[n+1]} B_{i+1}^n(t) + \frac{([i+n+3] - t)}{q^{i+2}[n+1]} B_{i+2}^n(t),$$

and, using the inductive hypothesis and (6.77), we have

$$B_{i+1}^{n+1}(t) = \frac{(x - [i])}{q^i[n+1]} B_i^n(x) + \frac{([i+n+2] - x)}{q^{i+1}[n+1]} B_{i+1}^n(x).$$

Again using (6.74), we see that the right side of the latter equation yields $B_i^{n+1}(x)$, which completes the proof. ∎

Example 6.4.1 We find from (6.73) and (6.74) that the linear B-spline with knots at $[i]$, $[i+1]$, and $[i+2]$ is

$$B_i^1(x) = \begin{cases} \dfrac{x-[i]}{q^i}, & [i] < x \le [i+1], \\ \dfrac{[i+2]-x}{q^{i+1}}, & [i+1] < x \le [i+2], \\ 0, & \text{otherwise.} \end{cases} \tag{6.78}$$

It is clear that $B_i^1(x)$ increases monotonically from zero for $x \le [i]$ to its maximum value of 1 at $x = [i+1]$. For $x \ge [i+1]$, $B_i^1(x)$ decreases monotonically to the value zero for $x \ge [i+2]$.

The quadratic B-spline with knots at the q-integers is

$$B_i^2(x) = \begin{cases} \dfrac{(x-[i])^2}{q^{2i}[2]}, & [i] < x \le [i+1], \\ \dfrac{[3]}{[2]^2} - (\beta_i(x))^2, & [i+1] < x \le [i+2], \\ \dfrac{([i+3]-x)^2}{q^{2i+3}[2]}, & [i+2] < x \le [i+3], \\ 0, & \text{otherwise,} \end{cases} \tag{6.79}$$

where

$$\beta_i(x) = \frac{([i+1]-x) + q([i+2]-x)}{q^{i+1}[2]}. \tag{6.80}$$

In deriving the above expression for B_i^2 in the interval $[\,[i+1],[i+2]\,]$, we find that the corresponding form in (6.37) for the case $q = 1$ is a useful guide. We note that

$$([i+1]-x) + q([i+2]-x) = 0 \quad \text{for} \quad x = \frac{[i+1]+q[i+2]}{1+q},$$

and we see that this value of x lies between $[i+1]$ and $[i+2]$. Thus the maximum value of $B_i^2(x)$ on the interval $[\,[i+1],[i+2]\,]$ is

$$\frac{[3]}{[2]^2} = \frac{1+q+q^2}{(1+q)^2},$$

which is attained at $x = ([i+1]+q[i+2])/(1+q)$, and this is the maximum value of $B_i^2(x)$ over all x. This spline increases monotonically from the value zero at $x = [i]$ to its maximum value, and then decreases monotonically to zero at $x = [i+3]$. ■

The B-splines with knots at the q-integers, with $q \neq 1$, are not symmetric about the midpoint of the interval of support as we found in (6.36) for the uniform B-splines. However, in the two theorems that follow, we give two generalizations of (6.36). We will replace $B_i^n(x)$ by $B_i^n(x;q)$ in Theorem 6.4.2, since this theorem involves B-splines with two different values of the parameter q.

Theorem 6.4.2 The B-splines with knots at the q-integers satisfy the relation

$$B_i^n(x;q) = B_i^n(q^{-2i-n}([2i+n+1]-x);1/q), \tag{6.81}$$

for all integers $n > 0$ and i, all real $q > 0$ and x, and for $n = 0$ and all x except for $x = [i]$ and $[i+1]$. ■

Theorem 6.4.3 The B-splines with knots at the q-integers satisfy the relation

$$B_i^n(x) = q^{-n(2i+n+1)/2}(1-(1-q)x)^n B_i^n\left(\frac{[2i+n+1]-x}{1-(1-q)x}\right), \tag{6.82}$$

for all integers $n > 0$ and i, all real $q > 0$ and x, and for $n = 0$ and all x except for $x = [i]$ and $[i+1]$. ■

Proofs of these theorems are given in Koçak and Phillips [30].

We can extend (6.75) to give

$$B_i^n(x) = B_{i+m}^n(q^m x + [m]), \tag{6.83}$$

and this is easily verified by induction on m. In particular, we obtain from (6.83) that

$$B_i^n([j]) = B_{i+m}^n([j+m]), \tag{6.84}$$

showing that although these B-splines are not translates of one another unless $q = 1$, all B-splines of the same degree n and fixed q take the same values at corresponding knots. We derive the values of the B-splines at the knots in the following theorem.

Theorem 6.4.4 Consider the B-splines whose knots are at the q-integers, so that $B_i^n(x)$ has interval of support $[\,[i],[i+n+1]\,]$. Then

$$B_0^n([j]) = \frac{1}{[n]!}\sum_{r=0}^{j-1}(-1)^r q^{r(r-1)/2}\begin{bmatrix} n+1 \\ r \end{bmatrix}[j-r]^n, \quad 1 \leq j \leq n, \tag{6.85}$$

and $B_0^n([j]) = 0$ otherwise.

Proof. In view of (6.84), if we know the values of $B_0^n([j])$ for $1 \leq j \leq n$, we can evaluate any spline B_i^n at the knots. We begin with the divided

difference representation of the general B-spline, (6.26), replace each t_j by $[j]$ and apply (1.113) to give

$$B_i^n(x) = \frac{\Delta_q^{n+1}([i]-x)_+^n}{q^{n(2i+n+1)/2}[n]!}, \tag{6.86}$$

where the q-difference operator acts on $([i]-x)_+^n$ as a function of i. We now put $i=0$ and $x=[j]$ in (6.86), and use (1.118) to expand the q-difference, giving

$$B_0^n([j]) = \frac{q^{-n(n+1)/2}}{[n]!}\sum_{r=0}^{n+1}(-1)^r q^{r(r-1)/2}\begin{bmatrix} n+1 \\ r \end{bmatrix}([n+1-r]-[j])_+^n.$$

Since it follows from Definition 4.1.1 and the properties of q-integers that

$$([n+1-r]-[j])_+^n = \begin{cases} 0, & r \geq n+1-j, \\ q^{jn}[n+1-r-j]^n, & 0 \leq r < n+1-j, \end{cases}$$

we obtain

$$B_0^n([j]) = \frac{q^{-n(n+1-2j)/2}}{[n]!}\sum_{r=0}^{n-j}(-1)^r q^{r(r-1)/2}\begin{bmatrix} n+1 \\ r \end{bmatrix}[n+1-r-j]^n.$$

Now replace j by $n+1-k$ to give

$$B_0^n([n+1-k]) = \frac{q^{n(n+1-2k)/2}}{[n]!}\sum_{r=0}^{k-1}(-1)^r q^{r(r-1)/2}\begin{bmatrix} n+1 \\ r \end{bmatrix}[k-r]^n.$$

On multiplying this last equation throughout by $q^{-n(n+1-2k)/2}$ and using the result (see Problem 6.4.4)

$$B_0^n([k]) = q^{-n(n+1-2k)/2}B_0^n([n+1-k]),$$

we find that

$$B_0^n([k]) = \frac{1}{[n]!}\sum_{r=0}^{k-1}(-1)^r q^{r(r-1)/2}\begin{bmatrix} n+1 \\ r \end{bmatrix}[k-r]^n,$$

which is just (6.85) with k in place of j. ■

Example 6.4.2 For $n \geq 1$ the B-spline B_0^n has n knots in the interior of its interval of support, namely, $x=[j]$, for $1 \leq j \leq n$. Let us apply Theorem 6.4.4 with $n=1,2,3$, and 4. We find that $B_0^{[1]}(1)=1$,

$$B_0^2([1]) = \frac{1}{[2]!}, \quad B_0^2([2]) = \frac{q}{[2]!},$$

$$B_0^3([1]) = \frac{1}{[3]!}, \quad B_0^3([2]) = \frac{2q(1+q)}{[3]!}, \quad B_0^3([3]) = \frac{q^3}{[3]!},$$

and

$$B_0^4([1]) = \frac{1}{[4]!}, \quad B_0^4([2]) = \frac{q(3+5q+3q^2)}{[4]!},$$

$$B_0^4([3]) = \frac{q^3(3+5q+3q^2)}{[4]!}, \quad B_0^4([4]) = \frac{q^6}{[4]!}.$$

We can verify that for any value of n above, the sum of the n coefficients is unity, in keeping with the general result stated in Problem 6.4.5. ■

In the last section we discussed two methods for computing an interpolatory quadratic spline, and a method for computing an interpolatory cubic spline, at equally spaced knots. We can adapt all of these algorithms to compute interpolatory splines with knots at the q-integers. We conclude this section by adapting the second of the methods for computing a quadratic spline. We thus interpolate the given function f at $x_0 = [0]$, $x_{N+1} = [N]$, and also at the points

$$x_i = \frac{[i-1] + q[i]}{1+q}, \quad 1 \le i \le N, \tag{6.87}$$

where, as shown in Example 6.4.1, x_i is the point where the B-spline B_{i-2}^2, with interval of support $[\,[i-2], [i+1]\,]$, has its maximum value, given by

$$B_{i-2}^2(x_i) = \frac{[3]}{[2]^2}. \tag{6.88}$$

Then, using the explicit representation of the quadratic B-spline in Example 6.4.1, we find that

$$B_{i-2}^2(x_{i-1}) = \frac{q^2}{[2]^3} \quad \text{and} \quad B_{i-2}^2(x_{i+1}) = \frac{q}{[2]^3}, \tag{6.89}$$

where x_{i-1} and x_{i+1} are given by (6.87).

We now derive a spline of the form

$$S(x) = \sum_{r=-2}^{N-1} a_r B_r^2(x), \quad 0 \le x \le [N], \tag{6.90}$$

that satisfies the above $N+2$ interpolatory conditions. Note that the first B-spline in the above summation, $B_{-2}^2(x)$, has an interval of support that includes knots at the q-integers $[-2]$ and $[-1]$. These have the values

$$[-2] = \frac{1-q^{-2}}{1-q} = -\frac{1+q}{q^2} \quad \text{and} \quad [-1] = \frac{1-q^{-1}}{1-q} = -\frac{1}{q}.$$

We proceed as in the case with equally spaced knots. First we find from (6.90), (6.88), and (6.89) that

$$\frac{q}{[2]^3}a_{i-3} + \frac{[3]}{[2]^2}a_{i-2} + \frac{q^2}{[2]^3}a_{i-1} = f(x_i), \quad 1 \le i \le N, \tag{6.91}$$

where x_i is defined in (6.87), and setting $S(x) = f(x)$ at $x = 0$ and $[N]$ gives the two further equations

$$\frac{q}{1+q}a_{-2} + \frac{1}{1+q}a_{-1} = f(0) \tag{6.92}$$

and

$$\frac{q}{1+q}a_{N-2} + \frac{1}{1+q}a_{N-1} = f([N]). \tag{6.93}$$

We eliminate the coefficient a_{-2} between (6.92) and (6.91) with $i = 1$, and also eliminate a_{N-1} between (6.93) and (6.91) with $i = N$. Thus we obtain a system of linear equations, consisting of those obtained from (6.91), which we will write as

$$\alpha a_{i-3} + \beta a_{i-2} + \gamma a_{i-1}, \quad 2 \le i \le N-1, \tag{6.94}$$

where

$$\alpha = \frac{q}{[2]^3}, \qquad \beta = \frac{[3]}{[2]^2}, \qquad \gamma = \frac{q^2}{[2]^3}, \tag{6.95}$$

together with a first and last equation. The first equation is

$$\delta a_{-1} + \gamma a_0 = f(x_1) - \frac{1}{[2]^2}f(0), \tag{6.96}$$

where γ is given above in (6.95) and

$$\delta = \frac{q}{[2]} + \frac{q}{[2]^3}, \tag{6.97}$$

and the last equation is

$$\alpha a_{N-3} + \epsilon a_{N-2} = f(x_N) - \frac{q^2}{[2]^2}f([N]), \tag{6.98}$$

where α is given above in (6.95) and

$$\epsilon = \frac{1}{[2]} + \frac{q^2}{[2]^3}. \tag{6.99}$$

This is a system of tridiagonal equations of the form

$$\mathbf{Ma} = \mathbf{b},$$

where $\mathbf{a}^T = [a_{-1}, a_0, \ldots, a_{N-2}]$ and $\mathbf{b}^T = [b_{-1}, b_0, \ldots, b_{N-2}]$. The first and last elements of $\mathbf{b}^T$ are

$$b_{-1} = f(x_1) - \frac{1}{(1+q)^2} f(x_0), \quad b_{N-2} = f(x_N) - \frac{q^2}{(1+q)^2} f(x_{N+1}),$$

and its other elements are $b_i = f(x_{i+2})$, for $0 \le i \le N-3$. Finally, $\mathbf{M}$ is the $N \times N$ tridiagonal matrix

$$\mathbf{M} = \begin{bmatrix} \delta & \gamma & & & & & \\ \alpha & \beta & \gamma & & & & \\ & \alpha & \beta & \gamma & & & \\ & & . & . & . & & \\ & & & . & . & . & \\ & & & & \alpha & \beta & \gamma \\ & & & & & \alpha & \epsilon \end{bmatrix}.$$

For $q > 0$ every element of $\mathbf{M}$ is positive, and we find that

$$\delta - \gamma = \frac{q(2+q+q^2)}{[2]^3} > 0, \qquad \epsilon - \alpha = \frac{1+q+2q^2}{[2]^3} > 0,$$

and

$$\beta - \alpha - \gamma = \frac{[4]}{[2]^3} = \frac{1+q^2}{[2]^2} > 0.$$

Thus the matrix $\mathbf{M}$ is strictly diagonally dominant, and is therefore nonsingular, for all $q > 0$. We solve the tridiagonal system to find the coefficients $a_{-1}, a_0, \ldots, a_{N-2}$, and then we find the two remaining coefficients a_{-2} and a_{N-1} from

$$a_{-2} = ((1+q)f(0) - a_{-1})/q \quad \text{and} \quad a_{N-1} = (1+q)f([N]) - q\, a_{N-2}.$$

Example 6.4.3 Let us construct a quadratic B-spline of the kind we have derived immediately above. Let us choose the function $f(x)$ as $e^{x/[N]}$, with $N = 5$, and let us take $q = 0.95$. We solve the tridiagonal system to find a_{-1}, $a_0, \ldots, a_3$, and then find a_{-2} and a_4, as described above. We find that the coefficients $a_{-2}, a_{-1}, \ldots, a_4$ are 0.8841719, 1.1100367, 1.3778033, 1.6917329, 2.0559284, 2.4742331, and 2.9501281, to seven decimal places. As a check on the accuracy, we compare the values of the spline and the exponential function at the interior knots $x = [1]$, $[2]$, $[3]$, and $[4]$:

x	$[1]$	$[2]$	$[3]$	$[4]$
$S(x)$	1.247353	1.538793	1.878500	2.270444
$e^{x/[5]}$	1.247354	1.538793	1.878499	2.270441

It is interesting to compare the accuracy of the above results, in which we used $q = 0.95$, with the results in Example 6.3.4, where we solved the same problem with $q = 1$. ■

Problem 6.4.1 Show that the B-splines with knots at the q-integers, defined by (6.73) and (6.74), satisfy $B_i^n([j]) = B_{i+m}^n([j+m])$.

Problem 6.4.2 Verify (6.86), which expresses a spline on the q-integers as a q-difference of a truncated power.

Problem 6.4.3 Show that for the quadratic B-spline defined in Example 6.4.1, the only nonzero values of the derivative at the knots are

$$\frac{d}{dx}B_i^2([i+1]) = \frac{2}{q^i(1+q)} \quad \text{and} \quad \frac{d}{dx}B_i^2([i+2]) = \frac{-2}{q^{i+1}(1+q)}.$$

Problem 6.4.4 Deduce from Theorem 6.4.3 that

$$B_0^n([n+1-j]) = q^{n(n+1-2j)/2}B_0^n([j]).$$

Problem 6.4.5 Deduce from (6.18) that

$$\sum_{j=1}^{n} B_0^n([j]) = 1.$$

Problem 6.4.6 Show, using (6.74) and (6.84), that

$$B_0^n([j]) = \frac{[j]}{[n]}B_0^{n-1}([j]) + \frac{q^{j-1}[n+1-j]}{[n]}B_0^{n-1}([j-1]).$$

Problem 6.4.7 Use the result in Problem 6.4.6 to show, by induction on n, that we can write

$$B_0^n([j]) = \frac{p_{j,n}(q)}{[n]!},$$

where $p_{j,n}(q)$ is a polynomial in q of degree $\frac{1}{2}(j-1)(2n-j)$ with nonnegative coefficients. Deduce from the result in Problem 6.4.4 that

$$p_{n+1-j,n}(q) = q^{n(n+1-2j)/2}\,p_{j,n}(q).$$

Problem 6.4.8 Construct a quadratic spline S of the form (6.90) that interpolates a given function at the $N+1$ knots $x = [i]$, for $0 \le i \le N$, and also satisfies the condition $S'(0) = f'(0)$.

7
Bernstein Polynomials

7.1 Introduction

This chapter is concerned with sequences of polynomials named after their creator S. N. Bernstein. Given a function f on $[0,1]$, we define the Bernstein polynomial

$$B_n(f;x) = \sum_{r=0}^{n} f\left(\frac{r}{n}\right) \binom{n}{r} x^r (1-x)^{n-r} \tag{7.1}$$

for each positive integer n. Thus there is a sequence of Bernstein polynomials corresponding to each function f. As we will see later in this chapter, if f is continuous on $[0,1]$, its sequence of Bernstein polynomials converges uniformly to f on $[0,1]$, thus giving a constructive proof of Weierstrass's Theorem 2.4.1, which we stated in Chapter 2. There are several proofs of this fundamental theorem, beginning with that given by K. Weierstrass [55] in 1885. (See the Notes in E. W. Cheney's text [7]. This contains a large number of historical references in approximation theory.) Bernstein's proof [3] was published in 1912. One might wonder why Bernstein created "new" polynomials for this purpose, instead of using polynomials that were already known to mathematics. Taylor polynomials are not appropriate; for even setting aside questions of convergence, they are applicable only to functions that are infinitely differentiable, and not to all continuous functions. We can also dismiss another obvious candidate, the interpolating polynomials for f constructed at equally spaced points. For the latter sequence of polynomials does *not* converge uniformly to f for *all* $f \in C[0,1]$, and the same is true of interpolation on any other fixed sequence of abscis-

sas. However, L. Fejér [19] used a method based on Hermite interpolation in a proof published in 1930, which we will discuss in the next section.

Later in this section we will consider how Bernstein discovered his polynomials, for this is not immediately obvious. We will also see that although the convergence of the Bernstein polynomials is slow, they have compensating "shape-preserving" properties. For example, the Bernstein polynomial of a convex function is itself convex.

It is clear from (7.1) that for all $n \geq 1$,

$$B_n(f;0) = f(0) \quad \text{and} \quad B_n(f;1) = f(1), \tag{7.2}$$

so that a Bernstein polynomial for f interpolates f at both endpoints of the interval $[0,1]$.

Example 7.1.1 It follows from the binomial expansion that

$$B_n(1;x) = \sum_{r=0}^{n} \binom{n}{r} x^r (1-x)^{n-r} = (x + (1-x))^n = 1, \tag{7.3}$$

so that the Bernstein polynomial for the constant function 1 is also 1. Since

$$\frac{r}{n}\binom{n}{r} = \binom{n-1}{r-1}$$

for $1 \leq r \leq n$, the Bernstein polynomial for the function x is

$$B_n(x;x) = \sum_{r=0}^{n} \frac{r}{n}\binom{n}{r} x^r (1-x)^{n-r} = x \sum_{r=1}^{n} \binom{n-1}{r-1} x^{r-1}(1-x)^{n-r}.$$

Note that the term corresponding to $r = 0$ in the first of the above two sums is zero. On putting $s = r - 1$ in the second summation, we obtain

$$B_n(x;x) = x \sum_{s=0}^{n-1} \binom{n-1}{s} x^s (1-x)^{n-1-s} = x, \tag{7.4}$$

the last step following from (7.3) with n replaced by $n-1$. Thus the Bernstein polynomial for the function x is also x. ■

We call B_n the *Bernstein operator*; it maps a function f, defined on $[0,1]$, to $B_n f$, where the function $B_n f$ evaluated at x is denoted by $B_n(f;x)$. The Bernstein operator is obviously linear, since it follows from (7.1) that

$$B_n(\lambda f + \mu g) = \lambda\, B_n f + \mu\, B_n g, \tag{7.5}$$

for all functions f and g defined on $[0,1]$, and all real λ and μ. We now require the following definition.

Definition 7.1.1 Let L denote a linear operator that maps a function f defined on $[a,b]$ to a function Lf defined on $[c,d]$. Then L is said to be a *monotone* operator or, equivalently, a *positive* operator if

$$f(x) \geq g(x), \quad x \in [a,b] \quad \Rightarrow \quad (Lf)(x) \geq (Lg)(x), \quad x \in [c,d], \tag{7.6}$$

where we have written $(Lf)(x)$ to denote the value of the function Lf at the point $x \in [a,b]$. ■

We can see from (7.1) that B_n is a monotone operator. It then follows from the monotonicity of B_n and (7.3) that

$$m \leq f(x) \leq M, \quad x \in [0,1] \quad \Rightarrow \quad m \leq B_n(f;x) \leq M, \quad x \in [0,1]. \tag{7.7}$$

In particular, if we choose $m = 0$ in (7.7), we obtain

$$f(x) \geq 0, \quad x \in [0,1] \quad \Rightarrow \quad B_n(f;x) \geq 0, \quad x \in [0,1]. \tag{7.8}$$

It follows from (7.3), (7.4), and the linear property (7.5) that

$$B_n(ax+b;x) = ax + b, \tag{7.9}$$

for all real a and b. We therefore say that the Bernstein operator *reproduces* linear polynomials. We can deduce from the following result that the Bernstein operator does not reproduce any polynomial of degree greater than one.

Theorem 7.1.1 The Bernstein polynomial may be expressed in the form

$$B_n(f;x) = \sum_{r=0}^{n} \binom{n}{r} \Delta^r f(0)\, x^r, \tag{7.10}$$

where Δ is the forward difference operator, defined in (1.67), with step size $h = 1/n$.

Proof. Beginning with (7.1), and expanding the term $(1-x)^{n-r}$, we have

$$B_n(f;x) = \sum_{r=0}^{n} f\left(\frac{r}{n}\right) \binom{n}{r} x^r \sum_{s=0}^{n-r} (-1)^s \binom{n-r}{s} x^s.$$

Let us put $t = r + s$. We may write

$$\sum_{r=0}^{n} \sum_{s=0}^{n-r} = \sum_{t=0}^{n} \sum_{r=0}^{t}, \tag{7.11}$$

since both double summations in (7.11) are over all lattice points (r,s) lying in the triangle shown in Figure 7.1. We also have

$$\binom{n}{r} \binom{n-r}{s} = \binom{n}{t} \binom{t}{r},$$

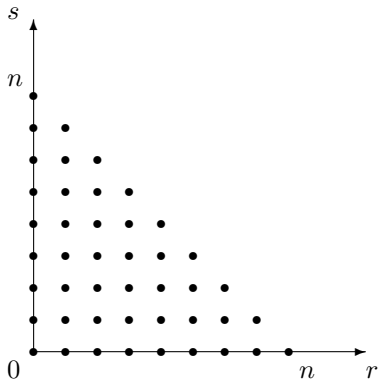

FIGURE 7.1. A triangular array of $\frac{1}{2}(n+1)(n+2)$ lattice points.

and so we may write the double summation as

$$\sum_{t=0}^{n} \binom{n}{t} x^t \sum_{r=0}^{t} (-1)^{t-r} \binom{t}{r} f\left(\frac{r}{n}\right) = \sum_{t=0}^{n} \binom{n}{t} \Delta^t f(0)\, x^t,$$

on using the expansion for a higher-order forward difference, as in Problem 1.3.7. This completes the proof. ■

In (1.80) we saw how differences are related to derivatives, showing that

$$\frac{\Delta^m f(x_0)}{h^m} = f^{(m)}(\xi), \tag{7.12}$$

where $\xi \in (x_0, x_m)$ and $x_m = x_0 + mh$. Let us put $h = 1/n$, $x_0 = 0$, and $f(x) = x^k$, where $n \geq k$. Then we have

$$n^r \Delta^r f(0) = 0 \quad \text{for} \quad r > k$$

and

$$n^k \Delta^k f(0) = f^{(k)}(\xi) = k!\,. \tag{7.13}$$

Thus we see from (7.10) with $f(x) = x^k$ and $n \geq k$ that

$$B_n(x^k; x) = a_0 x^k + a_1 x^{k-1} + \cdots + a_{k-1} x + a_k,$$

say, where $a_0 = 1$ for $k = 0$ and $k = 1$, and

$$a_0 = \binom{n}{k} \frac{k!}{n^k} = \left(1 - \frac{1}{n}\right)\left(1 - \frac{2}{n}\right) \cdots \left(1 - \frac{k-1}{n}\right)$$

for $k \geq 2$. Since $a_0 \neq 1$ when $n \geq k \geq 2$, this justifies our above statement that the Bernstein operator does not reproduce any polynomial of degree greater than one.

Example 7.1.2 With $f(x) = x^2$, we have

$$f(0) = 0, \quad \Delta f(0) = f\left(\frac{1}{n}\right) - f(0) = \frac{1}{n^2},$$

and we see from (7.13) that $n^2\Delta^2 f(0) = 2!$ for $n \geq 2$. Thus it follows from (7.10) that

$$B_n(x^2; x) = \binom{n}{1} \frac{x}{n^2} + \binom{n}{2} \frac{2x^2}{n^2} = \frac{x}{n} + \left(1 - \frac{1}{n}\right) x^2,$$

which may be written in the form

$$B_n(x^2; x) = x^2 + \frac{1}{n} x(1 - x). \tag{7.14}$$

Thus the Bernstein polynomials for x^2 converge uniformly to x^2 like $1/n$, very slowly. We will see from Voronovskaya's Theorem 7.1.10 that this rate of convergence holds for all functions that are twice differentiable. ■

We have already seen in (7.7) that if $f(x)$ is positive on $[0, 1]$, so is $B_n(f; x)$. We now show that if $f(x)$ is monotonically increasing, so is $B_n(f; x)$.

Theorem 7.1.2 The derivative of the Bernstein polynomial $B_{n+1}(f; x)$ may be expressed in the form

$$B'_{n+1}(f; x) = (n + 1) \sum_{r=0}^{n} \Delta f \left(\frac{r}{n+1}\right) \binom{n}{r} x^r (1 - x)^{n-r} \tag{7.15}$$

for $n \geq 0$, where Δ is applied with step size $h = 1/(n + 1)$. Furthermore, if f is monotonically increasing or monotonically decreasing on $[0, 1]$, so are all its Bernstein polynomials.

Proof. The verification of (7.15) is omitted because it is a special case of (7.16), concerning higher-order derivatives of the Bernstein polynomials, which we prove in the next theorem. To justify the above remark on monotonicity, we note that if f is monotonically increasing, its forward differences are nonnegative. It then follows from (7.15) that $B'_{n+1}(f; x)$ is nonnegative on $[0, 1]$, and so $B_{n+1}(f; x)$ is monotonically increasing. Similarly, we see that if f is monotonically decreasing, so is $B_{n+1}(f; x)$. ■

Theorem 7.1.3 For any integer $k \geq 0$, the kth derivative of $B_{n+k}(f; x)$ may be expressed in terms of kth differences of f as

$$B^{(k)}_{n+k}(f; x) = \frac{(n + k)!}{n!} \sum_{r=0}^{n} \Delta^k f \left(\frac{r}{n+k}\right) \binom{n}{r} x^r (1 - x)^{n-r} \tag{7.16}$$

for all $n \geq 0$, where Δ is applied with step size $h = 1/(n + k)$.

Proof. We write

$$B_{n+k}(f;x) = \sum_{r=0}^{n+k} f\left(\frac{r}{n+k}\right) \left(\begin{array}{c} n+k \\ r \end{array} \right) x^r (1-x)^{n+k-r}$$

and differentiate k times, giving

$$B_{n+k}^{(k)}(f;x) = \sum_{r=0}^{n+k} f\left(\frac{r}{n+k}\right) \left(\begin{array}{c} n+k \\ r \end{array} \right) p(x), \tag{7.17}$$

where

$$p(x) = \frac{d^k}{dx^k} x^r (1-x)^{n+k-r}.$$

We now use the Leibniz rule (1.83) to differentiate the product of x^r and $(1-x)^{n+k-r}$. First we find that

$$\frac{d^s}{dx^s} x^r = \begin{cases} \dfrac{r!}{(r-s)!} x^{r-s}, & r-s \geq 0, \\ \\ 0, & r-s < 0, \end{cases}$$

and

$$\frac{d^{k-s}}{dx^{k-s}} (1-x)^{n+k-r} = \begin{cases} (-1)^{k-s} \dfrac{(n+k-r)!}{(n+s-r)!} (1-x)^{n+s-r}, & r-s \leq n, \\ \\ 0, & r-s > n. \end{cases}$$

Thus the kth derivative of $x^r(1-x)^{n+k-r}$ is

$$p(x) = \sum_s (-1)^{k-s} \left(\begin{array}{c} k \\ s \end{array} \right) \frac{r!}{(r-s)!} \frac{(n+k-r)!}{(n+s-r)!} x^{r-s} (1-x)^{n+s-r}, \tag{7.18}$$

where the latter summation is over all s from 0 to k, subject to the constraints $0 \leq r-s \leq n$. We make the substitution $t = r-s$, so that

$$\sum_{r=0}^{n+k} \sum_s = \sum_{t=0}^{n} \sum_{s=0}^{k}. \tag{7.19}$$

A diagram may be helpful here. The double summations in (7.19) are over all lattice points (r,s) lying in the parallelogram depicted in Figure 7.2. The parallelogram is bounded by the lines $s=0$, $s=k$, $t=0$, and $t=n$, where $t = r-s$. We also note that

$$\left(\begin{array}{c} n+k \\ r \end{array} \right) \frac{r!}{(r-s)!} \frac{(n+k-r)!}{(n+s-r)!} = \frac{(n+k)!}{n!} \left(\begin{array}{c} n \\ r-s \end{array} \right). \tag{7.20}$$

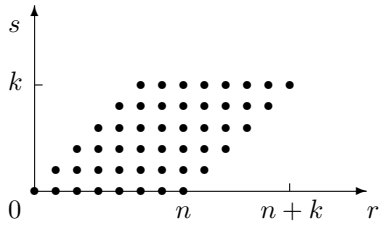

FIGURE 7.2. A parallelogram of $(n+1)(k+1)$ lattice points.

It then follows from (7.17), (7.18), (7.19), and (7.20) that the kth derivative of $B_{n+k}(f;x)$ is

$$\frac{(n+k)!}{n!}\sum_{t=0}^{n}\sum_{s=0}^{k}(-1)^{k-s}\left(\begin{array}{c} k \\ s \end{array}\right) f\left(\frac{t+s}{n+k}\right)\left(\begin{array}{c} n \\ t \end{array}\right) x^t(1-x)^{n-t}.$$

Finally, we note from Problem 1.3.7 that

$$\sum_{s=0}^{k}(-1)^{k-s}\left(\begin{array}{c} k \\ s \end{array}\right) f\left(\frac{t+s}{n+k}\right) = \Delta^k f\left(\frac{t}{n+k}\right),$$

where the operator Δ is applied with step size $h = 1/(n+k)$. This completes the proof. ■

By using the connection between differences and derivatives, we can deduce the following valuable result from Theorem 7.1.3.

Theorem 7.1.4 If $f \in C^k[0,1]$, for some $k \geq 0$, then

$$m \leq f^{(k)}(x) \leq M, \quad x \in [0,1] \quad \Rightarrow \quad c_k m \leq B_n^{(k)}(f;x) \leq c_k M, \quad x \in [0,1],$$

for all $n \geq k$, where $c_0 = c_1 = 1$ and

$$c_k = \left(1-\frac{1}{n}\right)\left(1-\frac{2}{n}\right)\cdots\left(1-\frac{k-1}{n}\right), \quad 2 \leq k \leq n.$$

Proof. We have already seen in (7.7) that this result holds for $k = 0$. For $k \geq 1$ we begin with (7.16) and replace n by $n-k$. Then, using (7.12) with $h = 1/n$, we write

$$\Delta^k f\left(\frac{r}{n}\right) = \frac{f^{(k)}(\xi_r)}{n^k}, \tag{7.21}$$

where $r/n < \xi_r < (r+k)/n$. Thus

$$B_n^{(k)}(f;x) = \sum_{r=0}^{n-k} c_k f^{(k)}(\xi_r) x^r (1-x)^{n-k-r},$$

and the theorem follows easily from the latter equation. One consequence of this result is that if $f^{(k)}(x)$ is of fixed sign on $[0,1]$, then $B_n^{(k)}(f;x)$ also has this sign on $[0,1]$. For example, if $f''(x)$ exists and is nonnegative on $[0,1]$, so that f is convex, then $B_n''(f;x)$ is also nonnegative and $B_n(f;x)$ is convex. ■

Bernstein's discovery of his polynomials was based on an ingenious probabilistic argument. Suppose we have an event that can be repeated and has only two possible outcomes, A and B. One of the simplest examples is the tossing of an unbiased coin, where the two possible outcomes, heads and tails, both occur with probability 0.5. More generally, consider an event where the outcome A happens with probability $x \in [0,1]$, and thus the outcome B happens with probability $1-x$. Then the probability of A happening precisely r times followed by B happening $n-r$ times is $x^r(1-x)^{n-r}$. Since there are $\binom{n}{r}$ ways of choosing the order of r outcomes out of n, the probability of obtaining r outcomes A and $n-r$ outcomes B in *any* order is given by

$$p_{n,r}(x) = \binom{n}{r} x^r (1-x)^{n-r}.$$

Thus we have

$$p_{n,r}(x) = \left(\frac{n-r+1}{r}\right)\left(\frac{x}{1-x}\right) p_{n,r-1}(x),$$

and it follows that

$$p_{n,r}(x) > p_{n,r-1}(x) \quad \text{if and only if} \quad r < (n+1)x.$$

We deduce that $p_{n,r}(x)$, regarded as a function of r, with x and n fixed, has a peak when $r = r_x \approx nx$, for large n, and is monotonically increasing for $r < r_x$ and monotonically decreasing for $r > r_x$. We already know that

$$\sum_{r=0}^{n} p_{n,r}(x) = B_n(1;x) = 1,$$

and in the sum

$$\sum_{r=0}^{n} p_{n,r}(x) f\left(\frac{r}{n}\right) = B_n(f;x),$$

where $f \in C[0,1]$ and n is large, the contributions to the sum from values of r sufficiently remote from r_x will be negligible, and the significant part of the sum will come from values of r close to r_x. Thus, for n large,

$$B_n(f;x) \approx f\left(\frac{r_x}{n}\right) \approx f(x),$$

and so $B_n(f;x) \to f(x)$ as $n \to \infty$. While this is by no means a rigorous argument, and is thus *not* a proof, it gives some insight into how Bernstein was motivated in his search for a proof of the Weierstrass theorem.

Example 7.1.3 To illustrate Bernstein's argument concerning the polynomials $p_{n,r}$, let us evaluate these polynomials when $n = 8$ and $x = 0.4$. The resulting values of $p_{n,r}(x)$ are given in the following table:

r	0	1	2	3	4	5	6	7	8
$p_{n,r}(x)$	0.02	0.09	0.21	0.28	0.23	0.12	0.04	0.01	0.00

In this case, the largest value of $p_{n,r}$ is attained for $r = r_x = 3$, consistent with our above analysis, which shows that $r_x \approx nx = 3.2$. ■

Theorem 7.1.5 Given a function $f \in C[0,1]$ and any $\epsilon > 0$, there exists an integer N such that

$$|f(x) - B_n(f;x)| < \epsilon, \quad 0 \le x \le 1,$$

for all $n \ge N$.

Proof. In other words, the above statement says that the Bernstein polynomials for a function f that is continuous on $[0,1]$ converge uniformly to f on $[0,1]$. The following proof is motivated by the plausible argument that we gave above.

We begin with the identity

$$\left(\frac{r}{n} - x\right)^2 = \left(\frac{r}{n}\right)^2 - 2\left(\frac{r}{n}\right)x + x^2,$$

multiply each term by $\binom{n}{r} x^r(1-x)^{n-r}$, and sum from $r = 0$ to n, to give

$$\sum_{r=0}^{n} \left(\frac{r}{n} - x\right)^2 \binom{n}{r} x^r(1-x)^{n-r} = B_n(x^2;x) - 2xB_n(x;x) + x^2 B_n(1;x).$$

It then follows from (7.3), (7.4), and (7.14) that

$$\sum_{r=0}^{n} \left(\frac{r}{n} - x\right)^2 \binom{n}{r} x^r(1-x)^{n-r} = \frac{1}{n}x(1-x). \tag{7.22}$$

For any fixed $x \in [0,1]$, let us estimate the sum of the polynomials $p_{n,r}(x)$ over all values of r for which r/n is not close to x. To make this notion precise, we choose a number $\delta > 0$ and let S_δ denote the set of all values of r satisfying $\left|\frac{r}{n} - x\right| \ge \delta$. We now consider the sum of the polynomials $p_{n,r}(x)$ over all $r \in S_\delta$. Note that $\left|\frac{r}{n} - x\right| \ge \delta$ implies that

$$\frac{1}{\delta^2}\left(\frac{r}{n} - x\right)^2 \ge 1. \tag{7.23}$$

Then, using (7.23), we have

$$\sum_{r\in S_\delta} \binom{n}{r} x^r(1-x)^{n-r} \le \frac{1}{\delta^2} \sum_{r\in S_\delta} \left(\frac{r}{n}-x\right)^2 \binom{n}{r} x^r(1-x)^{n-r}.$$

The latter sum is not greater that the sum of the same expression over *all* r, and using (7.22), we have

$$\frac{1}{\delta^2} \sum_{r=0}^{n} \left(\frac{r}{n}-x\right)^2 \binom{n}{r} x^r(1-x)^{n-r} = \frac{x(1-x)}{n\delta^2}.$$

Since $0 \le x(1-x) \le \frac{1}{4}$ on $[0,1]$, we have

$$\sum_{r\in S_\delta} \binom{n}{r} x^r(1-x)^{n-r} \le \frac{1}{4n\delta^2}. \tag{7.24}$$

Let us write

$$\sum_{r=0}^{n} = \sum_{r\in S_\delta} + \sum_{r\notin S_\delta},$$

where the latter sum is therefore over all r such that $|\frac{r}{n}-x| < \delta$. Having split the summation into these two parts, which depend on a choice of δ that we still have to make, we are now ready to estimate the difference between $f(x)$ and its Bernstein polynomial. Using (7.3), we have

$$f(x) - B_n(f;x) = \sum_{r=0}^{n} \left(f(x) - f\left(\frac{r}{n}\right)\right) \binom{n}{r} x^r(1-x)^{n-r},$$

and hence

$$f(x) - B_n(f;x) = \sum_{r\in S_\delta} \left(f(x) - f\left(\frac{r}{n}\right)\right) \binom{n}{r} x^r(1-x)^{n-r}$$

$$+ \sum_{r\notin S_\delta} \left(f(x) - f\left(\frac{r}{n}\right)\right) \binom{n}{r} x^r(1-x)^{n-r}.$$

We thus obtain the inequality

$$|f(x) - B_n(f;x)| \le \sum_{r\in S_\delta} \left|f(x) - f\left(\frac{r}{n}\right)\right| \binom{n}{r} x^r(1-x)^{n-r}$$

$$+ \sum_{r\notin S_\delta} \left|f(x) - f\left(\frac{r}{n}\right)\right| \binom{n}{r} x^r(1-x)^{n-r}.$$

Since $f \in C[0,1]$, it is bounded on $[0,1]$, and we have $|f(x)| \leq M$, for some $M > 0$. We can therefore write

$$\left| f(x) - f\left(\frac{r}{n}\right) \right| \leq 2M$$

for all r and all $x \in [0,1]$, and so

$$\sum_{r \in S_\delta} \left| f(x) - f\left(\frac{r}{n}\right) \right| \binom{n}{r} x^r (1-x)^{n-r} \leq 2M \sum_{r \in S_\delta} \binom{n}{r} x^r (1-x)^{n-r}.$$

On using (7.24) we obtain

$$\sum_{r \in S_\delta} \left| f(x) - f\left(\frac{r}{n}\right) \right| \binom{n}{r} x^r (1-x)^{n-r} \leq \frac{M}{2n\delta^2}. \tag{7.25}$$

Since f is continuous, it is also uniformly continuous, on $[0,1]$. Thus, corresponding to any choice of $\epsilon > 0$ there is a number $\delta > 0$, depending on ϵ and f, such that

$$|x - x'| < \delta \;\Rightarrow\; |f(x) - f(x')| < \frac{\epsilon}{2},$$

for all $x, x' \in [0,1]$. Thus, for the sum over $r \notin S_\delta$, we have

$$\begin{aligned} \sum_{r \notin S_\delta} \left| f(x) - f\left(\frac{r}{n}\right) \right| \binom{n}{r} x^r (1-x)^{n-r} &< \frac{\epsilon}{2} \sum_{r \notin S_\delta} \binom{n}{r} x^r (1-x)^{n-r} \\ &< \frac{\epsilon}{2} \sum_{r=0}^{n} \binom{n}{r} x^r (1-x)^{n-r}, \end{aligned}$$

and hence, again using (7.3), we find that

$$\sum_{r \notin S_\delta} \left| f(x) - f\left(\frac{r}{n}\right) \right| \binom{n}{r} x^r (1-x)^{n-r} < \frac{\epsilon}{2}. \tag{7.26}$$

On combining (7.25) and (7.26), we obtain

$$|f(x) - B_n(f;x)| < \frac{M}{2n\delta^2} + \frac{\epsilon}{2}.$$

It follows from the line above that if we choose $N > M/(\epsilon\delta^2)$, then

$$|f(x) - B_n(f;x)| < \epsilon$$

for all $n \geq N$, and this completes the proof. ■

Using the methods employed in the above proof, we can show, with a little greater generality, that if f is merely bounded on $[0,1]$, the sequence $(B_n(f;x))_{n=1}^{\infty}$ converges to $f(x)$ at any point x where f is continuous. We will now discuss some further properties of the Bernstein polynomials.

Theorem 7.1.6 If $f \in C^k[0,1]$, for some integer $k \geq 0$, then $B_n^{(k)}(f;x)$ converges uniformly to $f^{(k)}(x)$ on $[0,1]$.

Proof. We know from Theorem 7.1.5 that the above result holds for $k = 0$. For $k \geq 1$ we begin with the expression for $B_{n+k}^{(k)}(f;x)$ given in (7.16), and write

$$\Delta^k f\left(\frac{r}{n+k}\right) = \frac{f^{(k)}(\xi_r)}{(n+k)^k},$$

where $r/(n+k) < \xi_r < (r+k)/(n+k)$, as we did similarly in (7.21). We then approximate $f^{(k)}(\xi_r)$, writing

$$f^{(k)}(\xi_r) = f^{(k)}\left(\frac{r}{n}\right) + \left(f^{(k)}(\xi_r) - f^{(k)}\left(\frac{r}{n}\right)\right).$$

We thus obtain

$$\frac{n!(n+k)^k}{(n+k)!} B_{n+k}^{(k)}(f;x) = S_1(x) + S_2(x), \tag{7.27}$$

say, where

$$S_1(x) = \sum_{r=0}^{n} f^{(k)}\left(\frac{r}{n}\right) \binom{n}{r} x^r (1-x)^{n-r}$$

and

$$S_2(x) = \sum_{r=0}^{n} \left(f^{(k)}(\xi_r) - f^{(k)}\left(\frac{r}{n}\right)\right) \binom{n}{r} x^r (1-x)^{n-r}.$$

In $S_2(x)$, we can make $|\xi_r - \frac{r}{n}| < \delta$ for all r, for any choice of $\delta > 0$, by taking n sufficiently large. Also, given any $\epsilon > 0$, we can choose a positive value of δ such that

$$\left|f^{(k)}(\xi_r) - f^{(k)}\left(\frac{r}{n}\right)\right| < \epsilon,$$

for all r, because of the uniform continuity of $f^{(k)}$. Thus $S_2(x) \to 0$ uniformly on $[0,1]$ as $n \to \infty$. We can easily verify that

$$\frac{n!(n+k)^k}{(n+k)!} \to 1 \quad \text{as} \quad n \to \infty,$$

and we see from Theorem 7.1.5 with $f^{(k)}$ in place of f that $S_1(x)$ converges uniformly to $f^{(k)}(x)$ on $[0,1]$. This completes the proof. ■

As we have just seen, not only does the Bernstein polynomial for f converge to f, but derivatives converge to derivatives. This is a most remarkable property. In contrast, consider again the sequence of interpolating polynomials (p_n^*) for e^x that appear in Example 2.4.4. Although this sequence of polynomials converges uniformly to e^x on $[-1,1]$, this does *not*

hold for their derivatives, because of the oscillatory nature of the error of interpolation.

On comparing the complexity of the proofs of Theorems 7.1.5 and 7.1.6, it may seem surprising that the *additional* work required to complete the proof of Theorem 7.1.6 for $k \geq 1$ is so little compared to that needed to prove Theorem 7.1.5.

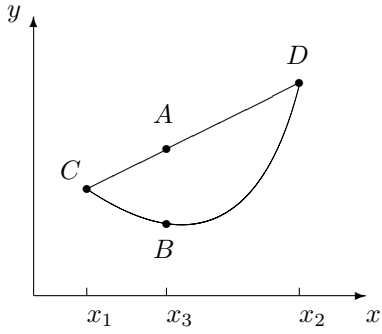

FIGURE 7.3. A and B are the points on the chord CD and on the graph of the convex function $y = f(x)$, respectively, with abscissa $x_3 = \lambda x_1 + (1-\lambda)x_2$.

We now state results concerning the Bernstein polynomials for a convex function f. First we define convexity and show its connection with second-order divided differences.

Definition 7.1.2 A function f is said to be convex on $[a, b]$ if for any $x_1, x_2 \in [a, b]$,

$$\lambda f(x_1) + (1-\lambda) f(x_2) \geq f(\lambda x_1 + (1-\lambda)x_2) \tag{7.28}$$

for any $\lambda \in [0, 1]$. Geometrically, this is just saying that a chord connecting any two points on the convex curve $y = f(x)$ is never below the curve. This is illustrated in Figure 7.3, where CD is such a chord, and the points A and B have y-coordinates $\lambda f(x_1) + (1-\lambda) f(x_2)$ and $f(\lambda x_1 + (1-\lambda)x_2)$, respectively. ■

If f is twice differentiable, f being convex is equivalent to f'' being nonnegative. Of course, functions can be convex without being differentiable. For example, we can have a convex polygonal arc.

Theorem 7.1.7 A function f is convex on $[a, b]$ if and only if all second-order divided differences of f are nonnegative.

Proof. Since a divided difference is unchanged if we alter the order of its arguments, as we see from the symmetric form (1.21), it suffices to consider the divided difference $f[x_0, x_1, x_2]$ where $a \leq x_0 < x_1 < x_2 \leq b$. Then we obtain from the recurrence relation (1.22) that

$$f[x_0, x_1, x_2] \geq 0 \quad \Leftrightarrow \quad f[x_1, x_2] \geq f[x_0, x_1]. \tag{7.29}$$

On multiplying the last inequality throughout by $(x_2 - x_1)(x_1 - x_0)$, which is positive, we find that both inequalities in (7.29) are equivalent to

$$(x_1 - x_0)(f(x_2) - f(x_1)) \geq (x_2 - x_1)(f(x_1) - f(x_0)),$$

which is equivalent to

$$(x_1 - x_0)f(x_2) + (x_2 - x_1)f(x_0) \geq (x_2 - x_0)f(x_1). \tag{7.30}$$

If we now divide throughout by $x_2 - x_0$ and write $\lambda = (x_2 - x_1)/(x_2 - x_0)$, we see that $x_1 = \lambda x_0 + (1 - \lambda)x_2$, and it follows from (7.30) that

$$\lambda f(x_0) + (1 - \lambda)f(x_2) \geq f(\lambda x_0 + (1 - \lambda)x_2),$$

thus completing the proof. ■

The proofs of the following two theorems are held over until Section 7.3, where we will state and prove generalizations of both results.

Theorem 7.1.8 If $f(x)$ is convex on $[0, 1]$, then

$$B_n(f; x) \geq f(x), \quad 0 \leq x \leq 1, \tag{7.31}$$

for all $n \geq 1$. ■

Theorem 7.1.9 If $f(x)$ is convex on $[0, 1]$,

$$B_{n-1}(f; x) \geq B_n(f; x), \quad 0 \leq x \leq 1, \tag{7.32}$$

for all $n \geq 2$. The Bernstein polynomials are equal at $x = 0$ and $x = 1$, since they interpolate f at these points. If $f \in C[0, 1]$, the inequality in (7.32) is *strict* for $0 < x < 1$, for a given value of n, unless f is linear in each of the intervals $\left[\frac{r-1}{n-1}, \frac{r}{n-1}\right]$, for $1 \leq r \leq n - 1$, when we have simply $B_{n-1}(f; x) = B_n(f; x)$. ■

Note that we have from Theorem 7.1.4 with $k = 2$ that if $f''(x) \geq 0$, and thus f is convex on $[0, 1]$, then $B_n(f; x)$ is also convex on $[0, 1]$. In Section 7.3 we will establish the stronger result that $B_n(f; x)$ is convex on $[0, 1]$, provided that f is convex on $[0, 1]$.

We conclude this section by stating two theorems concerned with estimating the error $f(x) - B_n(f; x)$. The first of these is the theorem due to Elizaveta V. Voronovskaya (1898–1972), which gives an asymptotic error term for the Bernstein polynomials for functions that are twice differentiable.

Theorem 7.1.10 Let $f(x)$ be bounded on $[0, 1]$. Then, for any $x \in [0, 1]$ at which $f''(x)$ exists,

$$\lim_{n\to\infty} n(B_n(f; x) - f(x)) = \frac{1}{2}x(1 - x)f''(x). \quad ■ \tag{7.33}$$

See Davis [10] for a proof of Voronovskaya's theorem.

Finally, there is the following result that gives an upper bound for the error $f(x)-B_n(f;x)$ in terms of the modulus of continuity, which we defined in Section 2.6.

Theorem 7.1.11 If f is bounded on $[0,1]$, then

$$\|f - B_n f\| \leq \frac{3}{2}\,\omega\left(\frac{1}{\sqrt{n}}\right), \tag{7.34}$$

where $\|\cdot\|$ denotes the maximum norm on $[0,1]$. ■

See Rivlin [48] for a proof of this theorem.

Example 7.1.4 Consider the Bernstein polynomial for $f(x) = |x - \frac{1}{2}|$,

$$B_n(f;x) = \sum_{r=0}^{n} \left|\frac{r}{n} - \frac{1}{2}\right| \binom{n}{r} x^r (1-x)^{n-r}.$$

The difference between $B_n(f;x)$ and $f(x)$ at $x = \frac{1}{2}$ is

$$\frac{1}{2^n} \sum_{r=0}^{n} \left|\frac{r}{n} - \frac{1}{2}\right| \binom{n}{r} = e_n,$$

say. Let us now choose n to be even. We note that

$$\left(\frac{1}{2} - \frac{r}{n}\right)\binom{n}{r} = \left(\frac{n-r}{n} - \frac{1}{2}\right)\binom{n}{n-r}$$

for all r, and that the quantities on each side of the above equation are zero when $r = n/2$. It follows that

$$2^n e_n = \sum_{r=0}^{n} \left|\frac{r}{n} - \frac{1}{2}\right| \binom{n}{r} = 2\sum_{r=0}^{n/2} \left(\frac{1}{2} - \frac{r}{n}\right)\binom{n}{r}. \tag{7.35}$$

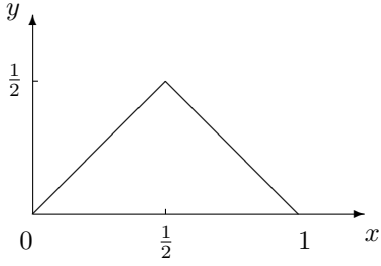

FIGURE 7.4. The function $f(x) = |x - \frac{1}{2}|$ on $[0,1]$.

Let us split the last summation into two. We obtain

$$\sum_{r=0}^{n/2} \binom{n}{r} = \frac{1}{2}\left(\binom{n}{n/2} + \sum_{r=0}^{n} \binom{n}{r} \right) = \frac{1}{2}\binom{n}{n/2} + 2^{n-1},$$

and since

$$\frac{r}{n}\binom{n}{r} = \binom{n-1}{r-1}, \qquad r \geq 1,$$

we find that

$$2\sum_{r=0}^{n/2} \frac{r}{n}\binom{n}{r} = 2\sum_{r=1}^{n/2} \binom{n-1}{r-1} = \sum_{r=1}^{n} \binom{n-1}{r-1} = 2^{n-1}.$$

It then follows from (7.35) that

$$e_n = \frac{1}{2^{n+1}}\binom{n}{n/2} \sim \frac{1}{\sqrt{2\pi n}}$$

for n large. The last step follows on using Stirling's formula for estimating $n!$ (see Problem 2.1.12). This shows that $\|f - B_n f\| \to 0$ at least as slowly as $1/\sqrt{n}$ for the function $f(x) = |x - \frac{1}{2}|$, where $\|\cdot\|$ denotes the maximum norm on $[0,1]$. ■

Problem 7.1.1 Show that

$$B_n(x^3; x) = x^3 + \frac{1}{n^2}x(1-x)(1+(3n-2)x),$$

for all $n \geq 3$.

Problem 7.1.2 Show that

$$B_n(e^{\alpha x}; x) = (xe^{\alpha/n} + (1-x))^n$$

for all integers $n \geq 1$ and all real α.

Problem 7.1.3 Deduce from Definition 7.1.2, using induction on n, that a function f is convex on $[a,b]$ if and only if

$$\sum_{r=0}^{n} \lambda_r f(x_r) \geq f\left(\sum_{r=0}^{n} \lambda_r x_r \right),$$

for all $n \geq 0$, for all $x_r \in [a,b]$, and for all $\lambda_r \geq 0$ such that

$$\lambda_0 + \lambda_1 + \cdots + \lambda_n = 1.$$

Problem 7.1.4 Find a function f and a real number λ such that f is a polynomial of degree two and $B_n f = \lambda f$. Also find a function f and a real number λ such that f is a polynomial of degree three and $B_n f = \lambda f$.

Problem 7.1.5 Verify Voronovskaya's Theorem 7.1.10 directly for the two functions x^2 and x^3.

7.2 The Monotone Operator Theorem

In the 1950s, H. Bohman [5] and P. P. Korovkin [31] obtained an amazing generalization of Bernstein's Theorem 7.1.5. They found that as far as convergence is concerned, the crucial properties of the Bernstein operator B_n are that $B_n f \to f$ uniformly on $[0,1]$ for $f = 1$, x, and x^2, and that B_n is a monotone linear operator. (See Definition 7.1.1.) We now state the Bohman–Korovkin theorem, followed by a proof based on that given by Cheney [7].

Theorem 7.2.1 Let (L_n) denote a sequence of monotone linear operators that map a function $f \in C[a,b]$ to a function $L_n f \in C[a,b]$, and let $L_n f \to f$ uniformly on $[a,b]$ for $f = 1$, x, and x^2. Then $L_n f \to f$ uniformly on $[a,b]$ for all $f \in C[a,b]$.

Proof. Let us define $\phi_t(x) = (t-x)^2$, and consider $(L_n\phi_t)(t)$. Thus we apply the linear operator L_n to ϕ_t, regarded as a function of x, and then evaluate $L_n\phi_t$ (which is also a function of x) at $x = t$. Since L_n is linear, we obtain

$$(L_n\phi_t)(t) = t^2(L_n g_0)(t) - 2t(L_n g_1)(t) + (L_n g_2)(t),$$

where

$$g_0(x) = 1, \qquad g_1(x) = x, \qquad g_2(x) = x^2.$$

Hence

$$(L_n\phi_t)(t) = t^2[(L_n g_0)(t) - 1] - 2t[(L_n g_1)(t) - t] + [(L_n g_2)(t) - t^2].$$

On writing $\|\cdot\|$ to denote the maximum norm on $[a,b]$, we deduce that

$$\|L_n\phi_t\| \le M^2\|L_n g_0 - g_0\| + 2M\|L_n g_1 - g_1\| + \|L_n g_2 - g_2\|,$$

where $M = \max(|a|,|b|)$. Since for $i = 0$, 1, and 2, each term $\|L_n g_i - g_i\|$ may be made as small as we please, by taking n sufficiently large, it follows that

$$(L_n\phi_t)(t) \to 0 \quad \text{as} \quad n \to \infty, \tag{7.36}$$

uniformly in t.

Now let f be any function in $C[a,b]$. Given any $\epsilon > 0$, it follows from the uniform continuity of f that there exists a $\delta > 0$ such that for all $t, x \in [a,b]$,

$$|t-x| < \delta \Rightarrow |f(t) - f(x)| < \epsilon. \tag{7.37}$$

On the other hand, if $|t-x| \ge \delta$, we have

$$|f(t) - f(x)| \le 2\|f\| \le 2\|f\|\frac{(t-x)^2}{\delta^2} = \alpha\phi_t(x), \tag{7.38}$$

say, where $\alpha = 2\|f\|/\delta^2 > 0$. Note that $\phi_t(x) \geq 0$. Then, from (7.37) and (7.38), we see that for *all* $t, x \in [a, b]$,

$$|f(t) - f(x)| \leq \epsilon + \alpha\phi_t(x),$$

and so

$$-\epsilon - \alpha\phi_t(x) \leq f(t) - f(x) \leq \epsilon + \alpha\phi_t(x). \tag{7.39}$$

At this stage we make use of the monotonicity of the linear operator L_n. We apply L_n to each term in (7.39), regarded as a function of x, and evaluate each resulting function of x at the point t, to give

$$\begin{aligned} -\epsilon(L_n g_0)(t) - \alpha(L_n\phi_t)(t) &\leq f(t)(L_n g_0)(t) - (L_n f)(t) \\ &\leq \epsilon(L_n g_0)(t) + \alpha(L_n\phi_t)(t). \end{aligned}$$

Observe that $(L_n\phi_t)(t) \geq 0$, since L_n is monotone and $\phi_t(x) \geq 0$. Thus we obtain the inequality

$$|f(t)(L_n g_0)(t) - (L_n f)(t)| \leq \epsilon\|L_n g_0\| + \alpha(L_n\phi_t)(t). \tag{7.40}$$

If we now write $L_n g_0 = 1 + (L_n g_0 - g_0)$, we obtain

$$\|L_n g_0\| \leq 1 + \|L_n g_0 - g_0\|,$$

and so derive the inequality

$$|f(t)(L_n g_0)(t) - (L_n f)(t)| \leq \epsilon(1 + \|L_n g_0 - g_0\|) + \alpha(L_n\phi_t)(t). \tag{7.41}$$

We now write

$$f(t) - (L_n f)(t) = [f(t)(L_n g_0)(t) - (L_n f)(t)] + [f(t) - f(t)(L_n g_0)(t)],$$

and hence obtain the inequality

$$\begin{aligned} |f(t) - (L_n f)(t)| \leq &|f(t)(L_n g_0)(t) - (L_n f)(t)| \\ &+ |f(t) - f(t)(L_n g_0)(t)|. \end{aligned} \tag{7.42}$$

In (7.41) we have already obtained an upper bound for the first term on the right of (7.42), and the second term satisfies the inequality

$$|f(t) - f(t)(L_n g_0)(t)| \leq \|f\| \, \|L_n g_0 - g_0\|. \tag{7.43}$$

Then, on substituting the two inequalities (7.41) and (7.43) into (7.42), we find that

$$|f(t) - (L_n f)(t)| \leq \epsilon + (\|f\| + \epsilon) \, \|L_n g_0 - g_0\| + \alpha(L_n\phi_t)(t). \tag{7.44}$$

On the right side of the above inequality we have ϵ plus two nonnegative quantities, each of which can be made less than ϵ for all n greater than some sufficiently large number N, and so

$$|f(t) - (L_n f)(t)| < 3\epsilon,$$

uniformly in t, for all $n > N$. This completes the proof. ■

Remark If we go through the above proof again, we can see that the following is a valid alternative version of the statement of Theorem 7.2.1. We will find this helpful when we discuss the Hermite–Fejér operator.

Let (L_n) denote a sequence of monotone linear operators that map functions $f \in C[a,b]$ to functions $L_n f \in C[a,b]$. Then if $L_n g_0 \to g_0$ uniformly on $[0,1]$, and $(L_n \phi_t)(t) \to 0$ uniformly in t on $[0,1]$, where g_0 and ϕ_t are defined in the proof of Theorem 7.2.1, it follows that $L_n f \to f$ uniformly on $[0,1]$ for all $f \in C[a,b]$. ■

Example 7.2.1 We see from Examples 7.1.1 and 7.1.2 that

$$B_n(1;x) = 1, \quad B_n(x;x) = x, \quad \text{and} \quad B_n(x^2;x) = x^2 + \frac{1}{n}x(1-x).$$

Thus $B_n(f;x)$ converges uniformly to $f(x)$ on $[0,1]$ for $f(x) = 1$, x, and x^2, and since the Bernstein operator B_n is also linear and monotone, it follows from the Bohman–Korovkin Theorem 7.2.1 that $B_n(f;x)$ converges uniformly to $f(x)$ on $[0,1]$ for all $f \in C[0,1]$, as we already found in Bernstein's Theorem 7.1.5. ■

We now recall the Hermite interpolating polynomial p_{2n+1}, defined by (1.38). If we write

$$q_{2n+1}(x) = \sum_{i=0}^{n} [a_i u_i(x) + b_i v_i(x)], \tag{7.45}$$

where u_i and v_i are defined in (1.39) and (1.40), then

$$q_{2n+1}(x_i) = a_i, \quad q'_{2n+1}(x_i) = b_i, \quad 0 \le i \le n. \tag{7.46}$$

If we now choose

$$a_i = f(x_i), \quad b_i = 0, \quad 0 \le i \le n, \tag{7.47}$$

where the x_i are the zeros of the Chebyshev polynomial T_{n+1} and f is a given function defined on $[-1,1]$, it follows that

$$q_{2n+1}(x) = \sum_{i=0}^{n} f(x_i) u_i(x) = (L_n f)(x), \tag{7.48}$$

say, where u_i is given by (2.103), and so

$$(L_n f)(x) = \left(\frac{T_{n+1}(x)}{n+1} \right)^2 \sum_{i=0}^{n} f(x_i) \frac{1 - x_i x}{(x - x_i)^2}. \tag{7.49}$$

We call L_n the Hermite–Fejér operator. It is clear that L_n is a linear operator, and since

$$0 \leq 1 - x_i x \leq 2 \quad \text{for} \quad -1 \leq x \leq 1, \tag{7.50}$$

for all i, we see that L_n is monotone. We also note that L_n reproduces the function 1, since the derivative of 1 is zero and $(L_n 1)(x)$ interpolates 1 on the Chebyshev zeros. It is also obvious that L_n does *not* reproduce the functions x and x^2, since their first derivatives are not zero on all the Chebyshev zeros. Let us apply L_n to the function $\phi_t(x) = (t-x)^2$. We obtain

$$(L_n \phi_t)(t) = \left(\frac{T_{n+1}(t)}{n+1} \right)^2 \sum_{i=0}^{n} (1 - x_i t),$$

and it follows from (7.50) that

$$|(L_n \phi_t)(t)| \leq \frac{2}{n+1},$$

so that $(L_n \phi_t)(t) \to 0$ uniformly in t on $[-1, 1]$. Thus, in view of the alternative statement of Theorem 7.2.1, given in the remark following the proof of the theorem, we deduce the following result as a special case of the Bohman–Korovkin Theorem 7.2.1.

Theorem 7.2.2 Let (L_n) denote the sequence of Hermite–Fejér operators, defined by (7.49). Then $L_n f \to f$ uniformly for all $f \in C[-1, 1]$. ■

Theorem 7.2.2, like Bernstein's Theorem 7.1.5, gives a constructive proof of the Weierstrass theorem. A direct proof of Theorem 7.2.2, which does not explicitly use the Bohman–Korovkin theorem, is given in Davis [10]. We will give another application of the Bohman–Korovkin theorem in the next section.

We can show (see Problem 7.2.1) that the only linear monotone operator that reproduces 1, x, and x^2, and thus all quadratic polynomials, is the identity operator. This puts into perspective the behaviour of the Bernstein operator, which reproduces linear polynomials, but does not reproduce x^2, and the Hermite–Fejér operator, which does not reproduce x or x^2.

Problem 7.2.1 Let L denote a linear monotone operator acting on functions $f \in C[a, b]$ that reproduces 1, x, and x^2. Show that $(L_n \phi_t)(t) = 0$, where $\phi_t(x) = (t-x)^2$. By working through the proof of Theorem 7.2.1 show that for a given $f \in C[a, b]$, (7.40) yields

$$|f(t) - (Lf)(t)| \leq \epsilon$$

for all $t \in [a, b]$ and any given $\epsilon > 0$. Deduce that $Lf = f$ for all $f \in C[a, b]$, and thus L is the identity operator.

Problem 7.2.2 Deduce from Theorem 7.2.2 that

$$\lim_{n\to\infty}\left(\frac{T_{n+1}(x)}{n+1}\right)^2\sum_{i=0}^{n}\frac{1-x_ix}{(x-x_i)^2}=1,$$

where the x_i denote the zeros of T_{n+1}.

7.3 On the q-Integers

In view of the many interesting properties of the Bernstein polynomials, it is not surprising that several generalizations have been proposed. In this section we discuss a generalization based on the q-integers, which are defined in Section 1.5. Let us write

$$B_n(f;x)=\sum_{r=0}^{n}f_r\begin{bmatrix} n \\ r \end{bmatrix}x^r\prod_{s=0}^{n-r-1}(1-q^sx) \tag{7.51}$$

for each positive integer n, where f_r denotes the value of the function f at $x=[r]/[n]$, the quotient of the q-integers $[r]$ and $[n]$, and $\begin{bmatrix} n \\ r \end{bmatrix}$ denotes a q-binomial coefficient, defined in (1.116). Note that an empty product in (7.51) denotes 1. When we put $q=1$ in (7.51), we obtain the classical Bernstein polynomial, defined by (7.1), and in this section we consistently write $B_n(f;x)$ to mean the *generalized* Bernstein polynomial, defined by (7.51). In Section 7.5, whenever we need to emphasize the dependence of the generalized Bernstein polynomial on the parameter q, we will write $B_n^q(f;x)$ in place of $B_n(f;x)$.

We see immediately from (7.51) that

$$B_n(f;0)=f(0)\quad\text{and}\quad B_n(f;1)=f(1), \tag{7.52}$$

giving interpolation at the endpoints, as we have for the classical Bernstein polynomials. It is shown in Section 8.2 that every q-binomial coefficient is a polynomial in q (called a Gaussian polynomial) with coefficients that are all positive integers. It is thus clear that B_n, defined by (7.51), is a linear operator and, with $0<q<1$, it is a monotone operator that maps functions defined on $[0,1]$ to P_n. The following theorem involves q-differences, which are defined in Section 1.5. This result yields Theorem 7.1.1 when $q=1$.

Theorem 7.3.1 The generalized Bernstein polynomial may be expressed in the form

$$B_n(f;x)=\sum_{r=0}^{n}\begin{bmatrix} n \\ r \end{bmatrix}\Delta_q^rf_0\,x^r, \tag{7.53}$$

where

$$\Delta_q^rf_j=\Delta_q^{r-1}f_{j+1}-q^{r-1}\Delta_q^{r-1}f_j,\quad r\ge1,$$

with $\Delta_q^0f_j=f_j=f([j]/[n])$.

Proof. Here we require the identity,

$$\prod_{s=0}^{n-r-1}(1-q^s x)=\sum_{s=0}^{n-r}(-1)^s q^{s(s-1)/2}\left[\begin{array}{c} n-r \\ s \end{array}\right] x^s, \tag{7.54}$$

which is equivalent to (8.12) and reduces to a binomial expansion when we put $q=1$. Beginning with (7.51), and expanding the term consisting of the product of the factors $(1-q^s x)$, we obtain

$$B_n(f;x)=\sum_{r=0}^{n} f_r \left[\begin{array}{c} n \\ r \end{array}\right] x^r \sum_{s=0}^{n-r}(-1)^s q^{s(s-1)/2}\left[\begin{array}{c} n-r \\ s \end{array}\right] x^s.$$

Let us put $t=r+s$. Then, since

$$\left[\begin{array}{c} n \\ r \end{array}\right]\left[\begin{array}{c} n-r \\ s \end{array}\right]=\left[\begin{array}{c} n \\ t \end{array}\right]\left[\begin{array}{c} t \\ r \end{array}\right],$$

we may write the latter double sum as

$$\sum_{t=0}^{n}\left[\begin{array}{c} n \\ t \end{array}\right] x^t \sum_{r=0}^{t}(-1)^{t-r} q^{(t-r)(t-r-1)/2}\left[\begin{array}{c} t \\ r \end{array}\right] f_r=\sum_{t=0}^{n}\left[\begin{array}{c} n \\ t \end{array}\right]\Delta_q^t f_0\, x^t,$$

on using the expansion for a higher-order q-difference, as in (1.121). This completes the proof. ■

We see from (1.33) and (1.113) that

$$\frac{\Delta_q^k f(x_0)}{q^{k(k-1)/2}\,[k]!}=f[x_0,x_1,\ldots,x_k]=\frac{f^{(k)}(\xi)}{k!},$$

where $x_j=[j]$ and $\xi\in(x_0,x_k)$. Thus q-differences of the monomial x^k of order greater than k are zero, and we see from Theorem 7.3.1 that for all $n\geq k$, $B_n(x^k;x)$ is a polynomial of degree k.

We deduce from Theorem 7.3.1 that

$$B_n(1;x)=1. \tag{7.55}$$

For $f(x)=x$ we have $\Delta_q^0 f_0=f_0=0$ and $\Delta_q^1 f_0=f_1-f_0=1/[n]$, and it follows from Theorem 7.3.1 that

$$B_n(x;x)=x. \tag{7.56}$$

For $f(x)=x^2$ we have $\Delta_q^0 f_0=f_0=0$, $\Delta_q^1 f_0=f_1-f_0=1/[n]^2$, and

$$\Delta_q^2 f_0=f_2-(1+q)f_1+qf_0=\left(\frac{[2]}{[n]}\right)^2-(1+q)\left(\frac{[1]}{[n]}\right)^2.$$

We then find from Theorem 7.3.1 that

$$B_n(x^2;x) = x^2 + \frac{x(1-x)}{[n]}. \tag{7.57}$$

The above expressions for $B_n(1;x)$, $B_n(x;x)$, and $B_n(x^2;x)$ generalize their counterparts given earlier for the case $q = 1$ and, with the help of Theorem 7.2.1, lead us to the following theorem on the convergence of the generalized Bernstein polynomials.

Theorem 7.3.2 Let (q_n) denote a sequence such that $0 < q_n < 1$ and $q_n \to 1$ as $n \to \infty$. Then, for any $f \in C[0,1]$, $B_n(f;x)$ converges uniformly to $f(x)$ on $[0,1]$, where $B_n(f;x)$ is defined by (7.51) with $q = q_n$.

Proof. We saw above from (7.55) and (7.56) that $B_n(f;x) = f(x)$ for $f(x) = 1$ and $f(x) = x$, and since $q_n \to 1$ as $n \to \infty$, we see from (7.57) that $B_n(f;x)$ converges uniformly to $f(x)$ for $f(x) = x^2$. Also, since $0 < q_n < 1$, it follows that B_n is a monotone operator, and the proof is completed by applying the Bohman–Korovkin Theorem 7.2.1. ■

We now state and prove the following generalizations of Theorems 7.1.8 and 7.1.9.

Theorem 7.3.3 If $f(x)$ is convex on $[0,1]$, then

$$B_n(f;x) \geq f(x), \quad 0 \leq x \leq 1, \tag{7.58}$$

for all $n \geq 1$ and for $0 < q \leq 1$.

Proof. For each $x \in [0,1]$, let us define

$$x_r = \frac{[r]}{[n]} \quad \text{and} \quad \lambda_r = \begin{bmatrix} n \\ r \end{bmatrix} x^r \prod_{s=0}^{n-r+1} (1 - q^s x), \quad 0 \leq r \leq n.$$

We see that $\lambda_r \geq 0$ when $0 < q \leq 1$ and $x \in [0,1]$, and note from (7.55) and (7.56), respectively, that

$$\lambda_0 + \lambda_1 + \cdots + \lambda_n = 1$$

and

$$\lambda_0 x_0 + \lambda_1 x_1 + \cdots + \lambda_n x_n = x.$$

Then we obtain from the result in Problem 7.1.3 that if f is convex on $[0,1]$,

$$B_n(f;x) = \sum_{r=0}^{n} \lambda_r f(x_r) \geq f\left(\sum_{r=0}^{n} \lambda_r x_r\right) = f(x),$$

and this completes the proof. ■

Theorem 7.3.4 If $f(x)$ is convex on $[0,1]$,

$$B_{n-1}(f;x) \geq B_n(f;x), \quad 0 \leq x \leq 1, \tag{7.59}$$

for all $n \geq 2$, where $B_{n-1}(f;x)$ and $B_n(f;x)$ are evaluated using the *same* value of the parameter q. The Bernstein polynomials are equal at $x = 0$ and $x = 1$, since they interpolate f at these points. If $f \in C[0,1]$, the inequality in (7.59) is *strict* for $0 < x < 1$ unless, for a given value of n, the function f is linear in each of the intervals $\left[\frac{[r-1]}{[n-1]}, \frac{[r]}{[n-1]}\right]$, for $1 \leq r \leq n-1$, when we have simply $B_{n-1}(f;x) = B_n(f;x)$.

Proof. In the proof given by Davis [10] for the special case of this theorem when $q = 1$, the difference between two consecutive Bernstein polynomials is expressed in terms of powers of $x/(1-x)$. This is not appropriate for $q \neq 1$, and our proof follows that given by Oruç and Phillips [40]. For $0 < q < 1$, let us write

$$\begin{aligned}
&(B_{n-1}(f;x) - B_n(f;x)) \prod_{s=0}^{n-1} (1-q^s x)^{-1} \\
&\quad = \sum_{r=0}^{n-1} f\left(\frac{[r]}{[n-1]}\right) \begin{bmatrix} n-1 \\ r \end{bmatrix} x^r \prod_{s=n-r-1}^{n-1} (1-q^s x)^{-1} \\
&\quad\quad - \sum_{r=0}^{n} f\left(\frac{[r]}{[n]}\right) \begin{bmatrix} n \\ r \end{bmatrix} x^r \prod_{s=n-r}^{n-1} (1-q^s x)^{-1}.
\end{aligned}$$

We now split the first of the above summations into two, writing

$$x^r \prod_{s=n-r-1}^{n-1} (1-q^s x)^{-1} = \psi_r(x) + q^{n-r-1}\psi_{r+1}(x),$$

say, where

$$\psi_r(x) = x^r \prod_{s=n-r}^{n-1} (1-q^s x)^{-1}. \tag{7.60}$$

On combining the resulting three summations, the terms in $\psi_0(x)$ and $\psi_n(x)$ cancel, and we obtain

$$(B_{n-1}(f;x) - B_n(f;x)) \prod_{s=0}^{n-1} (1-q^s x)^{-1} = \sum_{r=1}^{n-1} \begin{bmatrix} n \\ r \end{bmatrix} a_r \psi_r(x), \tag{7.61}$$

where

$$a_r = \frac{[n-r]}{[n]} f\left(\frac{[r]}{[n-1]}\right) + q^{n-r}\frac{[r]}{[n]} f\left(\frac{[r-1]}{[n-1]}\right) - f\left(\frac{[r]}{[n]}\right). \tag{7.62}$$

It is clear from (7.60) that each $\psi_r(x)$ is nonnegative on $[0,1]$ for $0 \le q \le 1$, and thus from (7.61), it will suffice to show that each a_r is nonnegative. Let us write

$$\lambda = \frac{[n-r]}{[n]}, \quad x_1 = \frac{[r]}{[n-1]}, \quad \text{and} \quad x_2 = \frac{[r-1]}{[n-1]}.$$

It then follows that

$$1 - \lambda = q^{n-r}\frac{[r]}{[n]} \quad \text{and} \quad \lambda x_1 + (1-\lambda)x_2 = \frac{[r]}{[n]},$$

and we see immediately, on comparing (7.62) and (7.28), that

$$a_r = \lambda f(x_1) + (1-\lambda)f(x_2) - f(\lambda x_1 + (1-\lambda)x_2) \ge 0,$$

and so $B_{n-1}(f;x) \ge B_n(f;x)$. We obviously have equality at $x = 0$ and $x = 1$, since all Bernstein polynomials interpolate f at these endpoints. The inequality will be strict for $0 < x < 1$ unless every a_r is zero; this can occur only when f is linear in each of the intervals between consecutive points $[r]/[n-1]$, $0 \le r \le n-1$, when we have $B_{n-1}(f;x) = B_n(f;x)$ for $0 < x < 1$. This completes the proof. ■

We now give an algorithm, first published in 1996 (see Phillips [42]), for evaluating the generalized Bernstein polynomials. When $q = 1$ it reduces to the well-known de Casteljau algorithm (see Hoschek and Lasser [26]) for evaluating the classical Bernstein polynomials.

Algorithm 7.3.1 This algorithm begins with the value of q and the values of f at the $n+1$ points $[r]/[n]$, $0 \le r \le n$, and computes $B_n(f;x) = f_0^{[n]}$, which is the final number generated by the algorithm.

input: q; $f([0]/[n]), f([1]/[n]), \ldots, f([n]/[n])$
 for $r = 0$ **to** n
  $f_r^{[0]} := f([r]/[n])$
 next r
 for $m = 1$ **to** n
  for $r = 0$ **to** $n - m$
   $f_r^{[m]} := (q^r - q^{m-1}x)f_r^{[m-1]} + x f_{r+1}^{[m-1]}$
  next r
 next m
output: $f_0^{[n]} = B_n(f;x)$ ■

The following theorem justifies the above algorithm.

Theorem 7.3.5 For $0 \le m \le n$ and $0 \le r \le n - m$, we have

$$f_r^{[m]} = \sum_{t=0}^{m} f_{r+t} \begin{bmatrix} m \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t-1} (q^r - q^s x), \tag{7.63}$$

and, in particular,

$$f_0^{[n]} = B_n(f;x). \tag{7.64}$$

Proof. We use induction on m. From the initial conditions in the algorithm, $f_r^{[0]} := f([r]/[n]) = f_r,\ 0 \le r \le n$, it is clear that (7.63) holds for $m = 0$ and $0 \le r \le n$. Note that an empty product in (7.63) denotes 1. Let us assume that (7.63) holds for some m such that $0 \le m < n$, and for all r such that $0 \le r \le n - m$. Then, for $0 \le r \le n - m - 1$, it follows from the algorithm that

$$f_r^{[m+1]} = (q^r - q^m x) f_r^{[m]} + x f_{r+1}^{[m]},$$

and using (7.63), we obtain

$$\begin{aligned} f_r^{[m+1]} = (\, q^r - q^m x) \sum_{t=0}^{m} f_{r+t} \begin{bmatrix} m \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t-1} (q^r - q^s x) \\ + \, x \sum_{t=0}^{m} f_{r+t+1} \begin{bmatrix} m \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t-1} (q^{r+1} - q^s x). \end{aligned}$$

The coefficient of f_r on the right of the latter equation is

$$(q^r - q^m x) \prod_{s=0}^{m-1} (q^r - q^s x) = \prod_{s=0}^{m} (q^r - q^s x),$$

and the coefficient of f_{r+m+1} is x^{m+1}. For $1 \le t \le m$, we find that the coefficient of f_{r+t} is

$$\begin{aligned} (q^r - q^m x) \begin{bmatrix} m \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t-1} (q^r - q^s x) + \begin{bmatrix} m \\ t-1 \end{bmatrix} x^t \prod_{s=0}^{m-t} (q^{r+1} - q^s x) \\ = a_t x^t \prod_{s=0}^{m-t-1} (q^r - q^s x), \end{aligned}$$

say. We see that

$$a_t = (q^r - q^m x) \begin{bmatrix} m \\ t \end{bmatrix} + q^{m-t} (q^{r+1} - x) \begin{bmatrix} m \\ t-1 \end{bmatrix}$$

and hence

$$a_t = q^r \left(\begin{bmatrix} m \\ t \end{bmatrix} + q^{m+1-t} \begin{bmatrix} m \\ t-1 \end{bmatrix} \right) - q^{m-t} x \left(q^t \begin{bmatrix} m \\ t \end{bmatrix} + \begin{bmatrix} m \\ t-1 \end{bmatrix} \right).$$

It is easily verified (see (8.7) and (8.8)) that

$$\begin{bmatrix} m \\ t \end{bmatrix} + q^{m+1-t} \begin{bmatrix} m \\ t-1 \end{bmatrix} = q^t \begin{bmatrix} m \\ t \end{bmatrix} + \begin{bmatrix} m \\ t-1 \end{bmatrix} = \begin{bmatrix} m+1 \\ t \end{bmatrix}$$

and thus

$$a_t = (q^r - q^{m-t}x) \begin{bmatrix} m+1 \\ t \end{bmatrix}.$$

Hence the coefficient of f_{r+t}, for $1 \le t \le m$, in the above expression for $f_r^{[m+1]}$ is

$$\begin{bmatrix} m+1 \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t} (q^r - q^s x),$$

and we note that this also holds for $t = 0$ and $t = m+1$. Thus we obtain

$$f_r^{[m+1]} = \sum_{t=0}^{m+1} f_{r+t} \begin{bmatrix} m+1 \\ t \end{bmatrix} x^t \prod_{s=0}^{m-t} (q^r - q^s x),$$

and this completes the proof by induction. ■

The above algorithm for evaluating $B_n(f;x)$ is not unlike Algorithm 1.1.1 (Neville–Aitken). In the latter algorithm, each quantity that is computed is, like the final result, an interpolating polynomial on certain abscissas. Similarly, in Algorithm 7.3.1, as we see in (7.63), each intermediate number $f_r^{[m]}$ has a form that resembles that of the final number $f_0^{[n]} = B_n(f;x)$. We now show that each $f_r^{[m]}$ can also be expressed simply in terms of q-differences, as we have for $B_n(f;x)$ in (7.53).

Theorem 7.3.6 For $0 \le m \le n$ and $0 \le r \le n-m$, we have

$$f_r^{[m]} = \sum_{s=0}^{m} q^{(m-s)r} \begin{bmatrix} m \\ s \end{bmatrix} \Delta_q^s f_r \, x^s. \tag{7.65}$$

Proof. We may verify (7.65) by induction on m, using the recurrence relation in Algorithm 7.3.1. Alternatively, we can derive (7.65) from (7.63) as follows. First we write

$$\prod_{s=0}^{m-t-1} (q^r - q^s x) = q^{(m-t)r} \prod_{s=0}^{m-t-1} (1 - q^s y),$$

where $y = x/q^r$, and we find with the aid of (7.54) that

$$\prod_{s=0}^{m-t-1} (q^r - q^s x) = \sum_{j=0}^{m-t} (-1)^j q^{j(j-1)/2+(m-t-j)r} \begin{bmatrix} m-t \\ j \end{bmatrix} x^j.$$

On substituting this into (7.63), we obtain

$$f_r^{[m]} = \sum_{t=0}^{m} f_{r+t} \begin{bmatrix} m \\ t \end{bmatrix} x^t \sum_{j=0}^{m-t} (-1)^j q^{j(j-1)/2+(m-t-j)r} \begin{bmatrix} m-t \\ j \end{bmatrix} x^j.$$

If we now let $s = j + t$, we may rewrite the above double summation as

$$\sum_{s=0}^{m} q^{(m-s)r} \begin{bmatrix} m \\ s \end{bmatrix} x^s \sum_{j=0}^{s} (-1)^j q^{j(j-1)/2} \begin{bmatrix} s \\ j \end{bmatrix} f_{r+s-j},$$

which, in view of (1.121), gives

$$f_r^{[m]} = \sum_{s=0}^{m} q^{(m-s)r} \begin{bmatrix} m \\ s \end{bmatrix} \Delta_q^s f_r \, x^s,$$

and this completes the proof. ■

Problem 7.3.1 Verify (7.65) directly by induction on m, using the recurrence relation in Algorithm 7.3.1.

Problem 7.3.2 Work through Algorithm 7.3.1 for the case $n = 2$, and so verify directly that $f_0^{[2]} = B_2(f; x)$.

7.4 Total Positivity

We begin this section by defining a totally positive matrix, and discuss the nature of linear transformations when the matrix is totally positive. We will apply these ideas in Section 7.5 to justify further properties of the Bernstein polynomials concerned with shape, such us convexity.

Definition 7.4.1 A real matrix $\mathbf{A}$ is called *totally positive* if all its minors are nonnegative, that is,

$$\mathbf{A} \begin{pmatrix} i_1, i_2, \ldots, i_k \\ j_1, j_2, \ldots, j_k \end{pmatrix} = \det \begin{bmatrix} a_{i_1,j_1} & \cdots & a_{i_1,j_k} \\ \vdots & & \vdots \\ a_{i_k,j_1} & \cdots & a_{i_k,j_k} \end{bmatrix} \geq 0, \tag{7.66}$$

for all $i_1 < i_2 < \cdots < i_k$ and all $j_1 < j_2 < \cdots < j_k$. We say that $\mathbf{A}$ is *strictly totally positive* if all its minors are *positive*, so that $\geq$ is replaced by $>$ in (7.66). ■

It follows, on putting $k = 1$ in (7.66), that a *necessary* condition for a matrix to be totally positive is that all its elements are nonnegative.

Theorem 7.4.1 A real matrix $\mathbf{A} = (a_{ij})$ is totally positive if

$$\mathbf{A} \begin{pmatrix} i, i+1, \ldots, i+k \\ j, j+1, \ldots, j+k \end{pmatrix} \geq 0 \quad \text{for all } i, j, \text{ and } k. \tag{7.67}$$

Similarly, the matrix $\mathbf{A}$ is strictly totally positive if the minors given in (7.67), which are formed from consecutive rows and columns, are all positive. ■

For a proof, see Karlin [28]. In view of Theorem 7.4.1, we can determine whether $\mathbf{A}$ is totally positive or strictly totally positive by testing the positivity of only those minors that are formed from consecutive rows and columns, rather than having to examine *all* minors.

Example 7.4.1 Let us consider the Vandermonde matrix

$$\mathbf{V} = \mathbf{V}(x_0, \ldots, x_n) = \begin{bmatrix} 1 & x_0 & x_0^2 & \cdots & x_0^n \\ 1 & x_1 & x_1^2 & \cdots & x_1^n \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_n & x_n^2 & \cdots & x_n^n \end{bmatrix}. \tag{7.68}$$

As we showed in Chapter 1 (see Problem 1.1.1),

$$\det \mathbf{V}(x_0, \ldots, x_n) = \prod_{i>j} (x_i - x_j). \tag{7.69}$$

Let $0 < x_0 < x_1 < \cdots < x_n$. Then we see from (7.69) that $\det \mathbf{V} > 0$, and we now prove that $\mathbf{V}$ is strictly totally positive. Using Theorem 7.4.1, it is sufficient to show that the minors

$$\det \begin{bmatrix} x_i^j & x_i^{j+1} & \cdots & x_i^{j+k} \\ x_{i+1}^j & x_{i+1}^{j+1} & \cdots & x_{i+1}^{j+k} \\ \vdots & \vdots & & \vdots \\ x_{i+k}^j & x_{i+k}^{j+1} & \cdots & x_{i+k}^{j+k} \end{bmatrix}$$

are positive for all nonnegative i, j, k such that $i+k, j+k \le n$. On removing common factors from its rows, the above determinant may be expressed as

$$(x_i \cdots x_{i+k})^j \det \mathbf{V}(x_i, \ldots, x_{i+k}) > 0,$$

since

$$\det \mathbf{V}(x_i, \ldots, x_{i+k}) = \prod_{i \le r < s \le i+k} (x_s - x_r) > 0.$$

This completes the proof that $\mathbf{V}$ is strictly totally positive. ■

If $\mathbf{A} = \mathbf{BC}$, where $\mathbf{A}$, $\mathbf{B}$, and $\mathbf{C}$ denote matrices of orders $m \times n, m \times k$, and $k \times n$, respectively, then

$$\mathbf{A} \begin{pmatrix} i_1, \ldots, i_p \\ j_1, \ldots, j_p \end{pmatrix} = \sum_{\beta_1 < \cdots < \beta_p} \cdots \sum \mathbf{B} \begin{pmatrix} i_1, \ldots, i_p \\ \beta_1, \ldots, \beta_p \end{pmatrix} \mathbf{C} \begin{pmatrix} \beta_1, \ldots, \beta_p \\ j_1, \cdots, j_p \end{pmatrix}. \tag{7.70}$$

This is known as the Cauchy–Binet determinant identity. It follows immediately from this most useful identity that the product of totally positive matrices is a totally positive matrix, and the product of strictly totally positive matrices is a strictly totally positive matrix.

Definition 7.4.2 Let $\mathbf{v}$ denote the sequence (v_i), which may be finite or infinite. Then we denote by $S^-(\mathbf{v})$ the number of strict sign changes in the sequence $\mathbf{v}$. ■

For example, $S^-(1,-2,3,-4,5,-6) = 5$, $S^-(1,0,0,1,-1) = 1$, and $S^-(1,-1,1,-1,1,-1,\ldots) = \infty$, where the last sequence alternates $+1$ and -1 indefinitely. It is clear that inserting zeros into a sequence, or deleting zeros from a sequence, does not alter the number of changes of sign. Also, deleting terms of a sequence does not increase the number of changes of sign. We use the same notation to denote sign changes in a function.

Definition 7.4.3 Let

$$v_i = \sum_{k=0}^{n} a_{ik} u_k, \quad i = 0, 1, \ldots, m,$$

where the u_k and the a_{ik}, and thus the v_i, are all real. This linear transformation is said to be *variation-diminishing* if

$$S^-(\mathbf{v}) \leq S^-(\mathbf{u}). \qquad ■$$

Definition 7.4.4 A matrix $\mathbf{A}$, which may be finite or infinite, is said to be *m-banded* if there exists an integer l such that $a_{ij} \neq 0$ implies that $l \leq j - i \leq l + m$. ■

This is equivalent to saying that all the nonzero elements of $\mathbf{A}$ lie on $m+1$ diagonals. We will say that a matrix $\mathbf{A}$ is *banded* if it is m-banded for some m. Note that every finite matrix is banded. We have already met 1-banded and 2-banded (tridiagonal) matrices in Chapter 6. In this section we will be particularly interested in 1-banded matrices, also called *bidiagonal* matrices, because of Theorem 7.4.3 below.

We now come to the first of the main results of this section.

Theorem 7.4.2 If $\mathbf{T}$ is a totally positive banded matrix and $\mathbf{v}$ is any vector for which $\mathbf{Tv}$ is defined, then

$$S^-(\mathbf{Tv}) \leq S^-(\mathbf{v}). \qquad ■$$

For a proof of this theorem see Goodman [22].

When we first encounter it, the question of whether a linear transformation is variation-diminishing may not seem very interesting. However, building on the concept of a variation-diminishing linear transformation, we will see in Section 7.5 that the number of sign changes in a function f defined on $[0,1]$ is not increased if we apply a Bernstein operator, and we say that Bernstein operators are shape-preserving. This property does not always hold, for example, for interpolating operators.

Example 7.4.2 Let $\mathbf{v}$ denote an infinite real sequence for which $S^-(\mathbf{v})$ is finite. Consider the sequence $\mathbf{w} = (w_i)_{i=0}^{\infty}$ defined by

$$w_i = v_i + v_{i-1} \quad \text{for} \quad i \geq 1, \quad \text{and} \quad w_0 = v_0.$$

Then $\mathbf{w} = \mathbf{T}\mathbf{v}$, where

$$\mathbf{T} = \begin{bmatrix} 1 & 0 & 0 & 0 & \cdots \\ 1 & 1 & 0 & 0 & \cdots \\ 0 & 1 & 1 & 0 & \cdots \\ 0 & 0 & 1 & 1 & \cdots \\ \vdots & \vdots & \vdots & \vdots & \vdots \end{bmatrix}.$$

Let us consider the minors of $\mathbf{T}$ constructed from consecutive rows and columns. Any such minor whose leading (top left) element is 0 has either a whole row or a whole column of zeros, and so the minor is zero. It is also not hard to see that any minor constructed from consecutive rows and columns whose leading element is 1 has itself the value 1. Thus, by Theorem 7.4.1, the matrix $\mathbf{T}$ is totally positive, and so we deduce from Theorem 7.4.2 that $S^-(\mathbf{w}) = S^-(\mathbf{T}\mathbf{v}) \leq S^-(\mathbf{v})$. ■

Theorem 7.4.3 A finite matrix is totally positive if and only if it is a product of 1-banded matrices with nonnegative elements. ■

For a proof of this theorem, see de Boor and Pinkus [13]. An immediate consequence of Theorem 7.4.3 is that the product of totally positive matrices is totally positive, as we have already deduced above from the Cauchy–Binet identity.

Example 7.4.3 To illustrate Theorem 7.4.3, consider the 1-banded factorization

$$\begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix} = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 0 \\ 0 & 1 & 2 \\ 0 & 0 & 2 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 1 \\ 0 & 0 & 1 \end{bmatrix}.$$

The four matrices in the above product are indeed all 1-banded matrices with nonnegative elements, and their product is totally positive. ■

We now state a theorem, and give a related example, concerning the factorization of a matrix into the product of lower and upper triangular matrices.

Theorem 7.4.4 A matrix $\mathbf{A}$ is strictly totally positive if and only if it can be expressed in the form $\mathbf{A} = \mathbf{L}\mathbf{U}$ where $\mathbf{L}$ is a lower triangular matrix, $\mathbf{U}$ is an upper triangular matrix, and both $\mathbf{L}$ and $\mathbf{U}$ are totally positive matrices. ■

For a proof, see Cryer [9].

Example 7.4.4 To illustrate Theorem 7.4.4, we continue Example 7.4.3, in which the matrix

$$\mathbf{A} = \begin{bmatrix} 1 & 1 & 1 \\ 1 & 2 & 4 \\ 1 & 3 & 9 \end{bmatrix}$$

is expressed as a product of four 1-banded matrices. If we multiply the first two of these 1-banded matrices, and also multiply the third and fourth, we obtain the **LU** factorization

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 1 & 2 & 1 \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 0 & 1 & 3 \\ 0 & 0 & 2 \end{bmatrix} = \mathbf{LU},$$

and it is easy to verify that **L** and **U** are both totally positive. ■

The matrix **A** in Example 7.4.4 is the 3×3 Vandermonde matrix $\mathbf{V}(1, 2, 3)$. In Section 1.2 we gave (see Theorem 1.2.3) the **LU** factorization of the general Vandermonde matrix.

Example 7.4.5 Let

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 1 \\ 1 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}.$$

Then we can easily verify that **A** is totally positive, and it is obviously not strictly totally positive. We give two different **LU** factorizations of **A**:

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where **L** is totally positive but **U** is not, and

$$\mathbf{A} = \mathbf{LU} = \begin{bmatrix} 1 & 0 & 0 \\ 1 & 1 & 0 \\ 0 & 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix},$$

where both **L** and **U** are totally positive. This example shows that we cannot replace "strictly totally positive" by "totally positive" in the statement of Theorem 7.4.4. ■

Definition 7.4.5 For a real-valued function f on an interval I, we define $S^-(f)$ to be the number of sign changes of f, that is,

$$S^-(f) = \sup S^-(f(x_0), \ldots, f(x_m)),$$

where the supremum is taken over all increasing sequences $(x_0, \ldots, x_m)$ in I, for all m.

Definition 7.4.6 We say that a sequence $(\phi_0, \ldots, \phi_n)$ of real-valued functions on an interval I is totally positive if for any points $0 < x_0 < \cdots < x_n$ in I, the collocation matrix $(\phi_j(x_i))_{i,j=0}^n$ is totally positive. ■

Theorem 7.4.5 Let $\psi_i(x) = \omega(x)\phi_i(x)$, for $0 \le i \le n$. Then, if $\omega(x) \ge 0$ on I and the sequence of functions $(\phi_0, \ldots, \phi_n)$ is totally positive on I, the sequence $(\psi_0, \ldots, \psi_n)$ is also totally positive on I.

Proof. This follows easily from the definitions. ■

Theorem 7.4.6 If $(\phi_0, \ldots, \phi_n)$ is totally positive on I, then for any numbers $a_0, \ldots, a_n$,

$$S^-(a_0\phi_n + \cdots + a_n\phi_n) \le S^-(a_0, \ldots, a_n). \quad ■$$

For a proof of this theorem see Goodman [22].

Definition 7.4.7 Let L denote a linear operator that maps each function f defined on an interval $[0,1]$ onto Lf defined on $[0,1]$. Then we say that L is variation-diminishing if

$$S^-(Lf) \le S^-(f). \quad ■$$

Problem 7.4.1 Show that an $n \times n$ matrix has $(2^n - 1)^2$ minors, of which $\frac{1}{4}n^2(n+1)^2$ are formed from consecutive rows and columns. How many minors are there in these two categories for an $m \times n$ matrix?

Problem 7.4.2 Show that the matrix

$$\begin{bmatrix} a_0 & a_1 & a_2 & a_3 & a_4 \\ 0 & a_0 & a_1 & a_2 & a_3 \\ 0 & 0 & a_0 & a_1 & a_2 \\ 0 & 0 & 0 & a_0 & a_1 \\ 0 & 0 & 0 & 0 & a_0 \end{bmatrix}$$

is totally positive, where $a_i \ge 0$ for all i, and $a_i^2 - a_{i-1}a_{i+1} \ge 0$ for $1 \le i \le 3$.

Problem 7.4.3 Let $\mathbf{v}$ denote an infinite real sequence for which $S^-(\mathbf{v})$ is finite. Consider the sequence $\mathbf{w}$ defined by

$$w_r = \sum_{s=0}^{r} v_s \quad \text{for} \quad r \ge 0.$$

Show that $S^-(\mathbf{w}) \le S^-(\mathbf{v})$.

Problem 7.4.4 Repeat Problem 7.4.3 with the sequence $\mathbf{w}$ defined by

$$w_r = \sum_{s=0}^{r} \binom{r}{s} v_s \quad \text{for} \quad r \geq 0,$$

again showing that $S^-(\mathbf{w}) \leq S^-(\mathbf{v})$.

Problem 7.4.5 Show that the matrix

$$\begin{bmatrix} (x_0+t_0)^{-1} & (x_0+t_1)^{-1} & 1 & x_0 \\ (x_1+t_0)^{-1} & (x_1+t_1)^{-1} & 1 & x_1 \\ (x_2+t_0)^{-1} & (x_2+t_1)^{-1} & 1 & x_2 \\ (x_3+t_0)^{-1} & (x_3+t_1)^{-1} & 1 & x_3 \end{bmatrix}$$

is totally positive if $0 < x_0 < x_1 < x_2 < x_3$ and $0 < t_0 < t_1$.

7.5 Further Results

This section is based on the work of Goodman, Oruç and Phillips [23]. We will use the theory of total positivity, developed in the last section, to justify shape-preserving properties of the generalized Bernstein polynomials. We will also show that if a function f is convex on $[0,1]$, then for each x in $[0,1]$ the generalized Bernstein polynomial $B_n(f;x)$ approaches $f(x)$ monotonically from above as the parameter q increases, for $0 < q \leq 1$.

In the last section we saw that for $0 < x_0 < x_1 < \cdots < x_n$, the Vandermonde matrix $\mathbf{V}(x_0, \ldots, x_n)$ is strictly totally positive. It then follows from Definition 7.4.6 that the sequence of monomials $(x^i)_{i=0}^n$ is totally positive on any interval $[0, \infty)$. We now make the change of variable $t = x/(1-x)$, and note that t is an increasing function of x. Thus, if $t_i = x_i/(1-x_i)$, and we now let $0 < x_0 < x_1 < \cdots < x_n < 1$, it follows that $0 < t_0 < t_1 < \cdots < t_n$.

Since the Vandermonde matrix $\mathbf{V}(t_0, \ldots, t_n)$ is strictly totally positive, it follows that the sequence of functions

$$\left(1, \frac{x}{1-x}, \frac{x^2}{(1-x)^2}, \ldots, \frac{x^n}{(1-x)^n}\right)$$

is totally positive on $[0,1]$. We also see from Theorem 7.4.5 that the sequence of functions

$$\left((1-x)^n, x(1-x)^{n-1}, x^2(1-x)^{n-2}, \ldots, x^{n-1}(1-x), x^n\right) \tag{7.71}$$

is totally positive on $[0,1]$. Since the $n+1$ functions in the sequence (7.71) are a basis for P_n, the subspace of polynomials of degree at most n, they are a basis for all the classical Bernstein polynomials of degree n, defined by (7.1), and we can immediately deduce the following powerful result from Theorem 7.4.6.

Theorem 7.5.1 Let $B_n(f;x)$ denote the classical Bernstein polynomial of degree n for the function f. Then

$$S^-(B_n f) \le S^-(f) \tag{7.72}$$

for all f defined on $[0,1]$, and thus the classical Bernstein operator B_n is variation-diminishing.

Proof. Using Theorem 7.4.6, we have

$$S^-(B_n f) \le S^-(f_0, f_1, \ldots, f_n) \le S^-(f),$$

where $f_r = f(r/n)$. ■

For each q such that $0 < q \le 1$, and each $n \ge 1$, we now define

$$P^q_{n,j}(x) = x^j \prod_{s=0}^{n-j-1} (1 - q^s x), \quad 0 \le x \le 1, \tag{7.73}$$

for $0 \le j \le n$. These functions are a basis for P_n, and are thus a basis for all the generalized Bernstein polynomials of degree n, defined by (7.51). We have already seen above that $(P^1_{n,0}, P^1_{n,1}, \ldots, P^1_{n,n})$ is totally positive on $[0,1]$, and we will show below that the same holds for $(P^q_{n,0}, P^q_{n,1}, \ldots, P^q_{n,n})$, for any q such that $0 < q \le 1$.

Since the functions defined in (7.73) are a basis for P_n, it follows that for any choice of q and r satisfying $0 < q, r \le 1$, there exists a nonsingular matrix $\mathbf{T}^{n,q,r}$ such that

$$\begin{bmatrix} P^q_{n,0}(x) \\ \vdots \\ P^q_{n,n}(x) \end{bmatrix} = \mathbf{T}^{n,q,r} \begin{bmatrix} P^r_{n,0}(x) \\ \vdots \\ P^r_{n,n}(x) \end{bmatrix}. \tag{7.74}$$

Theorem 7.5.2 For $0 < q \le r \le 1$ all elements of the matrix $\mathbf{T}^{n,q,r}$ are nonnegative.

Proof. We use induction on n. Since $\mathbf{T}^{1,q,r}$ is the 2×2 identity matrix, its elements are obviously nonnegative. Let us assume that the elements of $\mathbf{T}^{n,q,r}$ are all nonnegative for some $n \ge 1$. Then, since

$$P^q_{n+1,j+1}(x) = x P^q_{n,j}(x), \quad 0 \le j \le n, \tag{7.75}$$

for all q such that $0 < q \le 1$, we see from (7.75) and (7.74) that

$$\begin{bmatrix} P^q_{n+1,1}(x) \\ \vdots \\ P^q_{n+1,n+1}(x) \end{bmatrix} = \mathbf{T}^{n,q,r} \begin{bmatrix} P^r_{n+1,1}(x) \\ \vdots \\ P^r_{n+1,n+1}(x) \end{bmatrix}. \tag{7.76}$$

Also, we have

$$P^q_{n+1,0}(x) = (1-q^n x)P^q_{n,0}(x) = (1-q^n x)\sum_{j=0}^{n} t^{n,q,r}_{0,j} P^r_{n,j}(x), \qquad (7.77)$$

where $(t^{n,q,r}_{0,0}, t^{n,q,r}_{0,1}, \ldots, t^{n,q,r}_{0,n})$ denotes the first row of the matrix $\mathbf{T}^{n,q,r}$. If we now substitute

$$(1-q^n x)P^r_{n,j}(x) = P^r_{n+1,j}(x) + (r^{n-j} - q^n)P^r_{n+1,j+1}(x)$$

in the right side of (7.77) and simplify, we obtain

$$\begin{aligned} P^q_{n+1,0}(x) = {} & t^{n,q,r}_{0,0} P^r_{n+1,0}(x) + (1-q^n)t^{n,q,r}_{0,n} P^r_{n+1,n+1}(x) \\ & + \sum_{j=1}^{n} \left((r^{n+1-j} - q^n)t^{n,q,r}_{0,j-1} + t^{n,q,r}_{0,j}\right) P^r_{n+1,j}(x). \end{aligned} \qquad (7.78)$$

Then, on combining (7.76) and (7.78), we find that

$$\begin{bmatrix} P^q_{n+1,0}(x) \\ P^q_{n+1,1}(x) \\ \vdots \\ P^q_{n+1,n+1}(x) \end{bmatrix} = \begin{bmatrix} t^{n,q,r}_{0,0} & \mathbf{v}^T_{n+1} \\ & \\ \mathbf{0} & \mathbf{T}^{n,q,r} \end{bmatrix} \begin{bmatrix} P^r_{n+1,0}(x) \\ P^r_{n+1,1}(x) \\ \vdots \\ P^r_{n+1,n+1}(x) \end{bmatrix}, \qquad (7.79)$$

so that

$$\mathbf{T}^{n+1,q,r} = \begin{bmatrix} t^{n,q,r}_{0,0} & \mathbf{v}^T_{n+1} \\ & \\ \mathbf{0} & \mathbf{T}^{n,q,r} \end{bmatrix}. \qquad (7.80)$$

In the block matrix on the right side of (7.80) $\mathbf{0}$ denotes the zero vector with $n+1$ elements, and $\mathbf{v}^T_{n+1}$ is the row vector whose elements are the coefficients of $P^r_{n+1,1}(x), \ldots, P^r_{n+1,n+1}(x)$, given by (7.78). On substituting $x = 0$ in (7.80), it is clear that $t^{n,q,r}_{0,0} = 1$. We can deduce from (7.78) that if $0 < q \le r \le 1$, the elements of $\mathbf{v}^T_{n+1}$ are nonnegative, and this completes the proof by induction. ■

It follows from (7.80) and the definition of $\mathbf{v}^T_{n+1}$ that

$$t^{n+1,q,r}_{0,0} = t^{n,q,r}_{0,0} \qquad (7.81)$$

and

$$t^{n+1,q,r}_{0,j} = (r^{n+1-j} - q^n)t^{n,q,r}_{0,j-1} + t^{n,q,r}_{0,j}, \quad 1 \le j \le n. \qquad (7.82)$$

We will require this recurrence relation, which expresses the elements in the first row of $\mathbf{T}^{n+1,q,r}$ in terms of those in the first row of $\mathbf{T}^{n,q,r}$, in the proof of our next theorem. This shows that the transformation matrix $\mathbf{T}^{n,q,r}$ can be factorized as a product of 1-banded matrices. First we require the following lemma.

Lemma 7.5.1 For $m \geq 1$ and any real r and a, let $\mathbf{A}(m,a)$ denote the $m \times (m+1)$ matrix

$$\begin{bmatrix} 1 & r^m - a & & & & \\ & 1 & r^{m-1} - a & & & \\ & & \cdot & \cdot & & \\ & & & \cdot & \cdot & \\ & & & & 1 & r-a \end{bmatrix}.$$

Then

$$\mathbf{A}(m,a)\mathbf{A}(m+1,b) = \mathbf{A}(m,b)\mathbf{A}(m+1,a). \tag{7.83}$$

Proof. For $i = 0, \ldots, m-1$ the ith row of each side of (7.83) is

$$\left[0, \ldots, 0, 1, r^{m+1-i} + r^{m-i} - a - b, (r^{m-i} - a)(r^{m-i} - b), 0, \ldots, 0\right]. \quad \blacksquare$$

For $1 \leq j \leq n-1$, let $\mathbf{B}_j^{(n)}$ denote the 1-banded $(n+1) \times (n+1)$ matrix that has units on the main diagonal, and has the elements

$$r^j - q^{n-j}, r^{j-1} - q^{n-j}, \ldots, r - q^{n-j}, 0, \ldots, 0$$

on the diagonal above the main diagonal, where there are $n-j$ zeros at the lower end of that diagonal. Thus, for example,

$$\mathbf{B}_2^{(n)} = \begin{bmatrix} 1 & r^2 - q^{n-2} & & & & & \\ & 1 & r - q^{n-2} & & & & \\ & & 1 & & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & & 1 & \end{bmatrix}.$$

The matrix $\mathbf{B}_j^{(n+1)}$ can be expressed in a block form involving the matrix $\mathbf{B}_j^{(n)}$. We can verify that

$$\mathbf{B}_1^{(n+1)} = \begin{bmatrix} 1 & \mathbf{c}_0^T \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \tag{7.84}$$

and

$$\mathbf{B}_{j+1}^{(n+1)} = \begin{bmatrix} 1 & \mathbf{c}_j^T \\ \mathbf{0} & \mathbf{B}_j^{(n)} \end{bmatrix} \tag{7.85}$$

for $1 \leq j \leq n-1$, where each $\mathbf{c}_j^T$ is a row vector, $\mathbf{0}$ denotes the zero vector, and $\mathbf{I}$ is the unit matrix of order $n+1$.

Theorem 7.5.3 For $n \geq 2$ and any q, r, we have

$$\mathbf{T}^{n,q,r} = \mathbf{B}_1^{(n)} \mathbf{B}_2^{(n)} \cdots \mathbf{B}_{n-1}^{(n)}, \tag{7.86}$$

where $\mathbf{B}_j^{(n)}$ is the 1-banded matrix defined above.

Proof. We use induction on n. For all $n \geq 2$ let

$$\mathbf{S}^{n,q,r} = \mathbf{B}_1^{(n)} \mathbf{B}_2^{(n)} \cdots \mathbf{B}_{n-1}^{(n)}. \tag{7.87}$$

It is easily verified that

$$\mathbf{T}^{2,q,r} = \mathbf{S}^{2,q,r} = \mathbf{B}_1^{(2)} = \begin{bmatrix} 1 & r-q & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}.$$

Let us assume that for some $n \geq 2$, $\mathbf{T}^{n,q,r} = \mathbf{S}^{n,q,r}$. It follows from (7.84) and (7.85) that

$$\mathbf{S}^{n+1,q,r} = \begin{bmatrix} 1 & \mathbf{c}_0^T \\ \mathbf{0} & \mathbf{I} \end{bmatrix} \begin{bmatrix} 1 & \mathbf{c}_1^T \\ \mathbf{0} & \mathbf{B}_1^{(n)} \end{bmatrix} \cdots \begin{bmatrix} 1 & \mathbf{c}_{n-1}^T \\ \mathbf{0} & \mathbf{B}_{n-1}^{(n)} \end{bmatrix}. \tag{7.88}$$

If we carry out the multiplication of the n block matrices on the right of (7.88), then, using (7.86), we see that

$$\mathbf{S}^{n+1,q,r} = \begin{bmatrix} 1 & \mathbf{d}^T \\ \mathbf{0} & \mathbf{T}^{n,q,r} \end{bmatrix},$$

where $\mathbf{d}^T$ is a row vector. Thus it remains only to verify that the first rows of $\mathbf{T}^{n+1,q,r}$ and $\mathbf{S}^{n+1,q,r}$ are equal. Let us denote the first row of $\mathbf{S}^{n,q,r}$ by

$$\left[s_{0,0}^{n,q,r}, s_{0,1}^{n,q,r}, \ldots, s_{0,n}^{n,q,r} \right].$$

We will show that $s_{0,n}^{n,q,r} = 0$. Let us examine the product of the $n-1$ matrices on the right of (7.87). We can show by induction on j that for $1 \leq j \leq n-1$, the product $\mathbf{B}_1^{(n)} \mathbf{B}_2^{(n)} \cdots \mathbf{B}_j^{(n)}$ is j-banded, where the nonzero elements are on the main diagonal and the j diagonals above the main diagonal. (See Problem 7.5.2.) Thus $\mathbf{S}^{n,q,r}$ is $(n-1)$-banded, and so the last element in its first row, $s_{0,n}^{n,q,r}$, is zero.

Now let us write the matrix $\mathbf{B}_j^{(n+1)}$ in the block form

$$\mathbf{B}_j^{(n+1)} = \begin{bmatrix} \mathbf{A}(j, q^{n+1-j}) & \mathbf{O} \\ \mathbf{C}_j & \mathbf{D}_j \end{bmatrix},$$

where $\mathbf{A}(j, q^{n+1-j})$ is the $j \times (j+1)$ matrix defined in Lemma 7.5.1, $\mathbf{O}$ is the $j \times (n+1-j)$ zero matrix, $\mathbf{C}_j$ is $(n+2-j) \times (j+1)$, and $\mathbf{D}_j$ is $(n+2-j) \times (n+1-j)$. Thus

$$\mathbf{B}_1^{(n+1)}\mathbf{B}_2^{(n+1)} \cdots \mathbf{B}_j^{(n+1)} = \begin{bmatrix} \mathbf{A}(1, q^n) \cdots \mathbf{A}(j, q^{n+1-j}) & \mathbf{0}^T \\ \mathbf{F}_j & \mathbf{G}_j \end{bmatrix}, \quad (7.89)$$

where $\mathbf{A}(1, q^n) \cdots \mathbf{A}(j, q^{n+1-j})$ is $1 \times (j+1)$ and $\mathbf{0}^T$ is the zero vector with $n+1-j$ elements. In particular, on putting $j = n$ in (7.89), we see from (7.87) that the first row of $\mathbf{S}^{n+1,q,r}$ is

$$\left[\mathbf{A}(1, q^n)\mathbf{A}(2, q^{n-1}) \cdots \mathbf{A}(n-1, q^2)\mathbf{A}(n, q), 0\right] = \left[\mathbf{w}^T, 0\right], \quad (7.90)$$

say, where $\mathbf{w}^T$ is a row vector with $n+1$ elements. (We note in passing that this confirms our earlier observation that the last element of the first row of $\mathbf{S}^{n+1,q,r}$ is zero.) In view of Lemma 7.5.1, we may permute the quantities $q^n, q^{n-1}, \ldots, q$ in (7.90), leaving $\mathbf{w}^T$ unchanged. In particular, we may write

$$\mathbf{w}^T = \mathbf{A}(1, q^{n-1})\mathbf{A}(2, q^{n-2}) \cdots \mathbf{A}(n-1, q)\mathbf{A}(n, q^n). \quad (7.91)$$

Now, by comparison with (7.90), the product of the first $n-1$ matrices in (7.91) is the row vector containing the first n elements in the first row of $\mathbf{S}^{n,q,r}$, and thus

$$\begin{aligned} \mathbf{w}^T &= [s_{0,0}^{n,q,r}, \ldots, s_{0,n-1}^{n,q,r}] \begin{bmatrix} 1 & r^n - q^n & & \\ & \ddots & \ddots & \\ & & 1 & r - q^n \end{bmatrix} \\ &= [t_{0,0}^{n,q,r}, \ldots, t_{0,n-1}^{n,q,r}] \begin{bmatrix} 1 & r^n - q^n & & \\ & \ddots & \ddots & \\ & & 1 & r - q^n \end{bmatrix}. \end{aligned}$$

This gives

$$s_{0,0}^{n+1,q,r} = t_{0,0}^{n,q,r}$$

and

$$s_{0,j}^{n+1,q,r} = (r^{n+1-j} - q^n)t_{0,j-1}^{n,q,r} + t_{0,j}^{n,q,r}, \quad 1 \le j \le n,$$

where we note that $t_{0,n}^{n,q,r} = 0$. It then follows from (7.81) and (7.82) that

$$s_{0,j}^{n+1,q,r} = t_{0,j}^{n+1,q,r}, \quad 0 \le j \le n,$$

and since $s_{0,n+1}^{n+1,q,r} = 0 = t_{0,n+1}^{n+1,q,r}$, (7.86) holds for $n+1$. This completes the proof. ■

If $0 < q \le r^{n-1} \le 1$, all elements in the 1-banded matrices $\mathbf{B}_j^{(n)}$ on the right of (7.86) are nonnegative. Then, from Theorem 7.4.3, we immediately have the following result concerning the total positivity of $\mathbf{T}^{n,q,r}$.

Theorem 7.5.4 If $0 < q \le r^{n-1} \le 1$, the transformation matrix $\mathbf{T}^{n,q,r}$ is totally positive. ■

Theorem 7.5.4 has the following important consequence for the generalized Bernstein polynomials.

Theorem 7.5.5 For $0 < q \le 1$, the set of functions $(P^q_{n,0}, \ldots, P^q_{n,n})$, which are a basis for all generalized Bernstein polynomials of degree n, is totally positive on $[0,1]$.

Proof. Let $\mathbf{A}^q_n$ denote the collocation matrix $(P^q_{n,j}(x_i))^n_{i,j=0}$, where we have $0 \le x_0 < \cdots < x_n \le 1$. Then we see from (7.74) that

$$\mathbf{A}^q_n = \mathbf{T}^{n,1,q}\mathbf{A}^1_n. \tag{7.92}$$

For every q such that $0 < q \le 1$, $\mathbf{A}^q_n$ is the product of two totally positive matrices, and so is itself totally positive. It then follows from Definition 7.4.6 that $(P^q_{n,0}, \ldots, P^q_{n,n})$ is totally positive on $[0,1]$. ■

Let p denote any polynomial in P_n, and let q, r denote any real numbers such that $0 < q, r \le 1$. Since $(P^q_{n,0}, \ldots, P^q_{n,n})$ and $(P^r_{n,0}, \ldots, P^r_{n,n})$ are both bases for P_n, there exist real numbers $a^q_0, \ldots, a^q_n$ and $a^r_0, \ldots, a^r_n$ such that

$$p(x) = a^q_0 P^q_{n,0}(x) + \cdots + a^q_n P^q_{n,n}(x) = a^r_0 P^r_{n,0}(x) + \cdots + a^r_n P^r_{n,n}(x), \tag{7.93}$$

and we can deduce from (7.74) that

$$[a^q_0, a^q_1, \ldots, a^q_n]\mathbf{T}^{n,q,r} = [a^r_0, a^r_1, \ldots, a^r_n]. \tag{7.94}$$

If $0 < q \le r^{n-1}$, the matrix $\mathbf{T}^{n,q,r}$ is totally positive and (see Problem 7.5.1) so is its transpose. In particular, the matrix $\mathbf{T}^{n,r,1}$ is totally positive for all r such that $0 < r \le 1$. Thus we see from (7.94) and Theorem 7.4.2 that

$$S^-\left(a^1_0, \ldots, a^1_n\right) \le S^-\left(a^r_0, \ldots, a^r_n\right) \le S^-\left(a^q_0, \ldots, a^q_n\right),$$

where

$$p(x) = a^1_0 P^1_{n,0}(x) + \cdots + a^1_n P^1_{n,n}(x). \tag{7.95}$$

Since $(P^1_{n,0}, \ldots, P^1_{n,n})$ is totally positive, it follows from Theorem 7.4.6 that for $0 < q \le r^{n-1} \le 1$ and with p defined by (7.93) and (7.95),

$$S^-(p) \le S^-\left(a^1_0, \ldots, a^1_n\right) \le S^-\left(a^r_0, \ldots, a^r_n\right) \le S^-\left(a^q_0, \ldots, a^q_n\right). \tag{7.96}$$

We can now state a generalization of Theorem 7.5.1.

Theorem 7.5.6 Let $B^q_n(f;x)$ denote the generalized Bernstein polynomial that we denoted by $B_n(f;x)$ in (7.51). Then

$$S^-(B^q_n f) \le S^-(f) \tag{7.97}$$

on $[0,1]$, and thus the operator B^q_n is variation-diminishing.

Proof. Let us choose

$$p(x) = B_n^q(f;x) = a_0^q P_{n,0}^q(x) + \cdots + a_n^q P_{n,n}^q(x)$$

in (7.97). We have already noted that the q-binomial coefficient $\begin{bmatrix} n \\ r \end{bmatrix}$ is a polynomial in q with positive integer coefficients, and so is positive if $q > 0$. Thus, for $q > 0$,

$$S^-(B_n^q f) \le S^-(f_0, f_1, \ldots, f_n) \le S^-(f),$$

where $f_r = f([r]/[n])$. ■

Let p denote any linear polynomial; that is, $p \in P_1$. Then, since B_n^q reproduces linear polynomials, we may deduce the following result from Theorem 7.5.6.

Theorem 7.5.7 For any function f and any linear polynomial p, we have

$$S^-(B_n^q f - p) \le S^-(B_n^q(f-p)) \le S^-(f-p), \tag{7.98}$$

for $0 < q \le 1$. ■

The next two theorems readily follow from Theorem 7.5.7.

Theorem 7.5.8 Let f be monotonically increasing (decreasing) on $[0,1]$. Then the generalized Bernstein polynomial $B_n^q f$ is also monotonically increasing (decreasing) on $[0,1]$, for $0 < q \le 1$.

Proof. We have already proved this in Theorem 7.1.2 when $q = 1$. Let us replace p in (7.98) by the constant c. Then, if f is monotonically increasing on $[0,1]$,

$$S^-(B_n^q f - c) \le S^-(f-c) \le 1$$

for all choices of constant c, and thus $B_n^q f$ is monotonically increasing or decreasing. Since

$$B_n^q(f;0) = f(0) \le f(1) = B_n^q(f;1),$$

$B_n^q f$ must be monotonically *increasing*. On the other hand, if f is monotonically *decreasing*, we may replace f by $-f$, and repeat the above argument, concluding that $B_n^q f$ is monotonically decreasing. ■

Theorem 7.5.9 If f is convex on $[0,1]$, then $B_n^q f$ is also convex on $[0,1]$, for $0 < q \le 1$.

Proof. Let p denote any linear polynomial. Then if f is convex, the graph of p can intersect that of f at no more than two points, and thus $S^-(f-p) \le 2$. It follows from (7.98) that for any q such that $0 < q \le 1$,

$$S^-(B_n^q f - p) = S^-(B_n^q(f-p)) \le S^-(f-p) \le 2. \tag{7.99}$$

Suppose the graph of p intersects that of $B_n^q f$ at a and b. Then we have $p(a) = B_n^q(f;a)$ and $p(b) = B_n^q(f;b)$, where $0 < a < b < 1$, and we see from (7.99) that $B_n^q f - p$ cannot change sign in (a,b). As we vary a and b, a continuity argument shows that the sign of $B_n^q f - p$ on (a,b) is the same for all a and b, $0 < a < b < 1$. From the convexity of f we see that in the limiting case where $a = 0$ and $b = 1$, $0 \le p(x) - f(x)$ on $[0,1]$, so that

$$0 \le B_n^q(p - f;x) = p(x) - B_n^q(f;x), \quad 0 \le x \le 1,$$

and thus B_n^q is convex. ■

We conclude this section by proving that if f is convex, the generalized Bernstein polynomials $B_n^q f$, for n fixed, are monotonic in q.

Theorem 7.5.10 For $0 < q \le r \le 1$ and for f convex on $[0,1]$, we have

$$f(x) \le B_n^r(f;x) \le B_n^q(f;x), \quad 0 \le x \le 1. \tag{7.100}$$

Proof. It remains only to establish the second inequality in (7.100), since the first inequality has already been proved in Theorem 7.3.3. Let us write

$$\zeta_{n,j}^q = \frac{[j]}{[n]} \quad \text{and} \quad a_{n,j}^q = \left[\begin{array}{c} n \\ j \end{array} \right].$$

Then, for any function g on [0,1],

$$B_n^q(g;x) = \sum_{j=0}^{n} g(\zeta_{n,j}^q) a_{n,j}^q P_{n,j}^q(x) = \sum_{j=0}^{n} \sum_{k=0}^{n} g(\zeta_{n,j}^q) a_{n,j}^q t_{j,k}^{n,q,r} P_{n,k}^r(x),$$

and thus

$$B_n^q(g;x) = \sum_{k=0}^{n} P_{n,k}^r(x) \sum_{j=0}^{n} t_{j,k}^{n,q,r} g(\zeta_{n,j}^q) a_{n,j}^q. \tag{7.101}$$

With $g(x) = 1$, this gives

$$1 = \sum_{j=0}^{n} a_{n,j}^q P_{n,j}^q(x) = \sum_{k=0}^{n} P_{n,k}^r(x) \sum_{j=0}^{n} t_{j,k}^{n,q,r} a_{n,j}^q$$

and hence

$$\sum_{j=0}^{n} t_{j,k}^{n,q,r} a_{n,j}^q = a_{n,k}^r, \quad 0 \le k \le n. \tag{7.102}$$

On putting $g(x) = x$ in (7.101), we obtain

$$x = \sum_{j=0}^{n} \zeta_{n,j}^q a_{n,j}^q P_{n,j}^q(x) = \sum_{k=0}^{n} P_{n,k}^r(x) \sum_{j=0}^{n} t_{j,k}^{n,q,r} \zeta_{n,j}^q a_{n,j}^q.$$

Since

$$\sum_{j=0}^{n} \zeta_{n,j}^{r} a_{n,j}^{r} P_{n,j}^{r}(x) = x,$$

we have

$$\sum_{j=0}^{n} t_{j,k}^{n,q,r} \zeta_{n,j}^{q} a_{n,j}^{q} = \zeta_{n,k}^{r} a_{n,k}^{r}, \quad 0 \le k \le n. \tag{7.103}$$

Let us write

$$\lambda_j = \frac{t_{j,k}^{n,q,r} a_{n,j}^{q}}{a_{n,k}^{r}},$$

and we see from (7.102) and (7.103), respectively, that

$$\sum_{j=0}^{n} \lambda_j = 1 \quad \text{and} \quad \zeta_{n,k}^{r} = \sum_{j=0}^{n} \lambda_j \zeta_{n,j}^{q}.$$

It then follows from Problem 7.1.3 that if f is convex,

$$f(\zeta_{n,k}^{r}) = f\left(\sum_{j=0}^{n} \lambda_j \zeta_{n,j}^{q}\right) \le \sum_{j=0}^{n} \lambda_j f(\zeta_{n,j}^{q}),$$

which gives

$$f(\zeta_{n,k}^{r}) \le \sum_{j=0}^{n} (a_{n,k}^{r})^{-1} t_{j,k}^{n,q,r} a_{n,j}^{q} f(\zeta_{n,j}^{q}). \tag{7.104}$$

On substituting

$$P_{n,j}^{q}(x) = \sum_{k=0}^{n} t_{j,k}^{n,q,r} P_{n,k}^{r}(x),$$

obtained from (7.74), into

$$B_n^{q}(f;x) = \sum_{j=0}^{n} f(\zeta_{n,j}^{q}) a_{n,j}^{q} P_{n,j}^{q}(x),$$

we find that

$$B_n^{q}(f;x) = \sum_{k=0}^{n} a_{n,k}^{r} P_{n,k}^{r}(x) \sum_{j=0}^{n} (a_{n,k}^{r})^{-1} t_{j,k}^{n,q,r} f(\zeta_{n,j}^{q}) a_{n,j}^{q}.$$

It then follows from (7.104) that

$$B_n^{q}(f;x) \ge \sum_{k=0}^{n} a_{n,k}^{r} P_{n,k}^{r}(x) f(\zeta_{n,k}^{r}) = B_n^{r}(f;x),$$

and this completes the proof. ■

Problem 7.5.1 Given that

$$\det \mathbf{A} = \det \mathbf{A}^T,$$

for any square matrix $\mathbf{A}$, deduce from Definition 7.4.1 that if the matrix $\mathbf{A}$ is totally positive, so also is $\mathbf{A}^T$.

Problem 7.5.2 Let $\mathbf{A}_1, \mathbf{A}_2, \ldots$ denote $m \times m$ matrices that are 1-banded, and whose nonzero elements are on the main diagonal and the diagonal above the main diagonal. Show by induction on j that for $1 \leq j \leq m-1$, the product $\mathbf{A}_1\mathbf{A}_1 \cdots \mathbf{A}_j$ is a j-banded matrix.

8
Properties of the q-Integers

Had I been content to write only three or four pages on the topics in this chapter, such material could have been incorporated in Chapter 1, since it is required in Chapters 1, 5, 6, and 7. However, I wished to say a little more about q-integers than is strictly necessary for the applications in these earlier chapters, and thought it would be misleading to begin a book on approximation theory with even a very short chapter on such material. Therefore, I hope that the reader, having worked with q-integers and q-binomial coefficients in these earlier chapters, will endorse my decision to say a little more about them in this final chapter.

8.1 The q-Integers

It is convenient to begin by repeating the definitions of a q-integer, a q-factorial, and a q-binomial coefficient, which we stated in Section 1.5. Let $\mathbb{N}$ denote $\{0, 1, 2, \ldots\}$, the set of nonnegative integers.

Definition 8.1.1 Given a value of $q > 0$ we define $[r]$, where $r \in \mathbb{N}$, as

$$[r] = \begin{cases} (1-q^r)/(1-q), & q \neq 1, \\ r, & q = 1, \end{cases} \tag{8.1}$$

and call $[r]$ a q-integer. Clearly, we can extend this definition, allowing r to be any real number in (8.1). We then call $[r]$ a q-real. ■

For any given $q > 0$ let us define

$$\mathbb{N}_q = \{[r], \text{ with } r \in \mathbb{N}\}, \tag{8.2}$$

and we can see from Definition 8.1.1 that

$$\mathbb{N}_q = \{0, 1, 1+q, 1+q+q^2, 1+q+q^2+q^3, \dots\}. \tag{8.3}$$

Obviously, the set of q-integers $\mathbb{N}_q$ generalizes the set of nonnegative integers $\mathbb{N}$, which we recover by putting $q = 1$.

Definition 8.1.2 Given a value of $q > 0$ we define $[r]!$, where $r \in \mathbb{N}$, as

$$[r]! = \begin{cases} [r][r-1]\cdots[1], & r \geq 1, \\ 1, & r = 0, \end{cases} \tag{8.4}$$

and call $[r]!$ a q-factorial. ■

Definition 8.1.3 We define a q-binomial coefficient as

$$\left[\begin{array}{c} t \\ r \end{array} \right] = \frac{[t]\,[t-1]\cdots[t-r+1]}{[r]!}, \tag{8.5}$$

for all real t and integers $r \geq 0$, and as zero otherwise. ■

Since for the rest of this chapter we will be concerned with q-binomial coefficients for which $t = n \geq r \geq 0$, where n is an integer, it seems appropriate to make these the subject of a separate definition.

Definition 8.1.4 For any integers n and r, we define

$$\left[\begin{array}{c} n \\ r \end{array} \right] = \frac{[n][n-1]\cdots[n-r+1]}{[r]!} = \frac{[n]!}{[r]![n-r]!}, \tag{8.6}$$

for $n \geq r \geq 0$, and as zero otherwise. These are called *Gaussian polynomials*, named after C. F. Gauss. We will show in the next section that they are indeed polynomials. ■

The Gaussian polynomials satisfy the Pascal-type relations

$$\left[\begin{array}{c} n \\ r \end{array} \right] = \left[\begin{array}{c} n-1 \\ r-1 \end{array} \right] + q^r \left[\begin{array}{c} n-1 \\ r \end{array} \right] \tag{8.7}$$

and

$$\left[\begin{array}{c} n \\ r \end{array} \right] = q^{n-r} \left[\begin{array}{c} n-1 \\ r-1 \end{array} \right] + \left[\begin{array}{c} n-1 \\ r \end{array} \right]. \tag{8.8}$$

Although the identities (8.7) and (8.8) hold for all real n and all integers $r \geq 0$, we will be concerned with their application to integers $n \geq r \geq 0$ only. To verify (8.7) for integers $n \geq r \geq 0$ we write, using (8.6),

$$\left[\begin{array}{c} n-1 \\ r-1 \end{array} \right] + q^r \left[\begin{array}{c} n-1 \\ r \end{array} \right] = ([r] + q^r[n-r]) \frac{[n-1]!}{[r]![n-r]!}. \tag{8.9}$$

Since

$$[r] + q^r[n-r] = \frac{1-q^r}{1-q} + \frac{q^r(1-q^{n-r})}{1-q} = \frac{1-q^n}{1-q} = [n],$$

the right side of (8.9) indeed simplifies to give $\begin{bmatrix} n \\ r \end{bmatrix}$, thus justifying (8.7). We may verify (8.8) similarly. Note that when we put $q = 1$, both identities (8.7) and (8.8) reduce to the more familiar Pascal identity for ordinary binomial coefficients,

$$\begin{pmatrix} n \\ r \end{pmatrix} = \begin{pmatrix} n-1 \\ r-1 \end{pmatrix} + \begin{pmatrix} n-1 \\ r \end{pmatrix}. \tag{8.10}$$

It is obvious from the relation

$$\begin{pmatrix} n \\ r \end{pmatrix} = \frac{n!}{r!(n-r)!}, \qquad n \geq r \geq 0,$$

that the ordinary binomial coefficient is a positive rational number. However, we know that we can say more than this, that for $n \geq r \geq 0$ it is always a positive integer. We may deduce this from the Pascal identity (8.10), using induction on n. Likewise, we can verify from (8.6) that

$$\begin{bmatrix} n \\ r \end{bmatrix} = \frac{(1-q^{n-r+1})(1-q^{n-r+2})\cdots(1-q^n)}{(1-q)(1-q^2)\cdots(1-q^r)}, \tag{8.11}$$

and thus $\begin{bmatrix} n \\ r \end{bmatrix}$ is a rational function of the parameter q. However, just as the ordinary binomial coefficient is an integer rather than merely a rational number, we can deduce from either one of the Pascal-type identities (8.7) and (8.8) that the expression on the right side of equation (8.11) is a *polynomial* in q rather than a rational function of q. The details are given in the proof of Theorem 8.2.1 in the next section.

Consider the identity

$$\prod_{s=1}^{n}(1+q^{s-1}x) = \sum_{s=0}^{n} q^{s(s-1)/2} \begin{bmatrix} n \\ s \end{bmatrix} x^s, \tag{8.12}$$

which reduces to the binomial expansion

$$(1+x)^n = \sum_{s=0}^{n} \begin{pmatrix} n \\ s \end{pmatrix} x^s \tag{8.13}$$

when we set q equal to 1. To justify (8.12), we begin by writing

$$G_n(x) = (1+x)(1+qx)\cdots(1+q^{n-1}x) = \sum_{r=0}^{n} c_r x^r. \tag{8.14}$$

We replace x by qx and deduce from (8.14) that

$$(1+q^n x)G_n(x) = (1+x)G_n(qx),$$

so that

$$(1+q^n x)\sum_{r=0}^{n} c_r x^r = (1+x)\sum_{r=0}^{n} c_r (qx)^r.$$

Then we equate coefficients of x^s to obtain

$$c_s + q^n c_{s-1} = q^s c_s + q^{s-1} c_{s-1},$$

so that

$$c_s = q^{s-1}\left(\frac{1-q^{n-s+1}}{1-q^s}\right)c^{s-1} = q^{s-1}\frac{[n-s+1]}{[s]}c^{s-1},$$

for $1 \le s \le n$, and we also have $c_0 = 1$. It readily follows that

$$c_s = q^{s(s-1)/2}\frac{[n-s+1][n-s+2]\cdots[n]}{[s][s-1]\cdots[1]}c_0 = q^{s(s-1)/2}\begin{bmatrix} n \\ s \end{bmatrix}, \quad (8.15)$$

which verifies (8.12).

We now examine the expression

$$\prod_{s=1}^{n}(1+q^{s-1}x)^{-1} = \sum_{s=0}^{\infty}\begin{bmatrix} n+s-1 \\ s \end{bmatrix}(-x)^s \quad (8.16)$$

for the inverse of the q-binomial expansion (8.12). We can verify (8.16) by adapting the method we used above to verify (8.12). Let us write

$$H_n(x) = (1+x)^{-1}(1+qx)^{-1}\cdots(1+q^{n-1}x)^{-1} = \sum_{s=0}^{\infty} d_s x^s,$$

replace x by qx, and obtain the relation

$$(1+x)H_n(x) = (1+q^n x)H_n(qx),$$

so that

$$(1+x)\sum_{s=0}^{\infty} d_s x^s = (1+q^n x)\sum_{s=0}^{\infty} d_s (qx)^s.$$

On equating coefficients of x^s, we obtain

$$d_s = -\left(\frac{1-q^{n+s-1}}{1-q^s}\right)d_{s-1} = -\frac{[n+s-1]}{[s]}d_{s-1}$$

for $s \ge 1$, with $d_0 = 1$. We deduce that

$$d_s = (-1)^s\begin{bmatrix} n+s-1 \\ s \end{bmatrix},$$

which justifies (8.16).

The identities (8.12) and (8.16), which have applications in the theory of partitions, both go back to Euler. See Hardy and Wright [24].

Problem 8.1.1 Show that for any real numbers s, t, and u,

$$[s]\,[t+u]-[s+u][t]=q^t[u]\,[s-t].$$

Problem 8.1.2 Verify the Pascal-type identity (8.8).

Problem 8.1.3 Show that for all integers $n \geq r \geq 0$,

$$\left[\begin{array}{c} n \\ r \end{array}\right] = \left[\begin{array}{c} n \\ n-r \end{array}\right].$$

Problem 8.1.4 Replace r by $n-r$ in the identity (8.7). Apply the result of Problem 8.1.3 to both terms on the right side of the identity and so deduce the second Pascal identity (8.8) from the first.

Problem 8.1.5 Deduce from (8.5) that

$$\left[\begin{array}{c} -n \\ r \end{array}\right] = (-1)^r q^{-(2n+r-1)r/2} \left[\begin{array}{c} n+r-1 \\ r \end{array}\right],$$

for all $n \geq r \geq 0$.

Problem 8.1.6 With $G_n(x)$ as defined by (8.14), verify that the relation

$$G_n(x) = \sum_{s=0}^{n} q^{s(s-1)/2} \left[\begin{array}{c} n \\ s \end{array}\right] x^s, \tag{8.17}$$

whose coefficients were derived above in (8.15), holds for $n = 1$. Write $G_{n+1}(x) = (1+q^n x)G_n(x)$ and deduce that the coefficient of x^s in $G_{n+1}(x)$ is

$$q^{s(s-1)/2} \left[\begin{array}{c} n \\ s \end{array}\right] + q^n \cdot q^{(s-1)(s-2)/2} \left[\begin{array}{c} n \\ s-1 \end{array}\right].$$

Show, using (8.8), that this simplifies to give

$$q^{s(s-1)/2} \left(\left[\begin{array}{c} n \\ s \end{array}\right] + q^{n-s+1} \left[\begin{array}{c} n \\ s-1 \end{array}\right] \right) = q^{s(s-1)/2} \left[\begin{array}{c} n+1 \\ s \end{array}\right],$$

justifying that (8.17) holds when n is replaced by $n+1$, thus completing an induction argument that (8.12) holds for all $n \geq 1$.

Problem 8.1.7 By comparing coefficients of x^k on both sides of the identity $G_n(x)H_n(x) = 1$, show that for $1 \leq k \leq n$,

$$\sum_{t=0}^{k} (-1)^{k-t} q^{t(t-1)/2} \left[\begin{array}{c} n \\ t \end{array}\right] \left[\begin{array}{c} n+k-t-1 \\ k-t \end{array}\right] = 0.$$

Problem 8.1.8 Use induction on n to verify that

$$\sum_{i=0}^{s}(-1)^i q^{i(i-1)/2} \begin{bmatrix} n+1 \\ i \end{bmatrix} = (-1)^s q^{s(s+1)/2} \begin{bmatrix} n \\ s \end{bmatrix},$$

for $0 \le s \le n$.

8.2 Gaussian Polynomials

In this section we discuss properties of the Gaussian polynomials, which are defined above by (8.6).

Theorem 8.2.1 For $0 \le r \le n$, the q-binomial coefficient $\begin{bmatrix} n \\ r \end{bmatrix}$ is a polynomial of degree $r(n-r)$ in q, and all its coefficients are positive.

Proof. We will use induction on n. The above result is clearly true for $n = 0$. Now let us assume that the result holds for some fixed value of $n \ge 0$ and all $n+1$ values of r satisfying $0 \le r \le n$. Then let us consider

$$\begin{bmatrix} n+1 \\ r \end{bmatrix} = \begin{bmatrix} n \\ r-1 \end{bmatrix} + q^r \begin{bmatrix} n \\ r \end{bmatrix}, \tag{8.18}$$

which is just (8.7) with n replaced by $n+1$. From our inductive hypothesis, we note that both terms on the right of (8.18) are polynomials with positive coefficients, the first term being the zero polynomial when $r = 0$. The degree of the first term is $(r-1)(n+1-r)$, when $r > 0$, and the degree of the second term is

$$r + r(n-r) = r(n+1-r).$$

Thus $\begin{bmatrix} n+1 \\ r \end{bmatrix}$ is a polynomial of degree $r(n+1-r)$ with positive coefficients, and this completes the proof. ■

Example 8.2.1 Using (8.6), we find that

$$\begin{bmatrix} 5 \\ 2 \end{bmatrix} = \frac{[5]!}{[2]![3]!} = \frac{[5][4]}{[2][1]} = \frac{(1-q^5)(1-q^4)}{(1-q^2)(1-q)},$$

which simplifies to give

$$\begin{bmatrix} 5 \\ 2 \end{bmatrix} = 1 + q + 2q^2 + 2q^3 + 2q^4 + q^5 + q^6, \tag{8.19}$$

and we note the symmetry in the coefficients of this polynomial. ■

Definition 8.2.1 We say that a polynomial

$$p(x) = a_0 + a_1 x + \cdots + a_{m-1}x^{m-1} + a_m x^m \tag{8.20}$$

is *reciprocal* if its coefficients satisfy the condition

$$a_r = a_{m-r}, \quad 0 \le r \le m. \qquad \blacksquare$$

Definition 8.2.2 We say that the polynomial p in (8.20) is *unimodal* in its coefficients if for some integer k, $0 \le k \le m$,

$$a_0 \le a_1 \le \cdots \le a_k \ge a_{k+1} \ge \cdots \ge a_m. \qquad \blacksquare$$

Thus the polynomial in (8.19) is both reciprocal and unimodal in its coefficients. Since

$$x^m p\left(\frac{1}{x}\right) = a_m + a_{m-1}x + \cdots + a_1 x^{m-1} + a_0 x^m,$$

the property that a polynomial p of degree m is reciprocal is equivalent to saying that

$$x^m p\left(\frac{1}{x}\right) = p(x). \tag{8.21}$$

Let us now write

$$[j]' = \frac{1 - q^{-j}}{1 - q^{-1}}, \tag{8.22}$$

so that $[j]'$ is derived from $[j]$ by substituting $1/q$ for q. We note that

$$q^{j-1}[j]' = [j]. \tag{8.23}$$

Similarly, let us write $[r]'!$ and $\begin{bmatrix} n \\ r \end{bmatrix}'$ to denote the expressions we obtain when we substitute $1/q$ for q in $[r]!$ and $\begin{bmatrix} n \\ r \end{bmatrix}$, respectively. We then have

$$[r]'! = \left(\frac{1 - q^{-r}}{1 - q^{-1}}\right)\left(\frac{1 - q^{-r+1}}{1 - q^{-1}}\right)\cdots\left(\frac{1 - q^{-1}}{1 - q^{-1}}\right), \quad r \ge 1,$$

so that

$$q^{r(r-1)/2}\,[r]'! = [r]!. \tag{8.24}$$

We note that (8.24) holds for all $r \ge 0$, and since

$$\frac{1}{2}n(n-1) - \frac{1}{2}r(r-1) - \frac{1}{2}(n-r)(n-r-1) = r(n-r),$$

it follows from (8.6) and (8.24) that

$$q^{r(n-r)} \begin{bmatrix} n \\ r \end{bmatrix}' = \begin{bmatrix} n \\ r \end{bmatrix}. \tag{8.25}$$

Since the degree of the Gaussian polynomial $\begin{bmatrix} n \\ r \end{bmatrix}$ is $r(n-r)$, it follows from (8.25) and (8.21) that every Gaussian polynomial is reciprocal, as we found for the particular case given in Example 8.2.1.

Let us return to (8.14), from which we obtain

$$G_{i+j}(x) = G_i(x)(1+q^i x)\cdots(1+q^{i+j-1}x),$$

so that

$$G_{i+j}(x) = G_i(x)G_j(q^i x),$$

and using (8.17), we derive the relation

$$G_{i+j}(x) = \sum_{s=0}^{i} q^{s(s-1)/2} \begin{bmatrix} i \\ s \end{bmatrix} x^s \sum_{s=0}^{j} q^{s(s-1)/2} \begin{bmatrix} j \\ s \end{bmatrix} (q^i x)^s.$$

On equating powers of x^r in the last equation, we find that

$$q^{r(r-1)/2} \begin{bmatrix} i+j \\ r \end{bmatrix} = \sum_{t=0}^{r} q^{\alpha} \begin{bmatrix} i \\ t \end{bmatrix} \begin{bmatrix} j \\ r-t \end{bmatrix}, \tag{8.26}$$

where

$$\alpha = \frac{1}{2}t(t-1) + \frac{1}{2}(r-t)(r-t-1) + i(r-t),$$

and since

$$\alpha - \frac{1}{2}r(r-1) = (r-t)(i-t),$$

we see that (8.26) yields the identity

$$\begin{bmatrix} i+j \\ r \end{bmatrix} = \sum_{t=0}^{r} q^{(r-t)(i-t)} \begin{bmatrix} i \\ t \end{bmatrix} \begin{bmatrix} j \\ r-t \end{bmatrix}. \tag{8.27}$$

This is a q-analogue of the Chu–Vandermonde identity,

$$\begin{pmatrix} i+j \\ r \end{pmatrix} = \sum_{t=0}^{r} \begin{pmatrix} i \\ t \end{pmatrix} \begin{pmatrix} j \\ r-t \end{pmatrix}. \tag{8.28}$$

If we choose $i = 1$ and $j = n-1$, we obtain only two nonzero terms on the right of (8.27) for $r \geq 1$, and indeed this choice of i and j shows that (8.27) is a generalization of the Pascal identity (8.7).

Example 8.2.2 With $i = 4,\ j = 3$, and $r = 5$ in (8.27), we obtain

$$\begin{bmatrix} 7 \\ 5 \end{bmatrix} = q^6 \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} + q^2 \begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} + q^0 \begin{bmatrix} 4 \\ 4 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix}.$$

Since

$$\begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 3 \\ 3 \end{bmatrix} = (1+q+2q^2+q^3+q^4)\cdot 1,$$

$$\begin{bmatrix} 4 \\ 3 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} = (1 + q + q^2 + q^3) \cdot (1 + q + q^2),$$

and

$$\begin{bmatrix} 4 \\ 4 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} = 1 \cdot (1 + q + q^2),$$

the above expression simplifies to give

$$\begin{bmatrix} 7 \\ 5 \end{bmatrix} = 1 + q + 2q^2 + 2q^3 + 3q^4 + 3q^5 + 3q^6 + 2q^7 + 2q^8 + q^9 + q^{10}.$$

Alternatively, since $\begin{bmatrix} 7 \\ 5 \end{bmatrix} = \begin{bmatrix} 7 \\ 2 \end{bmatrix}$, we can apply (8.27) with $i = 4,\ j = 3,$ and $r = 2$ to give

$$\begin{bmatrix} 7 \\ 5 \end{bmatrix} = q^8 \begin{bmatrix} 4 \\ 0 \end{bmatrix} \begin{bmatrix} 3 \\ 2 \end{bmatrix} + q^3 \begin{bmatrix} 4 \\ 1 \end{bmatrix} \begin{bmatrix} 3 \\ 1 \end{bmatrix} + q^0 \begin{bmatrix} 4 \\ 2 \end{bmatrix} \begin{bmatrix} 3 \\ 0 \end{bmatrix},$$

which simplifies to give the same result for $\begin{bmatrix} 7 \\ 5 \end{bmatrix}$. ■

Gaussian polynomials arise in the theory of partitions. (See Andrews [1] or Hardy and Wright [24] for further material on partitions.) Let $p(s)$ denote the number of *partitions* of the positive integer s, meaning the number of ways of representing s as the sum of positive integers, which are called the *parts* of s. In counting the number of partitions, we ignore the order of the parts. For example, since

$$5 = 4+1 = 3+2 = 3+1+1 = 2+2+1 = 2+1+1+1 = 1+1+1+1+1,$$

we say that there are seven partitions of the positive integer 5, and write $p(5) = 7$.

Let us now consider a function whose power series, in powers of q, has $p(s)$ as the coefficient of q^s. This function, which was first obtained by Euler, is

$$\frac{1}{(1-q)(1-q^2)(1-q^3)\cdots} = \sum_{s=0}^{\infty} p(s) q^s, \tag{8.29}$$

where we conveniently define $p(0) = 1$. The function on the left side of (8.29) is called a *generating function*. Each partition of s supplies one unit to the coefficient of q^s. For example, the partition $5 = 3+1+1$ corresponds to constructing q^5 by multiplying q^3 chosen from the expansion

$$\frac{1}{1-q^3} = 1 + q^3 + (q^3)^2 + \cdots$$

by $q^{1+1} = q^2$ chosen from the expansion

$$\frac{1}{1-q} = 1 + q + q^2 + \cdots.$$

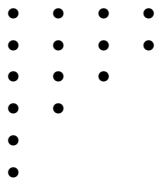

TABLE 8.1. The partition $15 = 4+4+3+2+1+1$.

We can represent a partition by an array of nodes, arranged in rows and columns. Each row of nodes represents one part, and the parts are arranged in decreasing order, with the largest part in the first row. This is called a *Ferrers graph*, named after N. M. Ferrers (1829–1903). Table 8.1 gives a Ferrers graph for the partition $15 = 4+4+3+2+1+1$. Many interesting results can be deduced with the aid of a Ferrers graph. For example, we can "read" a graph by columns rather than rows. Then Table 8.1 would represent the partition $15 = 6+4+3+2$. A partition read by rows and its corresponding partition read by columns are said to be *conjugate* to each other. By considering conjugate partitions we can see immediately that the number of partitions of s into exactly m parts is the same as the number of partitions of s whose largest part is m. A graph is said to be self-conjugate if its graph and the corresponding conjugate graph are equal.

Example 8.2.3 The partitions of 10 into 3 parts are $8+1+1$, $7+2+1$, $6+3+1$, $6+2+2$, $5+4+1$, $5+3+2$, $4+4+2$, and $4+3+3$, while the partitions of 10 into parts whose largest is 3 are $3+3+3+1$, $3+3+2+2$, $3+3+2+1+1$, $3+3+1+1+1+1$, $3+2+2+2+1$, $3+2+2+1+1+1$, $3+2+1+1+1+1+1$, and $3+1+1+1+1+1+1+1$. There are eight of each kind of partition, in accord with the result deduced above. As an exercise, the reader is encouraged to determine which pairs of partitions, one being chosen from each of the above sets with eight members, are conjugate to each other. ■

Let $p(m,n;s)$ denote the number of partitions of s into at most m parts of size at most n, and let $p'(m,n;s)$ denote the number of partitions of s into *exactly* m parts of size at most n. These are defined for all integers $m \geq 1$ and $n \geq 1$, and $p'(m,n;s)$ is obviously zero for $s < m$. It follows immediately from these definitions that

$$p'(m,n;s) = p(m,n;s) - p(m-1,n;s). \tag{8.30}$$

If we extend the definition of $p(m,n;s)$ to $m = 0$, writing $p(0,n;s) = 0$, (8.30) will hold for all $m,n \geq 1$. We also write $p''(m,n;s)$ to denote the number of partitions of s into exactly m parts, the largest being *exactly* n. It follows from this and the definition of $p'(m,n;s)$ that

$$p''(m,n;s) = p'(m,n;s) - p'(m,n-1;s), \tag{8.31}$$

and (8.31) will hold for all $m, n \geq 1$ if we extend the definition of $p'(m, n; s)$ by writing $p'(m, 0; s) = 0$.

Example 8.2.4 From the partitions of 5 given just above (8.29), we may verify that $p(4, 3; 5) = 4$, $p(3, 3; 5) = 3$, $p'(4, 3; 5) = 1$, $p'(4, 2; 5) = 1$, and $p''(4, 3; 5) = 0$. ■

We are now ready to derive a generating function for $p''(m, n; s)$ and hence, via (8.30) and (8.31), obtain generating functions for $p'(m, n; s)$ and $p(m, n; s)$. Our approach is to consider the Ferrers graphs for the partitions enumerated by $p''(m, n; s)$. Such partitions are either in the set S_1, say, whose smallest part is 1, or in the set S_2 whose smallest part is greater than 1. The number of partitions in the set S_1 (think of removing the last part, of size 1, from all such partitions) is just the number of partitions of $s - 1$ into $m - 1$ parts, with largest part n, which is $p''(m - 1, n; s - 1)$. If we remove the first column from the Ferrers graphs of all partitions in the set S_2, we see that the number of partitions in S_2 is the same as the number of partitions of $s - m$ into m parts, with largest part $n - 1$, which is $p''(m, n - 1; s - m)$. We have thus established the recurrence relation

$$p''(m, n; s) = p''(m - 1, n; s - 1) + p''(m, n - 1; s - m), \tag{8.32}$$

and it is clear from the definition of $p''(m, n; s)$ that

$$p''(m, n; s) = 0 \quad \text{if} \quad s < m + n - 1. \tag{8.33}$$

We will show by induction that

$$q^{m+n-1} \begin{bmatrix} m+n-2 \\ m-1 \end{bmatrix} = \sum_{s=m+n-1}^{\infty} p''(m, n; s) q^s. \tag{8.34}$$

To verify (8.34), let us write

$$q^{m+n-1} \begin{bmatrix} m+n-2 \\ m-1 \end{bmatrix} = \sum_{s=m+n-1}^{\infty} a(m, n; s) q^s. \tag{8.35}$$

Now it follows from (8.7) that

$$q^{m+n-1} \begin{bmatrix} m+n-2 \\ m-1 \end{bmatrix} = q^{m+n-1} \begin{bmatrix} m+n-3 \\ m-2 \end{bmatrix} + q^{2m+n-2} \begin{bmatrix} m+n-3 \\ m-1 \end{bmatrix},$$

and on equating coefficients of q^s in the last equation, we obtain

$$a(m, n; s) = a(m - 1, n; s - 1) + a(m, n - 1; s - m). \tag{8.36}$$

We note that this recurrence relation for $a(m, n; s)$ is the same as the recurrence relation (8.32) for $p''(m, n; s)$, and since

$$a(1, 1; 1) = p''(1, 1; 1) = 1,$$

we can say that $a(m,n;s) = p''(m,n;s)$ for all $m,n \geq 1$ such that $m+n=2$ and all s such that $m+n-1 \leq s \leq mn$. We now assume that $a(m,n;s) = p''(m,n;s)$ for all $m,n \geq 1$ such that $m+n=k$, for some integer $k \geq 2$ and all s such that $m+n-1 \leq s \leq mn$. It follows from the recurrence relations (8.32) and (8.36) that $a(m,n;s) = p''(m,n;s)$ for all $m,n \geq 1$ such that $m+n=k+1$ and all s such that $m+n-1 \leq s \leq mn$. Thus, by induction, the sequences $a(m,n;s)$ and $p''(m,n;s)$ are the same, which justifies (8.34). In view of (8.31) and the relation

$$q^{m+n-1} \begin{bmatrix} m+n-2 \\ m-1 \end{bmatrix} = q^m \begin{bmatrix} m+n-1 \\ m \end{bmatrix} - q^m \begin{bmatrix} m+n-2 \\ m \end{bmatrix}$$

we readily deduce that

$$q^m \begin{bmatrix} m+n-1 \\ m \end{bmatrix} = \sum_{s=m}^{\infty} p'(m,n;s) q^s. \tag{8.37}$$

Similarly, we deduce from (8.30) and the relation

$$q^m \begin{bmatrix} m+n-1 \\ m \end{bmatrix} = \begin{bmatrix} m+n \\ m \end{bmatrix} - \begin{bmatrix} m+n-1 \\ m-1 \end{bmatrix}$$

that

$$\begin{bmatrix} m+n \\ m \end{bmatrix} = \sum_{s=0}^{\infty} p(m,n;s) q^s. \tag{8.38}$$

Observe that since

$$\begin{bmatrix} m+n \\ m \end{bmatrix} = \frac{[m+n]!}{[m]![n]!},$$

the generating function for $p(m,n;s)$ is symmetric in m and n. We deduce that

$$p(m,n;s) = p(n,m;s),$$

which also follows by considering conjugate Ferrers graphs and employing a similar argument to that used just before Example 8.21.

Next we determine a one-to-one correspondence between the partitions of s into at most m parts of size at most n and the partitions of $mn-s$ of the same kind. We obtain this by "subtracting" the Ferrers graph of such a partition of s from the Ferrers graph of the partition of mn into m parts of size n, the latter being a rectangular array of nodes. Table 8.2 illustrates this correspondence for partitions into at most 7 parts of size at most 5, between the two partitions $15 = 4+4+3+2+1+1$ and $20 = 5+4+4+3+2+1+1$. In the general case, corresponding to any partition of s into at most m parts of size at most n,

$$s = c_1 + c_2 + \cdots + c_m,$$

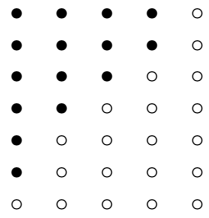

TABLE 8.2. The partitions $15 = 4+4+3+2+1+1$ and $20 = 5+4+4+3+2+1+1$.

where $n \geq c_1 \geq c_2 \geq \cdots \geq c_m \geq 0$, we have the partition

$$mn - s = (n - c_m) + (n - c_{m-1}) + \cdots + (n - c_1) = d_1 + d_2 + \cdots + d_m, \quad (8.39)$$

where each d_r is defined by $d_r = n - c_{m-r+1}$. Since

$$n \geq d_1 \geq d_2 \geq \cdots \geq d_m \geq 0$$

(8.39) does indeed define a partition of $mn - s$ into at most m parts of size at most n. From this one-to-one correspondence between such partitions of s and $mn - s$, we deduce that

$$p(m, n; s) = p(m, n; mn - s). \quad (8.40)$$

From (8.38), this shows that every Gaussian polynomial is reciprocal, as we showed by other means earlier in this section. It can also be shown that every Gaussian polynomial is unimodal. This is equivalent to showing that

$$p(m, n; s - 1) \leq p(m, n; s)$$

for $0 \leq s \leq \frac{1}{2}mn$. Somewhat surprisingly, there does not seem to be any simple proof of this depending on combinatorial arguments of the sort we have used in this section, and we therefore omit the proof. See Andrews [1].

Problem 8.2.1 Consider the $n+1$ Gaussian polynomials $\begin{bmatrix} n \\ r \end{bmatrix}$, where n is fixed and $0 \leq r \leq n$. Show that the greatest degree of these polynomials is $\frac{1}{4}n^2$ when n is even, this being attained when $r = \frac{1}{2}n$, and the greatest degree is $\frac{1}{4}(n^2 - 1)$ when n is odd, attained when $r = \frac{1}{2}(n \pm 1)$.

Problem 8.2.2 Repeat the proof given in the text that $\begin{bmatrix} n \\ r \end{bmatrix}$ is a polynomial of degree $r(n - r)$, but apply the induction argument to the second Pascal-type relation (8.8) instead of (8.7).

Problem 8.2.3 By reversing the order of summation in (8.27), obtain the following alternative expression for the Chu–Vandermonde identity:

$$\begin{bmatrix} i+j \\ r \end{bmatrix} = \sum_{t=0}^{r} q^{t(i+t-r)} \begin{bmatrix} i \\ r-t \end{bmatrix} \begin{bmatrix} j \\ t \end{bmatrix}.$$

Problem 8.2.4 Display the Ferrers graphs of the self-conjugate partitions of 5 and 10.

Problem 8.2.5 Show that the number of self-conjugate partitions of s is the same as the number of partitions of s into distinct odd numbers.

Problem 8.2.6 Show that the generating function for self-conjugate partitions is the infinite product

$$(1+x)(1+x^3)(1+x^5)(1+x^7)\cdots.$$

Problem 8.2.7 Show that $p''(m,n,m+n-1) = p''(n,m,m+n-1) = 1$ for all $m,n \geq 1$. What are the values of $p''(m,n,m+n)$, $p''(m,n,m+n+1)$ and $p''(m,n,m+n+2)$?

Problem 8.2.8 Show that

$$p(m,n;s) = \sum_{i=1}^{m}\sum_{j=1}^{n} p''(i,j;s),$$

noting that $p''(i,j;s)$ is zero if $s < i+j-1$.

Problem 8.2.9 Show that $p(6,6;18) = 58$ by determining the largest coefficient in the Gaussian polynomial $\begin{bmatrix} 12 \\ 6 \end{bmatrix}$, using (8.27).

Problem 8.2.10 Show that

$$\begin{bmatrix} 2n \\ n \end{bmatrix} = \sum_{t=0}^{n} q^{t^2} \begin{bmatrix} n \\ t \end{bmatrix}^2.$$

References

[1] G. E. Andrews. *The Theory of Partitions*, Cambridge University Press, Cambridge, 1998.

[2] D. L. Berman. On the best grid system for parabolic interpolation, *Izv. Vyss. Učebn. Zaved. Matematika* **4**, 20–25, 1963. (In Russian.)

[3] S. N. Bernstein. Démonstration du théorème de Weierstrass fondée, *Comm. Kharkov Math. Soc.* **13**, 1–2, 1912.

[4] S. N. Bernstein. Sur la limitation des valeurs d'une polynôme $P(x)$ de degré n sur tout un segment par ses valeurs en $n+1$ points du segment, *Izv. Akad. Nauk SSSR* **7**, 1025–1050, 1931.

[5] H. Bohman. On approximation of continuous and of analytic functions, *Arkiv för Matematik* **2**, 43–56, 1952.

[6] L. Brutman. On the Lebesgue function for polynomial interpolation, *SIAM J. Numer. Anal.* **14**, 694–704, 1978.

[7] E. W. Cheney. *Introduction to Approximation Theory*, McGraw–Hill, New York, 1966.

[8] K. C. Chung and T. H. Yao. On lattices admitting unique Lagrange interpolation, *SIAM J. Numer. Anal.* **15**, 735–743, 1977.

[9] C. Cryer. The *LU* factorization of totally positive matrices, *Linear Alg. Appl.* **7**, 83–92, 1973.

[10] P. J. Davis. *Interpolation and Approximation*, Dover, New York, 1976.

[11] Philip J. Davis and Philip Rabinowitz. *Methods of Numerical Integration*, Academic Press, New York, 1975.

[12] C. de Boor and A. Pinkus. Proof of the conjectures of Bernstein and Erdős concerning the optimal nodes for polynomial interpolation, *J. Approx. Theory* **24**, 289–303, 1978.

[13] C. de Boor and A. Pinkus. The approximation of totally positive banded matrices by a strictly banded positive one, *Linear Algebra Appl.* **42**, 81–98, 1982.

[14] Deutsch, Frank. *Best Approximation in Inner Product Spaces*, Springer-Verlag, New York, 2001.

[15] C. H. Edwards, Jr. *The Historical Development of the Calculus*, Springer-Verlag, New York, 1979.

[16] H. Ehlich and K. Zeller. Auswertung der Normen von Interpolationsoperatoren, *Math. Annalen* **164**, 105–112, 1966.

[17] P. Erdős. Problems and results on the theory of interpolation I, *Acta Math. Acad. Sci. Hungar.* **9**, 381–388, 1958.

[18] P. Erdős. Problems and results on the theory of interpolation II, *Acta Math. Acad. Sci. Hungar.* **12**, 235–244, 1961.

[19] L. Fejér. Über Weierstrassche Approximation besonders durch Hermitesche Interpolation, *Math. Annalen* **102**, 707–725, 1930.

[20] Israel Gohberg and Israel Koltracht. Triangular factors of Cauchy and Vandermonde matrices, *Integral Equat. Oper. Th.* **26**, 46–59, 1996.

[21] H. H. Goldstine. *A History of Numerical Analysis from the 16th through the 19th Century*, Springer-Verlag, New York, 1977.

[22] Tim N. T. Goodman. Total positivity and the shape of curves, *Total Positivity and its Applications (Jaca, 1994)*, *Math. Appl.* **359**, 157–186, Kluwer, Dordrecht, 1996.

[23] Tim N. T. Goodman, Halil Oruç, and G. M. Phillips. Convexity and generalized Bernstein polynomials, *Proc. Edin. Math. Soc.* **42**, 179–190, 1999.

[24] G. H. Hardy and E. M. Wright. *An Introduction to the Theory of Numbers, 5th Edition*, Clarendon Press, Oxford, 1979.

[25] Nicholas J. Higham. *Accuracy and Stability of Numerical Algorithms*, Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[26] J. Hoschek and D. Lasser. *Fundamentals of Computer Aided Geometric Design*, A. K. Peters, Wellesley, Massachusetts, 1993.

[27] John M. Howie. *Real Analysis*, Springer-Verlag, London, 2001.

[28] S. Karlin. *Total Positivity*, Stanford University Press, Stanford, 1968.

[29] T. A. Kilgore. A characterization of the Lagrange interpolating projection with minimal Tchebycheff norm, *J. Approx. Theory* **24**, 273–288, 1978.

[30] Z. F. Koçak and G. M. Phillips. B-splines with geometric knot spacings, *BIT* **34**, 388–399, 1994.

[31] P. P. Korovkin. On convergence of linear positive operators in the space of continuous functions, *Doklady Akademii Nauk SSSR* **90**, 961–964, 1953.

[32] S. L. Lee and G. M. Phillips. Polynomial interpolation at points of a geometric mesh on a triangle, *Proc. Roy. Soc. Edin.* **108A**, 75–87, 1988.

[33] S. L. Lee and G. M. Phillips. Pencils, Pappus' theorem and polynomial interpolation, *Math. Medley* **20**, 68–78, 1992.

[34] G. G. Lorentz. *Bernstein Polynomials. Second edition*, Chelsea, New York, 1986.

[35] G. G. Lorentz. *Approximation of Functions*, Holt, Rinehart and Winston, New York, 1966.

[36] F. W. Luttmann and T. J. Rivlin. Some numerical experiments in the theory of polynomial interpolation, *IBM J. Res. Develop.* **9**, 187–191, 1965.

[37] E. A. Maxwell. *The Methods of Projective Geometry based on the Use of General Homogeneous Coordinates*, Cambridge University Press, 1957.

[38] E. A. Maxwell. *General Homogeneous Coordinates in Spaces of Three Dimensions*, Cambridge University Press, 1959.

[39] Halil Oruç. *Generalized Bernstein Polynomials and Total Positivity*, Ph.D. Thesis, School of Mathematical and Computational Sciences, University of St. Andrews, 1998.

[40] Halil Oruç and G. M. Phillips. A generalization of the Bernstein polynomials, *Proc. Edin. Math. Soc.* **42**, 403–413, 1999.

[41] Halil Oruç and G. M. Phillips. Explicit factorization of the Vandermonde matrix, *Linear Algebra and Applications* **315**, 113–123, 2000.

[42] G. M. Phillips. A de Casteljau algorithm for generalized Bernstein polynomials, *BIT* **36**, 232–236, 1996.

[43] G. M. Phillips. Bernstein polynomials based on the q-integers, *The heritage of P. L. Chebyshev: a Festschrift in honor of the 70th birthday of T. J. Rivlin, Annals of Numerical Math.* **1–4**, 511–518, 1997.

[44] G. M. Phillips. *Two Millennia of Mathematics. From Archimedes to Gauss*, Springer-Verlag, New York, 2000.

[45] G. M. Phillips and P. J. Taylor. *Theory and Applications of Numerical Analysis, 2nd Edition*, Academic Press, London, 1996.

[46] M. J. D. Powell. On the maximum errors of polynomial approximations defined by interpolation and by least squares criteria, *Comput. J.* **79**, 404–407, 1967.

[47] M. J. D. Powell. *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1981.

[48] T. J. Rivlin. *An Introduction to the Approximation of Functions*, Dover, New York, 1981.

[49] T. J. Rivlin. *The Chebyshev Polynomials*, John Wiley & Sons, New York, 1974.

[50] I. J. Schoenberg. On polynomial interpolation at the points of a geometric progression, *Proc. Roy. Soc. Edin.* **90A**, 195–207, 1981.

[51] B. Sendov and V. A. Popov. *The Averaged Moduli of Smoothness*, John Wiley & Sons, Chichester, 1988.

[52] D. D. Stancu. The remainder of certain linear approximation formulas in two variables, *J. Soc. Indust. Appl. Math. Ser. B Numer. Anal.* **1**, 137–163, 1964.

[53] A. H. Stroud. *Approximate Calculation of Multiple Integrals*, Prentice–Hall, Englewood Cliffs, New Jersey, 1971.

[54] H. W. Turnbull. *James Gregory Tercentenary Memorial Volume*, Published for the Royal Society of Edinburgh by G. Bell & Sons, London, 1939.

[55] K. Weierstrass. Über die analytische Darstellbarkeit sogenannter willkürlicher Funktionen einer reelen Veränderlichen, *Sitzungsberichte der Akademie zu Berlin*, 633–639, 789–805, 1885.

Index