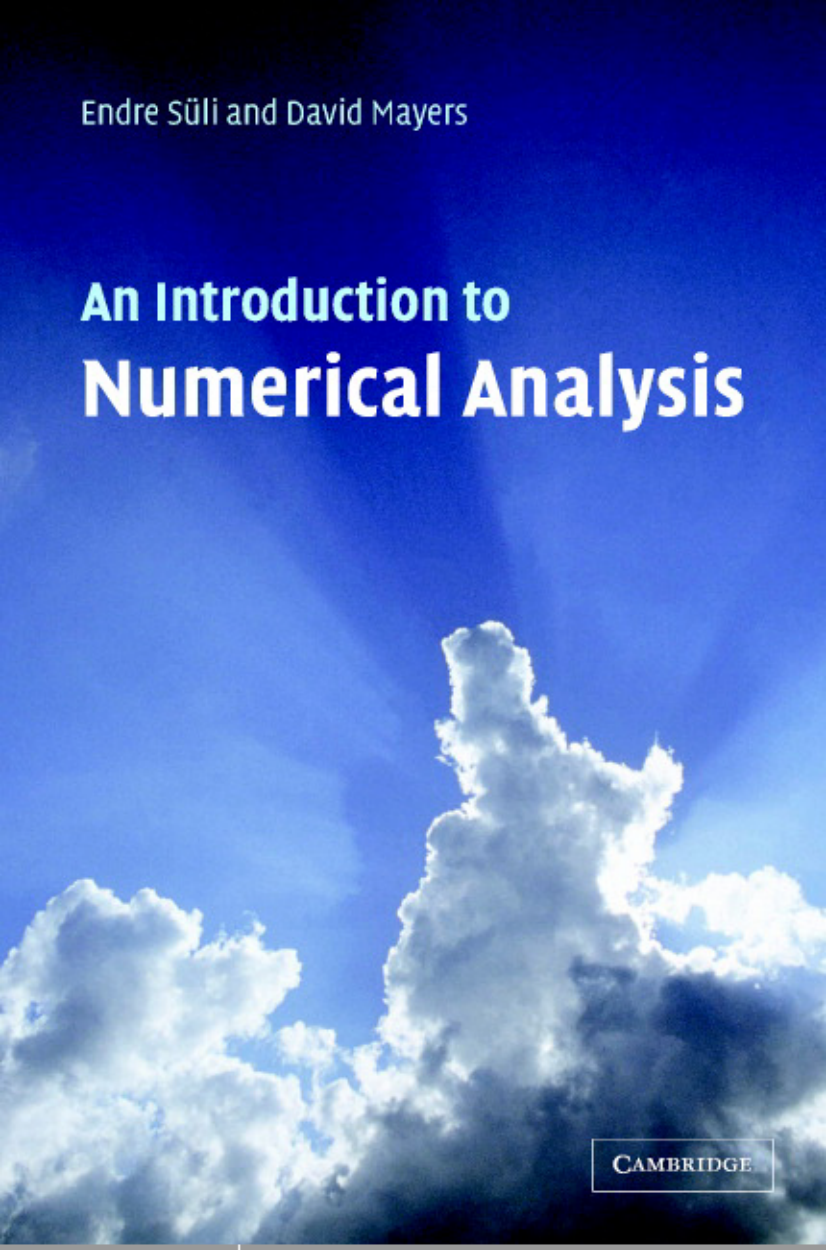
Endre Süli and David Mayers

An Introduction to Numerical Analysis



CAMBRIDGE

CAMBRIDGE

more information - www.cambridge.org/9780521810265

This page intentionally left blank

An Introduction to Numerical Analysis

Endre Süli and David F. Mayers

University of Oxford

CAMBRIDGE
UNIVERSITY PRESS

CAMBRIDGE UNIVERSITY PRESS
Cambridge, New York, Melbourne, Madrid, Cape Town, Singapore, São Paulo

Cambridge University Press
The Edinburgh Building, Cambridge CB2 2RU, United Kingdom

Published in the United States of America by Cambridge University Press, New York

www.cambridge.org

Information on this title: www.cambridge.org/9780521810265

First published in print format 2003

ISBN-13 978-0-511-07653-4 eBook (EBL)
ISBN-10 0-511-07653-3 eBook (EBL)

ISBN-13 978-0-521-81026-5 hardback
ISBN-10 0-521-81026-4 hardback

ISBN-13 978-0-521-00794-8 paperback
ISBN-10 0-521-00794-1 paperback

Contents

Preface

This book has grown out of printed notes which accompanied lectures given by ourselves and our colleagues over many years to undergraduate mathematicians at Oxford. During those years the contents and the arrangement of the lectures have changed substantially, and this book has a wider scope than is currently taught. It contains mathematics which, in an ideal world, would be part of the equipment of any well-educated mathematician.

Numerical analysis is the branch of mathematics concerned with the theoretical foundations of numerical algorithms for the solution of problems arising in scientific applications. The subject addresses a variety of questions ranging from the approximation of functions and integrals to the approximate solution of algebraic, transcendental, differential and integral equations, with particular emphasis on the stability, accuracy, efficiency and reliability of numerical algorithms. The purpose of this book is to provide an elementary introduction into this active and exciting field, and is aimed at students in the second year of a university mathematics course.

The book addresses a wide range of numerical problems in algebra and analysis. Chapter 2 deals with the solution of systems of linear equations, a process which can be completed in a finite number of arithmetical operations. In the rest of the book the solution of a problem is sought as the limit of an infinite sequence; in that sense the output of the numerical algorithm is an 'approximate' solution. This need not, however, mean any relaxation of the usual standards of rigorous analysis. The idea of convergence of a sequence of real numbers (x_n) to a real number ξ is very familiar: given any positive value of ε there exists a positive integer N such that $|x_n - \xi| < \varepsilon$ for all n such that $n > N$. In such a situation one can obtain as accurate an approximation to ξ as

required by calculating sufficiently many members of the sequence, or just one member, sufficiently far along. A 'pure mathematician' would prefer the exact answer, ξ, but the sorts of guaranteed accurate approximations which will be discussed here are entirely satisfactory in real-life applications.

Numerical analysis brings two new ideas to the usual discussion of convergence of sequences. First, we need, not just the existence of N, but a good estimate of how large it is; and it may be too large for practical calculations. Second, rather than being asked for the limit of a given sequence, we are usually given the existence of the limit ξ (or its approximate location on the real line) and then have to construct a sequence which converges to it. If the rate of convergence is slow, so that the value of N is large, we must then try to construct a better sequence, one that converges to ξ more rapidly. These ideas have direct applications in the solution of a single nonlinear equation in Chapter 1, the solution of systems of nonlinear equations in Chapter 4 and the calculation of the eigenvalues and eigenvectors of a matrix in Chapter 5.

The next six chapters are concerned with polynomial approximation, and show how, in various ways, we can construct a polynomial which approximates, as accurately as required, a given continuous function. These ideas have an obvious application in the evaluation of integrals, where we calculate the integral of the approximating polynomial instead of the integral of the given function.

Finally, Chapters 12 to 14 deal with the numerical solution of ordinary differential equations, with Chapter 14 presenting the fundamentals of the finite element method. The results of Chapter 14 can be readily extended to linear second-order partial differential equations.

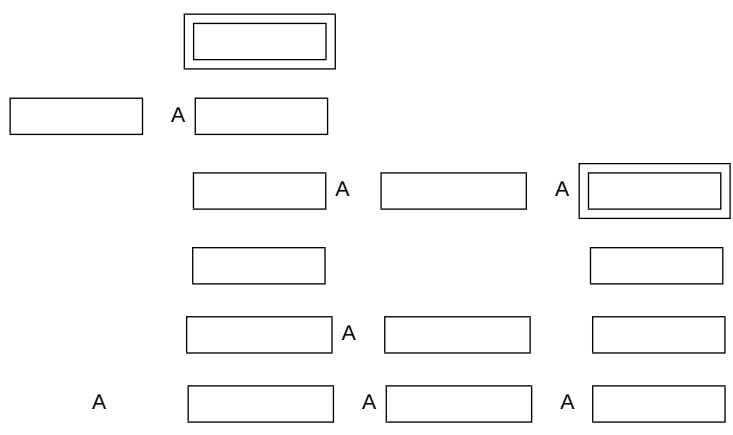
We have tried to make the coverage as complete as is consistent with remaining quite elementary. The limitations of size are most obvious in Chapter 12 on the solution of initial value problems for ordinary differential equations. This is an area where a number of excellent books are available, at least one of which is published in two weighty volumes. Chapter 12 does not describe or analyse anything approaching all the available methods, but we hope we have included some of those in most common use.

There is a selection of Exercises at the end of each chapter. All these exercises are theoretical; students are urged to apply all the methods described to some simple examples to see what happens. A few of the exercises will be found to require some heavy algebraic manipulation; these have been included because we assume that readers will have ac-

cess to some computer algebra system such as Maple or Mathematica, which then make the algebraic work almost trivial. Those involved in teaching courses based on this book may obtain copies of LaTeX files containing solutions to these exercises by applying to the publisher by email ( ). Although the material presented in this book does not presuppose the reader's acquaintance with mathematical software packages, the importance of these cannot be overemphasised. In Appendix B, a brief set of pointers is provided to relevant software repositories.

Our treatment is intended to maintain a reasonably high standard of rigour, with many theorems and formal proofs. The main prerequisite is therefore some familiarity with elementary real analysis. Appendix A lists the standard theorems (labelled **Theorem A.1**, **A.2**, . . . , **A7**) which are used in the book, together with proofs of one or two of them which might be less familiar. Some knowledge of basic matrix algebra is assumed. We have also used some elementary ideas from the theory of normed linear spaces in a number of places; complete definitions and examples are given. Some prior knowledge of these areas would be helpful, although not essential.

The chart below indicates how the chapters of the book are interrelated. They show, in particular, how Chapters 1 to 5 form a largely self-contained unit, as do Chapters 6 to 10.



We have included some historical notes throughout the book. As well as hoping to stimulate an interest in the development of the subject, these notes show how wide a historical range even this elementary book covers. Many of the methods were developed by the great mathematicians of the seventeenth and eighteenth centuries, including Newton, Euler and Gauss, but what is usually known as Gaussian elimination for the solution of systems of linear equations was known to the Chinese two thousand years ago. At the other end of the historical scale, the analysis of the eigenvalue problem, and the numerical solution of differential equations, are much more recent, and are due to mathematicians who are still very much alive. Many of our historical notes are based on the excellent biographical database at the history of mathematics website

We have tried to eradicate as many typographical errors from the text as possible; however, we are mindful that some may have escaped our attention. We plan to post any typos reported to us on

We wish to express our gratitude to Professor Bill Morton for setting us off on this *tour de force*, to David Tranah at Cambridge University Press for encouraging us to persist with the project, and to the staff of the Press for not only improving the appearance of the book and eliminating a number of typographical errors, but also for correcting and improving some of our mathematics. We also wish to thank our colleagues at the Oxford University Computing Laboratory, particularly Nick Trefethen, Mike Giles and Andy Wathen, for keeping our spirits up, and to Paul Houston at the Department of Mathematics and Computer Science of the University of Leicester for his help with the final example in the book.

Above all, we are grateful to our families for their patience, support and understanding: this book is dedicated to them.

ES & DFM *Oxford, September 2002.*

1

Solution of equations by iteration

1.1 Introduction

Equations of various kinds arise in a range of physical applications and a substantial body of mathematical research is devoted to their study. Some equations are rather simple: in the early days of our mathematical education we all encountered the single *linear* equation $ax+b=0$, where a and b are real numbers and $a \neq 0$, whose solution is given by the formula $x = -b/a$. Many equations, however, are *nonlinear:* a simple example is $ax^2+bx+c=0$, involving a quadratic polynomial with real coefficients a, b, c, and $a \neq 0$. The two solutions to this equation, labelled x_+ and x_-, are found in terms of the coefficients of the polynomial from the familiar formulae

$$x_+ = \frac{-b+\sqrt{b^2-4ac}}{2a}, \qquad x_- = \frac{-b-\sqrt{b^2-4ac}}{2a}. \tag{1.1}$$

It is less likely that you have seen the more intricate formulae for the solution of cubic and quartic polynomial equations due to the sixteenth century Italian mathematicians Niccolo Fontana Tartaglia (1499–1557) and Lodovico Ferrari (1522–1565), respectively, which were published by Girolamo Cardano (1501–1576) in 1545 in his *Artis magnae sive de regulis algebraicis liber unus.* In any case, if you have been led to believe that similar expressions involving radicals (roots of sums of products of coefficients) will supply the solution to any polynomial equation, then you should brace yourself for a surprise: no such closed formula exists for a general polynomial equation of degree n when $n \geq 5$. It transpires that for each $n \geq 5$ there exists a polynomial equation of degree n with

Theorem 1.1 *Let* f *be a real-valued function, defined and continuous on a bounded closed interval* [a, b] *of the real line. Assume, further, that* f(a)f(b) 0*; then, there exists* ξ *in* [a, b] *such that* f(ξ) = 0.

Proof If f(a) = 0 or f(b) = 0, then ξ = a or ξ = b, respectively, and the proof is complete. Now, suppose that f(a)f(b) = 0. Then, f(a)f(b) < 0; in other words, 0 belongs to the open interval whose endpoints are f(a) and f(b). By the Intermediate Value Theorem (Theorem A.1), there exists ξ in the open interval (a, b) such that f(ξ) = 0. □

To paraphrase Theorem 1.1, if a continuous function f has opposite signs at the endpoints of the interval [a, b], then the equation f(x) = 0 has a solution in (a, b). The converse statement is, of course, false. Consider, for example, a continuous function defined on [a, b] which changes sign in the open interval (a, b) an even number of times, with f(a)f(b) = 0; then, f(a)f(b) > 0 even though f(x) = 0 has solutions inside [a, b]. Of course, in the latter case, there exist an even number of subintervals of (a, b) at the endpoints of each of which f *does* have opposite signs. However, finding such subintervals may not always be easy.

To illustrate this last point, consider the rather pathological function

$$f: x \quad \frac{1}{2} \quad \frac{1}{1 + M\ x \quad 1.05}, \tag{1.2}$$

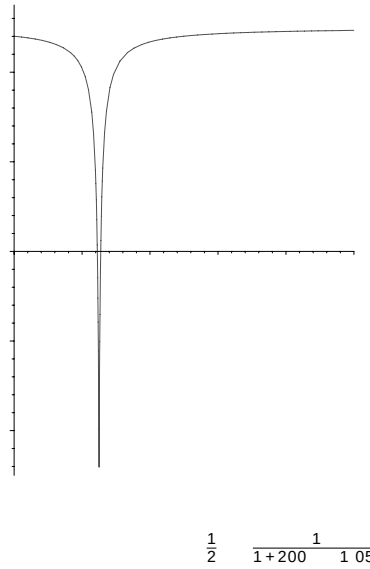
depicted in Figure 1.1 for x in the closed interval [0.8, 1.8] and M = 200. The solutions x = 1.05 (1/M) and x = 1.05 + (1/M) to the equation f(x) = 0 are only a distance 2/M apart and, for large and positive M, locating them computationally will be a challenging task.

Remark 1.1 *If you have access to the mathematical software package Maple, plot the function* f *by typing*

!""# $ % %# $! &'"

at the Maple command line, and then repeat this experiment by choosing M = 2000, 20000, 200000, 2000000, *and* 20000000 *in place of the number* 200. *What do you observe? For the last two values of* M*, replot the function* f *for* x *in the subinterval* [1.04999, 1.05001].

An alternative sufficient condition for the existence of a solution to the equation f(x) = 0 is arrived at by rewriting it in the equivalent form x g(x) = 0 where g is a certain real-valued function, defined



$\frac{1}{2}$ $\frac{1}{1+200\quad 1\ 05}$

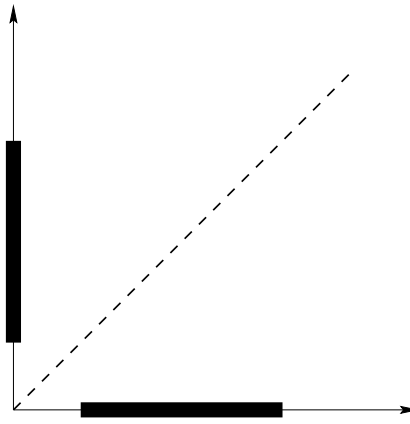
and continuous on $[a, b]$; the choice of g and its relationship with f will be clarified below through examples. Upon such a transformation the problem of solving the equation $f(x) = 0$ is converted into one of finding ξ such that $\xi \quad g(\xi) = 0$.

Theorem 1.2 (Brouwer's Fixed Point Theorem) *Suppose that* g *is a real-valued function, defined and continuous on a bounded closed interval* $[a, b]$ *of the real line, and let* $g(x) \quad [a, b]$ *for all* $x \quad [a, b]$. *Then, there exists* ξ *in* $[a, b]$ *such that* $\xi = g(\xi)$*; the real number* ξ *is called a* **fixed point of the function** g.

Proof Let $f(x) = x \quad g(x)$. Then, $f(a) = a \quad g(a) \quad 0$ since $g(a) \quad [a, b]$ and $f(b) = b \quad g(b) \quad 0$ since $g(b) \quad [a, b]$. Consequently, $f(a) f(b) \quad 0$, with f defined and continuous on the closed interval $[a, b]$. By Theorem 1.1 there exists $\xi \quad [a, b]$ such that $0 = f(\xi) = \xi \quad g(\xi)$. □

Figure 1.2 depicts the graph of a function $x \quad g(x)$, defined and continuous on a closed interval $[a, b]$ of the real line, such that $g(x)$ belongs to $[a, b]$ for all x in $[a, b]$. The function g has three fixed points in the interval $[a, b]$: the x-coordinates of the three points of intersection of the graph of g with the straight line $y = x$.

Of course, any equation of the form $f(x) = 0$ can be rewritten in the



Although the ability to verify the existence of a solution to the equation $f(x) = 0$ is important, none of what has been said so far provides a *method* for solving this equation. The following definition is a first step in this direction: it will lead to the construction of an algorithm for computing an approximation to the fixed point ξ of the function g, and will thereby supply an approximate solution to the equivalent equation $f(x) = 0$.

Definition 1.1 *Suppose that* g *is a real-valued function, defined and continuous on a bounded closed interval* $[a, b]$ *of the real line, and assume that* $g(x) \quad [a, b]$ *for all* $x \quad [a, b]$. *Given that* $x \quad [a, b]$, *the recursion defined by*

$$x \quad = g(x\ ), \qquad k = 0, 1, 2, \ldots , \tag{1.3}$$

is called a **simple iteration***; the numbers* $x\ , k \quad 0$, *are referred to as* **iterates**.

If the sequence $(x\ )$ defined by (1.3) converges, the limit must be a fixed point of the function g, since g is continuous on a closed interval. Indeed, writing $\xi = \lim_{\to \infty} x\ ,$ we have that

$$\xi = \lim_{\to \infty} x \quad = \lim_{\to \infty} g(x\ ) = g \quad \lim_{\to \infty} x \quad = g(\xi)\,, \tag{1.4}$$

where the second equality follows from (1.3) and the third equality is a consequence of the continuity of g.

A sufficient condition for the convergence of the sequence $(x\ )$ is provided by our next result which represents a refinement of Brouwer's Fixed Point Theorem, under the additional assumption that the mapping g is a contraction.

Definition 1.2 (Contraction) *Suppose that* g *is a real-valued function, defined and continuous on a bounded closed interval* $[a, b]$ *of the real line. Then,* g *is said to be a* **contraction** *on* $[a, b]$ *if there exists a constant* L *such that* $0 < L < 1$ *and*

$$g(x) \quad g(y) \quad L\, x \quad y \qquad x, y \quad [a, b]\,. \tag{1.5}$$

Remark 1.2 *The terminology 'contraction' stems from the fact that when (1.5) holds with* $0 < L < 1$, *the distance* $g(x) \quad g(y)$ *between the images of the points* x, y *is (at least* $1/L$ *times)* smaller *than the distance*

x y *between* x *and* y. *More generally, when* L *is any positive real number, (1.5) is referred to as a* **Lipschitz condition**.

Armed with Definition 1.2, we are now ready to state the main result of this section.

Theorem 1.3 (Contraction Mapping Theorem) *Let* g *be a real-valued function, defined and continuous on a bounded closed interval* [a, b] *of the real line, and assume that* g(x) [a, b] *for all* x [a, b]. *Suppose, further, that* g *is a contraction on* [a, b]. *Then,* g *has a unique fixed point* ξ *in the interval* [a, b]. *Moreover, the sequence* (x) *defined by (1.3) converges to* ξ *as* k *for any starting value* x *in* [a, b].

Proof The existence of a fixed point ξ for g is a consequence of Theorem 1.2. The uniqueness of this fixed point follows from (1.5) by contradiction: for suppose that g has a second fixed point, η, in [a, b]. Then,

$$\xi \quad \eta \;=\; g(\xi) \quad g(\eta) \qquad L\ \xi \quad \eta\ ,$$

i.e., (1 L) ξ η 0. As 1 L > 0, we deduce that η = ξ.

Let x be any element of [a, b] and consider the sequence (x) defined by (1.3). We shall prove that (x) converges to the fixed point ξ. According to (1.5) we have that

$$x \quad \xi \;=\; g(x \_\ ) \quad g(\xi) \qquad L\ x \_ \quad \xi\ , \qquad k \quad 1\,,$$

from which we then deduce by induction that

$$x \quad \xi \qquad L \quad x \quad \xi\ , \qquad k \quad 1\,. \tag{1.6}$$

As L (0, 1), it follows that lim → ∞ L = 0, and hence we conclude that lim → ∞ x ξ = 0. □

Let us illustrate the Contraction Mapping Theorem by an example.

Example 1.2 *Consider the equation* f(x) = 0 *on the interval* [1, 2] *with* f(x) = e 2x 1, *as in Example 1.1. Recall from Example 1.1 that this equation has a solution,* ξ, *in the interval* [1, 2], *and* ξ *is a fixed point of the function* g *defined on* [1, 2] *by* g(x) = ln(2x + 1).

1 1 H : , + B# ' 6 #.* I B $
D C & H, #@ * F 6D ,
6 6 1 1 , 9 6
1 6 J G 6 , , ,
64

Table 1.1. *The sequence* (x_k) *defined by (1.8).*

k	x_k
0	1.00000
1	1.09861
2	1.19392
3	1.22668
4	1.25177
5	1.25885
6	1.26078
7	1.26130
8	1.26144
9	1.26148
10	1.26149

Now, the function g is defined and continuous on the interval $[1,2]$, and g is differentiable on $(1,2)$. Thus, by the Mean Value Theorem (Theorem A.3), for any x, y in $[1,2]$ we have that

$$g(x) - g(y) = g'(\eta)(x - y) = g'(\eta)\,x - y \tag{1.7}$$

for some η that lies between x and y and is therefore in the interval $[1,2]$. Further, $g'(x) = 2/(2x+1)$ and $g''(x) = -4/(2x+1)^2$. As $g''(x) < 0$ for all x in $[1,2]$, g' is monotonic decreasing on $[1,2]$. Hence $g'(1) \geq g'(\eta) \geq g'(2)$, *i.e.*, $g'(\eta) \in [2/5, 2/3]$. Thus we deduce from (1.7) that

$$|g(x) - g(y)| \leq L\,|x - y| \qquad x, y \in [1,2]\,,$$

with $L = 2/3$. According to the Contraction Mapping Theorem, the sequence (x_k) defined by the simple iteration

$$x_{k+1} = \ln(2x_k + 1)\,, \qquad k = 0, 1, 2, \ldots, \tag{1.8}$$

converges to ξ for any starting value x_0 in $[1,2]$. Let us choose $x_0 = 1$, for example, and compute the next 11 iterates, say. The results are shown in Table 1.1. Even though we have carried six decimal digits, after 11 iterations only the first two decimal digits of the iterates x_k appear to have settled; thus it seems likely that $\xi = 1.26$ to two decimal digits.

You may now wonder how many iterations we should perform in (1.8)

to ensure that all six decimals have converged to their correct values. In order to answer this question, we need to carry out some analysis.

Theorem 1.4 *Consider the simple iteration (1.3) where the function* g *satisfies the hypotheses of the Contraction Mapping Theorem on the bounded closed interval* [a, b]. *Given* x [a, b] *and a certain tolerance* ε > 0, *let* k (ε) *denote the smallest positive integer such that* x *is no more than* ε *away from the (unknown) fixed point* ξ, *i.e.,* x ξ ε, *for all* k k (ε). *Then,*

$$k\ (\varepsilon) \quad \frac{\ln\ x \quad x \quad \ln(\varepsilon(1 \quad L))}{\ln(1/L)} \quad +1\,, \tag{1.9}$$

where, for a real number x, [x] *signifies the largest integer less than or equal to* x.

Proof From (1.6) in the proof of Theorem 1.3 we know that

$$x \quad \xi \quad L \quad x \quad \xi\,, \quad k \quad 1\,.$$

Using this result with k = 1, we obtain

$$\begin{aligned} x \quad \xi &= x \quad x \ + x \quad \xi \\ &\quad x \quad x \ + x \quad \xi \\ &\quad x \quad x \ + L\, x \quad \xi\,. \end{aligned}$$

Hence

$$x \quad \xi \quad \frac{1}{1 \quad L}\, x \quad x\ .$$

By substituting this into (1.6) we get

$$x \quad \xi \quad \frac{L}{1 \quad L}\, x \quad x\ . \tag{1.10}$$

Thus, in particular, x ξ ε provided that

$$L\ \frac{1}{1 \quad L}\ x \quad x \quad \varepsilon\,.$$

On taking the (natural) logarithm of each side in the last inequality, we find that x ξ ε for all k such that

$$k \quad \frac{\ln\ x \quad x \quad \ln(\varepsilon(1 \quad L))}{\ln(1/L)}\,.$$

Therefore, the smallest integer k (ε) such that x ξ ε for all

$k \quad k(\varepsilon)$ cannot exceed the expression on the right-hand side of the inequality (1.9). □

This result provides an upper bound on the maximum number of iterations required to ensure that the error between the kth iterate x and the (unknown) fixed point ξ is below the prescribed tolerance ε. Note, in particular, from (1.9), that if L is close to 1, then $k(\varepsilon)$ may be quite large for any fixed ε. We shall revisit this point later on in the chapter.

Example 1.3 *Now we can return to Example 1.2 to answer the question posed there about the maximum number of iterations required, with starting value* $x = 1$*, to ensure that the last iterate computed is correct to six decimal digits.*

Letting $\varepsilon = 0.5 \quad 10^{-}$ and recalling from Example 1.2 that $L = 2/3$, the formula (1.9) yields $k(\varepsilon) \quad [32.778918] + 1$, so we have that $k(\varepsilon) \quad 33$. In fact, 33 is a somewhat pessimistic overestimate of the number of iterations required: computing the iterates x successively shows that already x is correct to six decimal digits, giving $\xi = 1.256431$.

Condition (1.5) can be rewritten in the following equivalent form:

$$\frac{g(x) \quad g(y)}{x \quad y} \qquad L \qquad x, y \quad [a, b], \quad x = y,$$

with $L \quad (0, 1)$, which can, in turn, be rephrased by saying that the absolute value of the slope of the function g does not exceed $L \quad (0, 1)$. Assuming that g is a differentiable function on the open interval (a, b), the Mean Value Theorem (Theorem A.3) tells us that

$$\frac{g(x) \quad g(y)}{x \quad y} = g'(\eta)$$

for some η that lies between x and y and is therefore contained in the interval (a, b).

We shall therefore adopt the following assumption that is somewhat stronger than (1.5) but is easier to verify in practice:

$$\begin{aligned} &g \text{ is differentiable on } (a, b) \text{ and} \\ &L \quad (0, 1) \text{ such that } g'(x) \qquad L \text{ for all } x \quad (a, b). \end{aligned} \tag{1.11}$$

Consequently, Theorem 1.3 still holds when (1.5) is replaced by (1.11).

We note that the requirement in (1.11) that g be differentiable is

indeed more demanding than the Lipschitz condition (1.5): for example, $\mathtt{g}(\mathtt{x}) = \mathtt{x}$

If the conditions of Theorem 1.5 are satisfied in the vicinity of a fixed point ξ, then the sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, will converge to ξ for any starting value x_0 that is sufficiently close to ξ. If, on the other hand, the conditions of Theorem 1.5 are violated, there is no guarantee that any sequence (x_k) defined by the iteration $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, will converge to the fixed point ξ for any starting value x_0 near ξ. In order to distinguish between these two cases, we introduce the following definition.

Definition 1.3 *Suppose that* g *is a real-valued function, defined and continuous on the bounded closed interval* $[a, b]$*, such that* $g(x) \in [a, b]$

If (1.15) holds with $\mu = 0$*, then the sequence* (x_k) *is said to converge to* ξ **superlinearly**.

If (1.15) holds with $\mu \in (0,1)$ *and* $\varepsilon_k = |x_k - \xi|$, $k = 0, 1, 2, \ldots$, *then* (x_k) *is said to converge to* ξ **linearly**, *and the number* $\rho = -\log_{10}\mu$ *is then called the* **asymptotic rate of convergence** *of the sequence. If (1.15) holds with* $\mu = 1$ *and* $\varepsilon_k = |x_k - \xi|$, $k = 0, 1, 2, \ldots$, *the rate of convergence is slower than linear and we say that the sequence converges to* ξ **sublinearly**.

The words 'at least' in this definition refer to the fact that we only have inequality in $|x_k - \xi| \le \varepsilon_k$, which may be all that can be ascertained in practice. Thus, it is really the sequence of bounds ε_k that converges linearly.

For a linearly convergent sequence the asymptotic rate of convergence ρ measures the number of correct decimal digits gained in one iteration; in particular, the number of iterations required in order to gain one more correct decimal digit is at most $[1/\rho] + 1$. Here $[1/\rho]$ denotes the largest integer that is less than or equal to $1/\rho$.

Under the hypotheses of Theorem 1.5, the equalities (1.14) will hold with $\mu = g'(\xi) \in [0, 1)$, and therefore the sequence (x_k) generated by the simple iteration will converge to the fixed point ξ linearly or superlinearly.

Example 1.4 *Given that* α *is a fixed positive real number, consider the function* g *defined on the interval* $[0, 1]$ *by*

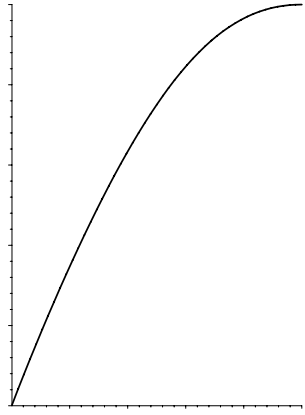
$$g(x) = \begin{cases} 2^{-} & \text{for } 0 < x \le 1, \\ 0 & \text{for } x = 0. \end{cases}$$

As $\lim_{x \to 0} g(x) = 0$, the function g is continuous on $[0, 1]$. Moreover, g is strictly monotonic increasing on $[0, 1]$ and $g(x) \in [0, 1/2] \subset [0, 1]$ for all x in $[0, 1]$. We note that $\xi = 0$ is a fixed point of g (cf. Figure 1.3).

Consider the sequence (x_k) defined by $x_{k+1} = g(x_k)$, $k \ge 0$, with $x_0 = 1$. It is a simple matter to show by induction that $x_k = 2^{-k}$, $k \ge 0$. Thus we deduce that (x_k) converges to $\xi = 0$ as $k \to \infty$. Since

$$\lim_{k \to \infty} \frac{x_{k+1}}{x_k^{\alpha}} = \mu = \begin{cases} 1 & \text{for } 0 < \alpha < 1, \\ - & \text{for } \alpha = 1, \\ 0 & \text{for } \alpha > 1, \end{cases}$$

we conclude that for $\alpha \in (0, 1)$ the sequence (x_k) converges to $\xi = 0$ sublinearly. For $\alpha = 1$ it converges to $\xi = 0$ linearly with asymptotic rate



of the sequence (x_k) is $\rho = -\log |g'(\xi)|$. Evidently, a small value of $|g'(\xi)|$ corresponds to a large positive value of ρ and will result in more rapid convergence, while if $|g'(\xi)| < 1$ but $|g'(\xi)|$ is very close to 1, ρ will be a small positive number and the sequence will converge very slowly.

Next, we discuss the behaviour of the iteration (1.3) in the vicinity of an *unstable fixed point* ξ. If $|g'(\xi)| > 1$, then the sequence (x_k) defined by (1.3) does not converge to ξ from any starting value x_0; the next theorem gives a rigorous proof of this fact.

Theorem 1.6 *Suppose that $\xi = g(\xi)$, where the function g has a continuous derivative in some neighbourhood of ξ, and let $|g'(\xi)| > 1$. Then, the sequence (x_k) defined by $x_{k+1} = g(x_k)$, $k = 0, 1, 2, \ldots$, does not converge to ξ from any starting value x_0, $x_0 \neq \xi$.*

Proof Suppose that $x_k = \xi$. As in the proof of Theorem 1.5, we can see that there is an interval $I = [\xi - \delta, \xi + \delta]$, $\delta > 0$, in which $|g'(x)| \geq L > 1$ for some constant L. If x_k lies in this interval, then

$$|x_{k+1} - \xi| = |g(x_k) - g(\xi)| = |(x_k - \xi)g'(\eta_k)| \geq L|x_k - \xi|,$$

for some η_k between x_k and ξ. If x_{k+1} lies in I the same argument shows that

$$|x_{k+2} - \xi| \geq L|x_{k+1} - \xi| \geq L^2|x_k - \xi|,$$

and so on. Evidently, after a finite number of steps some member of the sequence $x_k, x_{k+1}, x_{k+2}, \ldots$ must be outside the interval I, since $L > 1$. Hence there can be no value of $K = K(\delta)$ such that $|x_k - \xi| \leq \delta$ for all $k \geq K$, and the sequence therefore does not converge to ξ. □

Example 1.5 *In this example we explore the simple iteration (1.3) for g defined by*

$$g(x) = -(x^2 + c)$$

where c is a fixed constant.

The fixed points of the function g are the solutions of the quadratic equation $x^2 + x + c = 0$, which are $-\frac{1}{2}(1 \pm \sqrt{1 - 4c})$. If $c > \frac{1}{4}$ there are no solutions (in the set $\mathbb{R}$ of real numbers, that is!), if $c = \frac{1}{4}$ there is one solution in $\mathbb{R}$, and if $c < \frac{1}{4}$ there are two.

Suppose now that `c <`

Table 1.2. *The sequences* (u) *and* (v) *in Example 1.6.*

$$	$))
$	)
& $) $$	
' $% ($%	
($($$	
% $' '	
) $& %(	
$)'(	
$ $ &( )	
$ '&	

is still larger than 0.9. Although (u) eventually converges faster than v , we find that u = (0.99) becomes smaller than v = (k + 1)⁻ when

$$k > \frac{10}{\ln(1/0.99)} \ln(k+1)\,.$$

This first happens when k = 9067, at which point u and v are both roughly 10⁻ . In this rather extreme example the concept of asymptotic rate of convergence is not useful, since for any practical purposes (v) converges faster than (u).

1.3 Iterative solution of equations

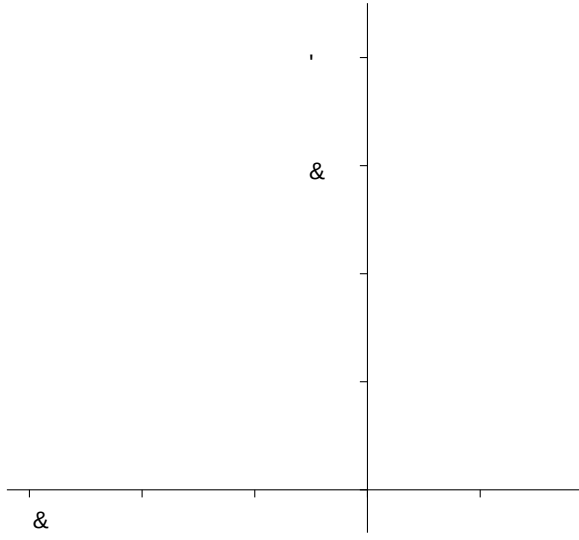
In this section we apply the idea of simple iteration to the solution of equations. Given a real-valued continuous function f, we wish to construct a sequence (x), using iteration, which converges to a solution of f(x) = 0. We begin with an example where it is easy to derive various such sequences; in the next section we shall describe a more general approach.

Example 1.7 *Consider the problem of determining the solutions of the equation* f(x) = 0*, where* f: x e x 2*.*

Since f′(x) = e 1 the function f is monotonic increasing for positive x and monotonic decreasing for negative values of x. Moreover,

$$\begin{aligned} f(1) &= e \quad 3 < 0, \\ f(2) &= e \quad 4 > 0, \\ f(\ 1) &= e^{-} \quad 1 < 0, \\ f(\ 2) &= e^{-} \ > 0. \end{aligned} \tag{1.16}$$

Hence the equation $f(x) = 0$ has exactly one positive solution, which lies in the interval $(1, 2)$, and exactly one negative solution, which lies in the interval $(\ 2,\ 1)$. This is illustrated in Figure 1.4, which shows the graphs of the functions $x \quad e$ and $x \quad x + 2$ on the same axes. We shall write ξ for the positive solution and ξ for the negative solution.



will converge to the positive solution, ξ, provided that the starting value x is sufficiently close to it. As $0 < g'(\xi) < 1/3$, the asymptotic rate of convergence of (x) to ξ is certainly greater than $\log 3$.

On the other hand, $g'(\xi) > 1$ since $2 < \xi < 1$, so the sequence (x) defined by (1.17) cannot converge to the solution ξ. It is not difficult to prove that for $x > \xi$ the sequence (x) converges to ξ while if $x < \xi$ the sequence will decrease monotonically until $x \quad 2$ for some k, and then the iteration breaks down as $g(x)$ becomes undefined.

The equation $f(x) = 0$ may also be written in the form $x = e \quad 2$, suggesting the sequence (x) defined by the iteration

$$x \quad = e \quad 2, \qquad k = 0, 1, 2, \ldots .$$

In this case $g(x) = e \quad 2$ and $g'(x) = e$. Hence $g'(\xi) > 1$, $g'(\xi) < e^{-}$, showing that the sequence (x) may converge to ξ, but cannot converge to ξ. It is quite straightforward to show that the sequence converges to ξ for any $x < \xi$, but diverges to $+$ when $x > \xi$.

As a third alternative, consider rewriting the equation $f(x) = 0$ as $x = g(x)$ where the function g is defined by $g(x) = x(e \quad x)/2$; the fixed points of the associated iteration $x \quad = g(x)$ are the solutions ξ and ξ of $f(x) = 0$, and also the point 0. For this iteration neither of the fixed points, ξ or ξ, is stable, and the sequence (x) either converges to 0 or diverges to .

Evidently the given equation may be written in many different forms, leading to iterations with different properties.

1.4 Relaxation and Newton's method

In the previous section we saw how various ingenious devices lead to iterations which may or may not converge to the desired solutions of a given equation $f(x) = 0$. We would obviously benefit from a more generally applicable iterative method which would, except possibly in special cases, produce a sequence (x) that always converges to a required solution. One way of constructing such a sequence is by relaxation.

[illegible]

Definition 1.5 *Suppose that* f *is a real-valued function, defined and continuous in a neighbourhood of a real number* ξ. **Relaxation** *uses the sequence* (x_k) *defined by*

$$x_{k+1} = x_k - \lambda f(x_k), \qquad k = 0, 1, 2, \ldots, \tag{1.18}$$

where

corresponding to

$$\lambda = \frac{4}{2M + \alpha}.$$

On defining $g(x) = x - \lambda f(x)$, we then deduce that

$$g'(x) \leq \vartheta < 1, \quad x \in [\xi - \delta, \xi + \delta]. \tag{1.19}$$

Thus we can apply Theorem 1.5 to conclude that the sequence (x_k) defined by the relaxation iteration (1.18) converges to ξ, provided that x_0 is in the interval $[\xi - \delta, \xi + \delta]$. The asymptotic rate of convergence of the relaxation iteration (1.18) to ξ is at least $-\log \vartheta$. □

We can now extend the idea of relaxation by allowing λ to be a continuous function of x in a neighbourhood of ξ rather than just a constant. This suggests an iteration

$$x_{k+1} = x_k - \lambda(x_k) f(x_k), \qquad k = 0, 1, 2, \ldots,$$

corresponding to a simple iteration with $g(x) = x - \lambda(x) f(x)$. If the sequence (x_k) converges, the limit ξ will be a solution of $f(x) = 0$, except possibly when $\lambda(\xi) = 0$. Moreover, as we have seen, the ultimate rate of convergence is determined by $g'(\xi)$. Since $f(\xi) = 0$, it follows that $g'(\xi) = 1 - \lambda(\xi) f'(\xi)$, and (1.19) suggest using a function λ which makes $1 - \lambda(\xi) f'(\xi)$ small. The obvious choice is $\lambda(x) = 1/f'(x)$, and leads us to Newton's method.

Definition 1.6 *Newton's method for the solution of* $f(x) = 0$ *is defined by*

$$x_{k+1} = x_k - \frac{f(x_k)}{f'(x_k)}, \qquad k = 0, 1, 2, \ldots, \tag{1.20}$$

with prescribed starting value x_0. *We implicitly assume in the defining formula (1.20) that* $f'(x_k) \neq 0$ *for all* $k \geq 0$.

[illegible]

Newton's method is a simple iteration with $g(x) = x - f(x)/f'(x)$. Its geometric interpretation is illustrated in Figure 1.5: the tangent to the curve $y = f(x)$ at the point $(x_n, f(x_n))$ is the line with the equation $y - f(x_n)$

We note that unlike the definition of linear convergence where μ was required to belong to the interval $(0,1)$, all we demand here is that $\mu > 0$. The reason is simple: when $q > 1$, (1.21) implies suitably rapid decay of the sequence (ε_k) irrespective of the size of μ.

Example 1.8 *Let $c > 1$ and $q > 1$. The sequence (x_k) defined by $x_k = c^{-q^k}$, $k = 0, 1, 2, \ldots$, converges to 0 with order q.*

Theorem 1.8 (Convergence of Newton's method) *Suppose that f is a continuous real-valued function with continuous second derivative f'', defined on the closed interval $I = [\xi - \delta, \xi + \delta]$, $\delta > 0$, such that $f(\xi) = 0$ and $f'(\xi) = 0$. Suppose further that there exists a positive constant A such that*

$$\frac{f''(x)}{f'(y)} \leq A \qquad x, y \in I .$$

If $\xi - x_0 \leq h$, where h is the smaller of δ and $1/A$, then the sequence (x_k) defined by Newton's method (1.20) converges quadratically to ξ.

Proof Suppose that $\xi - x_0 \leq h = \min\{\delta, 1/A\}$, so that $x_0 \in I$. Then, by Taylor's Theorem (Theorem A.4), expanding about the point $x_k \in I$,

$$0 = f(\xi) = f(x_k) + (\xi - x_k) f'(x_k) + \frac{(\xi - x_k)^2}{2} f''(\eta_k), \tag{1.22}$$

for some η_k between ξ and x_k, and therefore in the interval I. Recalling (1.20), this shows that

$$\xi - x_{k+1} = \frac{(\xi - x_k)^2 f''(\eta_k)}{2 f'(x_k)}. \tag{1.23}$$

Since $\xi - x_0 \leq -$, we have $|\xi - x_1| \leq |\xi - x_0|$. As we are given that $|\xi - x_0| \leq h$ it follows by induction that $|\xi - x_k| \leq 2^{-k} h$ for all $k \geq 0$; hence (x_k) converges to ξ as $k \to \infty$.

Now, η_k lies between ξ and x_k, and therefore (η_k) also converges to ξ as $k \to \infty$. Since f' and f'' are continuous on I, it follows from (1.23) that

$$\lim_{k \to \infty} \frac{x_{k+1} - \xi}{(x_k - \xi)^2} = \frac{f''(\xi)}{2 f'(\xi)}, \tag{1.24}$$

which, according to Definition 1.7, implies quadratic convergence of the sequence (x_k) to ξ with $\mu = |f''(\xi)/2f'(\xi)|$, $\mu \in (0, A/2]$. □

The conditions of the theorem implicitly require that $f'(\xi) = 0$, for otherwise the quantity $f''(x)/f'(y)$ could not be bounded in a neighbourhood of ξ

1.5 The secant method

So far we have considered iterations which can be written in the form $x_{k} = g(x_{k-1})$, $k \geq 0$, so that the new value is expressed in terms of the old one. It is also possible to define an iteration of the form $x_{k} = g(x_{k-1}, x_{k-2})$, $k \geq 1$, where the new value is expressed in terms of two previous values. In particular, we shall consider two applications of this idea, leading to the secant method and the method of bisection, respectively.

Remark 1.3 *We note in passing that one can consider more general iterative methods of the form*

$$x_{k} = g(x_{k-1}, x_{k-2}, \ldots, x_{k-\ell}), \qquad k = \ell, \ell+1, \ldots,$$

with $\ell \geq 1$ fixed; here, we shall confine ourselves to the simplest case when $\ell = 1$ as this is already sufficiently illuminating.

Using Newton's method to solve a nonlinear equation $f(x) = 0$ requires explicit knowledge of the first derivative f' of the function f. Unfortunately, in many practical situations f' is not explicitly available or it can only be obtained at high computational cost. In such cases, the value $f'(x_{k-1})$ in (1.20) can be approximated by a difference quotient; that is,

$$f'(x_{k-1}) \approx \frac{f(x_{k-1}) - f(x_{k-2})}{x_{k-1} - x_{k-2}}.$$

Replacing $f'(x_{k-1})$ in (1.20) by this difference quotient leads us to the following definition.

Definition 1.8 *The* **secant method** *is defined by*

$$x_{k} = x_{k-1} - f(x_{k-1}) \frac{x_{k-1} - x_{k-2}}{f(x_{k-1}) - f(x_{k-2})}, \qquad k = 1, 2, 3, \ldots, \quad (1.25)$$

where x_{0} and x_{1} are given starting values. It is implicitly assumed here that $f(x_{k-1}) - f(x_{k-2}) \neq 0$ for all $k \geq 1$.

The method is illustrated in Figure 1.6. The new iterate x_{k} is obtained from x_{k-1} and x_{k-2} by drawing the chord joining the points $P(x_{k-2}, f(x_{k-2}))$ and $Q(x_{k-1}, f(x_{k-1}))$, and using as x_{k} the point at which this chord intersects the x-axis. If x_{k-2} and x_{k-1} are close together and f

Table 1.3. *Comparison of the secant method and Newton's method for the solution of* e x 2 = 0.

1	/ 0	
	&	%&$(&
	&%%%%(	'%'
&	%" $	'% $&
'	(& $$	'% $&
(	'()'(	
%	'% $	
)	'% $&	

where ϑ is between x and ξ, and ϕ lies between x and x – . Hence, if x – I and x I , then also ϑ I and ϕ I . Therefore,

$$\xi \quad x \qquad \xi \quad x \quad 1 \quad \frac{5\alpha/4}{3\alpha/4} = - \xi \quad x \quad . \tag{1.29}$$

Thus, x I and the sequence (x) converges to ξ at least linearly, with rate at least log (3/2), provided that x I and x I . □

In fact, it can be shown that

$$\lim_{\to \infty} \frac{x \qquad \xi}{x \qquad \xi} = \mu \tag{1.30}$$

where μ is a positive constant and q = –(1 + 5) 1.6, so that the convergence of the sequence (x) to ξ is faster than linear, but not as fast as quadratic. (See Exercise 10.)

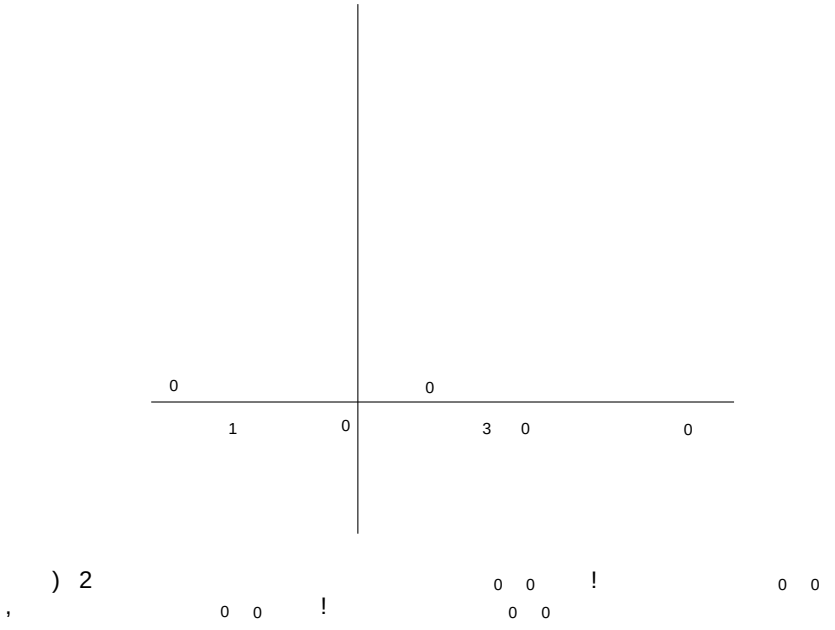
This is illustrated in Table 1.3, which compares two iterative methods for the solution of f(x) = 0 with f: x e x 2; the first is the secant method, starting from x = 1, x = 3, while the second is Newton's method starting from x = 1.

This experiment shows the faster convergence of Newton's method, but it must be remembered that each iteration of Newton's method requires the calculation of both f(x) and f′(x), while each iteration of the secant method requires the calculation of f(x) only (as f(x –) has already been computed). In our examples the computations are quite trivial, but in a practical situation the calculation of each value of f(x) and f′(x) may demand a substantial amount of work, and then

each iteration of Newton's method is likely to involve at least twice as much work as one iteration of the secant method.

1.6 The bisection method

Suppose that f is a real-valued function defined and continuous on a bounded closed interval $[a, b]$ of the real line and such that $f(\xi) = 0$ for some $\xi \quad [a, b]$. A very simple iterative method for the solution of the nonlinear equation $f(x) = 0$ can be constructed by beginning with an interval $[a$



$(b - a)/2$. The bisection method is therefore very robust, though Newton's method will always win once the current iterate is sufficiently close to ξ.

If the initial interval $[a, b]$ contains more than one solution, the limit of the bisection method will depend on the positions of these solutions. Figure 1.7 illustrates a possible situation, where $[a, b]$ contains three solutions. Since $f(c)$ has the same sign as $f(b)$ the second interval is $[a, c]$, and the sequence (c) of midpoints defined by (1.31) converges to the solution ξ . If however the initial interval is $[a, b']$ the sequence of midpoints converges to the solution ξ .

1.7 Global behaviour

We have already seen how an iteration will often converge to a limit if the starting value is sufficiently close to that limit. The behaviour of the iteration, when started from an arbitrary starting value, can be very complicated. In this section we shall consider two examples. No theorems will be stated: our aim is simply to illustrate various kinds of behaviour.

First consider the simple iteration defined by

$$x_{k+1} = g(x_k), \quad k = 0, 1, 2, \ldots, \qquad \text{where } g(x) = ax(1 - x), \quad (1.33)$$

which is often known as the **logistic equation**. We require the constant a to lie in the range $0 < a \leq 4$, for then if the starting value x_0 is in the interval $[0, 1]$, then all members of the sequence (x_k) also lie in $[0, 1]$. The function g has two fixed points: $x = 0$ and $x = 1 - 1/a$. The fixed point at 0 is stable if $0 < a < 1$, and the fixed point at $1 - 1/a$ is stable if $1 < a < 3$. The behaviour of the iteration for these values of a is what might be expected from this information, but for larger values of the parameter a the behaviour of the sequence (x_k) becomes increasingly complicated.

For example, when $a = 3.4$ there is no stable fixed point, and from any starting point the sequence eventually oscillates between two values, which are 0.45 and 0.84 to two decimal digits. These are the two stable fixed points of the double iteration

$$x_{k+1} = g^{(2)}(x_k), \qquad g^{(2)}(x) = g(g(x)) = a^2 x(1 - x)[1 - ax(1 - x)]. \quad (1.34)$$

When $3 < a < 1 + \sqrt{6}$, the fixed points of $g^{(2)}$ are the two fixed points of g, that is 0 and $1 - 1/a$, and also

$$\frac{1}{2}\left(1 + \frac{1}{a}\right) \pm \frac{1}{2a}\sqrt{a^2 - 2a - 3}. \quad (1.35)$$

This behaviour is known as a stable two-cycle (see Exercise 12).

When $a > 1 + \sqrt{6}$ all the fixed points of $g^{(2)}$ are unstable. For example, when $a = 3.5$ all sequences (x_k) defined by (1.33) tend to a stable 4-cycle, taking successive values 0.50, 0.87, 0.38 and 0.83.

For larger values of the parameter a the sequences become chaotic. For example, when $a = 3.99$ there are no stable fixed points or limit-cycles, and the members of any sequence appear random. In fact it can be shown that for such values of a the members of the sequence are *dense* in a subinterval of $[0, 1]$: there exist real numbers α and β, $\alpha < \beta$, such that any subinterval of (α, β), however small, contains an infinite subsequence of (x_k). For the value $a = 3.99$ the maximal interval (α, β) is $(0.00995, 0.99750)$ to five decimal digits. Starting from $x_0 = 0.75$ we find that the interval $(0.70, 0.71)$, for example, contains the subsequence

$$x_{13}, x_{14}, x_{15}, x_{16}, x_{17}, \ldots. \quad (1.36)$$

The sequence does not show any apparent regular behaviour. The calculation is extremely sensitive: if we replace x_0 by $x_0 + \delta x_0$, and write

that from any starting value Newton's method eventually converges to a solution, which might be . However, it is certainly *not* true that the sequence converges to the solution closest to the starting point; indeed, if this were true, no sequence could converge to . It is easy to see why the behaviour is much more complicated than this.

The Newton iteration converges to the solution at 0 from any point in the interval ($0.327, 0.445$). As we see from Figure 1.8, the iteration will converge exactly to 0 in one iteration if we start from the x-coordinate of any of the points a , a and a ; at each of these three points the tangent to the curve passes through the origin. Since f is continuous, this means that there is an open interval surrounding each of these points from which the Newton iteration will converge to 0. The maximal such intervals are (1.555, 1.487), $(1.735, 1.817)$ and $(3.514, 3.529)$ to three decimal digits. In the same way, there are several points at which the tangent to the curve passes through the point $(A\ , 0)$, where A is the x-coordinate of the point a . Starting from one of these points, the Newton iteration will evidently converge exactly to the solution at 0 in two steps; surrounding each of these points there is an open interval from which the iteration will converge to 0.

Now suppose we define the sets $S\ , m = 1, 0, 1, 3, \ , \ $, where S consists of those points from which the Newton iteration converges to the zero at m. Then, an extension of the above argument shows that each of the sets S is the union of an infinite number of disjoint open intervals. The remarkable property of these sets is that, if ξ is a boundary point of one of the sets S , then it is also a boundary point of all the other sets as well. This means that any neighbourhood of such a point ξ, however small, contains an infinite number of members of each of the sets S . For example, we have seen that the iteration starting from any point in the interval ($0.327, 0.445$) converges to 0. We find that the end of this interval lies between 0.4457855 and 0.4457860; Table 1.4 shows the limits of various Newton iterations starting from points near this boundary. Each of these points is, of course, itself surrounded by an open interval which gives the same limit.

1.8 Notes

Theorem 1.2 is a special case of Brouwer's Fixed Point Theorem. Luitzen Egbertus Jan Brouwer (1881–1966) was professor of set theory, function theory and axiomatics at the University of Amsterdam, and made major contributions to topology. Brouwer was a mathematical genius with

Table 1.4. *Limit of Newton's method near a boundary point.*

o	3
"() '	
"() '(	
"() (	
"() ((	
"() %	
"() %(	
"())	
"())(	
"()	
"() (	
"() $	.
"() $(	&
"()$	

strong mystical and philosophical leanings. For an historical overview of Brouwer's life and work we refer to the recent book of Dirk Van Dalen, *Mystic, Geometer, and Intuitionist. The Life of L.E.J. Brouwer: the Dawning Revolution*, Clarendon Press, Oxford, 1999.

The Contraction Mapping Theorem, as stated here, is a simplified version of Banach's fixed point theorem. Stefan Banach founded modern functional analysis and made outstanding contributions to the theory of topological vector spaces, measure theory, integration, the theory of sets, and orthogonal series. For an inspiring account of Banach's life and times, see R. Kaluza, *Through the Eyes of a Reporter: the Life of Stefan Banach*, Birkhäuser, Boston, MA, 1996.

In our definitions of linear convergence and convergence with order q, we followed Definitions 2.1 and 2.2 in Chapter 4 of

, *Numerical Analysis: an Introduction*, Birkhäuser, Boston, MA, 1997.

Exciting surveys of the history of Newton's method are available in T. Ypma, Historical development of the Newton–Raphson method, *SIAM Rev.* **37**, 531–551, 1995, H. Goldstine, *History of Numerical Analysis from the Sixteenth through the Nineteenth Century*, Springer, New York, 1977; and in Chapter 6 of Jean-Luc Chabert (Editor), *A History of Algorithms from the Pebble to the Microchip*, Springer, New York, 1999. As

1 * ' , #.@ "- % CO 6 B D C *# % #@ >
:5 5 " B D4

is noted in these sources, Newton's *De analysi per aequationes numero terminorum infinitas*, probably dating from mid-1669, is sometimes regarded as the historical source of the method, despite the fact that, surprisingly, there is no trace in this tract of the familiar recurrence relation $x = x - f(x)/f'(x)$ bearing Newton's name, nor is there a mention of the idea of derivative. Instead, the paper contains an example of a cubic polynomial whose roots are found by purely algebraic and rather complicated substitutions. In 1690, Joseph Raphson (1648–1715) in the Preface to his *Analysis aequationum universalis* describes his version of Newton's method as 'not only, I believe, not of the same origin, but also, certainly, not with the same development' as Newton's method. Further improvements to the method, and its form as we know it today, were given by Thomas Simpson in his *Essays in Mathematicks* (1740). Simpson presents it as 'a new method for the solution of equations' using the 'method of fluxions', *i.e.*, derivatives. It is argued in Ypma's article that Simpson's contributions to this subject have been underestimated, and 'it would seem that the Newton–Raphson–Simpson method is a designation more nearly representing facts of history of this method which lurks inside millions of modern computer programs and is printed with Newton's name attached in so many textbooks'.

The convergence analysis of Newton's method was initiated in the first half of the twentieth century by L.V. Kantorovich. More recently, Smale, Dedieu and Shub, and others have provided significant insight into the properties of Newton's method. A full discussion of the global behaviour of the logistic equation (1.33), and other examples, will be found in P.G. Drazin, *Nonlinear Systems*, Cambridge University Press, Cambridge, 1992, particularly Chapters 1 and 3.

The secant method is also due to Newton (cf. Section 3 of Ypma's paper cited above), and is found in a collection of unpublished notes termed 'Newton's Waste Book' written around 1665.

In this chapter, we have been concerned with the iterative solution of equations for a real-valued function of a single real variable. In Chapter 4, we shall discuss the iterative solution of nonlinear systems of equations

of the form $(\ ) = \mathbf{0}$ where $:$. There, corresponding to the case of $n = 2$, we shall say more about the solution of equations of the form $f(z) = 0$ where f is a complex-valued function of a single complex variable z.

This chapter has been confined to generally applicable iterative methods for the solution of a single nonlinear equation of the form $f(x) = 0$ for a real-valued function f of a single real variable. In particular, we have not discussed specialised methods for the solution of polynomial equations or the various techniques for locating the roots of polynomi-

Show that if the starting value is positive, the iteration converges to the positive solution, and if the starting value is negative it converges to the negative solution. Obtain approximate expressions for x_k if (i) $x_0 = 100$ and (ii) $x_0 = -100$, and describe the subsequent behaviour of the iteration. About how many iterations would be required to obtain the solution to six decimal digits in these two cases?

1.4 Consider the iteration

$$x_{k+1} = x_k - \frac{[f(x_k)]^2}{f(x_k + f(x_k)) - f(x_k)}, \qquad k = 0, 1, 2, \ldots,$$

for the solution of $f(x) = 0$. Explain the connection with Newton's method, and show that (x_k) converges quadratically if x_0 is sufficiently close to the solution. Apply this method to the same example as in Example 1.7, $f(x) = e^x - 2$, and verify quadratic convergence beginning from $x_0 = 1$. Experiment with calculations beginning from $x_0 = 10$ and from $x_0 = -10$, and account for their behaviour.

1.5 It is sometimes said that Newton's method converges quadratically, and therefore in the successive approximations to the solution the number of correct digits doubles each time. Explain why this is not generally correct. Suppose that $f''(x)$ is defined and continuous in a neighbourhood of ξ and that x_k agrees with the solution ξ to m decimal digits; give an estimate of the number of correct decimal digits in x_{k+1}.

Illustrate your estimate by using Newton's method to determine the positive zero of $f(x) = e^x - 1.000000005$, which is close to 0.0001; use $x_0 = 0.0005$.

1.6 Suppose that $f(\xi) = f'(\xi) = 0$, so that f has a double root at ξ, and that f'' is defined and continuous in a neighbourhood of ξ. If (x_k) is a sequence obtained by Newton's method, show that

$$\xi - x_{k+1} = -\frac{(\xi - x_k)^2 f''(\eta_k)}{f'(x_k)} = -(\xi - x_k)\frac{f''(\eta_k)}{f''(\chi_k)},$$

where η_k and χ_k both lie between ξ and x_k. Suppose, further, that $0 < m < |f''(x)| < M$ for all x in the interval $[\xi - \delta, \xi + \delta]$ for some $\delta > 0$, where $M < 2m$; show that if x_0 lies in this interval the iteration converges to ξ, and that convergence is

linear, with rate log 2. Verify this conclusion by finding the solution of e $= 1 + x$, beginning from $x = 1$.

1.7 Extend the result of the previous exercise to a case where f has a triple root at ξ, so that $f(\xi) = f'(\xi) = f''(\xi) = 0$.

1.8 Suppose that the function f has a continuous second derivative, that $f(\xi) = 0$, and that in the interval $[X, \xi]$, with $X < \xi$, $f'(x) > 0$ and $f''(x) < 0$. Show that the Newton iteration, starting from any x in $[X, \xi]$, converges to ξ.

1.9 The secant method is used to determine solutions of the equation $x \quad 1 = 0$. Starting from $x = 1 + \varepsilon$, $x = \quad 1 + \varepsilon$, show that $x = -\varepsilon + \quad (\varepsilon)$, and determine x , x and x , neglecting terms of order $\quad (\varepsilon)$. Explain why, at least for sufficiently small values of ε, the sequence (x) converges to the solution 1.

Repeat the calculation with x and x interchanged, so that $x = \quad 1 + \varepsilon$ and $x = 1 + \varepsilon$, and show that the sequence now converges to the solution 1.

1.10 Write the secant iteration in the form

$$x \quad = \frac{x \quad f(x \; - \;) \quad x \; - \; f(x \;)}{f(x \; - \;) \quad f(x \;)}, \qquad k = 1, 2, 3, \ldots .$$

Supposing that f has a continuous second derivative in a neighbourhood of the solution ξ of $f(x) = 0$, and that $f'(\xi) > 0$ and $f''(\xi) > 0$, define

$$\phi(x \; , x \; - \;) = \frac{x \quad \xi}{(x \quad \xi)(x \; - \quad \xi)},$$

where x has been expressed in terms of x and $x \; - \;$. Find an expression for

$$\psi(x \; - \;) = \lim_{\to} \phi(x \; , x \; - \;),$$

and then determine $\lim \quad_{\to} \psi(x \; - \;)$. Deduce that

$$\lim \quad_{\to} \phi(x \; , x \; - \;) = f''(\xi)/2f'(\xi).$$

Now assume that

$$\lim_{\to \infty} \frac{x \quad \xi}{x \quad \xi} = A.$$

Show that $q \quad 1 \quad 1/q = 0$, and hence that $q = -(1 + \quad 5)$.

Deduce finally that

$$\lim_{n \to \infty} \frac{x_{n+1} - \xi}{x_n - \xi} = \frac{f''(\xi)}{2f'(\xi)}$$

2

Solution of systems of linear equations

2.1 Introduction

In Chapter 1 we considered the solution of a single equation of the form $f(x) = 0$ where f is a real-valued function defined and continuous on a closed interval of the real line. The simplest example of this kind is the linear equation $ax = b$ where a and b are given real numbers, with $a = 0$, whose solution is

$$x = a^{-} b, \tag{2.1}$$

trivially. Of course, we could have expressed the solution as $x = b/a$ as in Chapter 1, but, as you will see in a moment, writing $x = a^{-} b$ is much more revealing in the present context. In this chapter we shall consider a different generalisation of this elementary problem:

Let A be an n n matrix with a as its entry in row i and column j and a given column vector of size n with jth entry b ; find a column vector of size n such that A = .

Denoting by x the ith entry of the vector , we can also write A = in the following expanded form:

$$\begin{array}{l} a\ x + a\ x + \quad + a\ x = b\ , \\ a\ x + a\ x + \quad + a\ x = b\ , \\ \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \\ a\ x + a\ x + \quad + a\ x = b\ . \end{array} \tag{2.2}$$

Recall that in order to ensure that for real numbers a and b the single linear equation $ax = b$ has a unique solution, we need to assume that $a = 0$. In the case of the simultaneous system (2.2) of n linear equations in n unknowns we shall have to make an analogous assumption on the matrix A.

To do so, we introduce the following definition.

Definition 2.1 *The set of all* $m \quad n$ *matrices with real entries is denoted by* $\quad^{\times}$ *. A matrix of size* $n \quad n$ *will be called a square matrix of order* n*, or simply a matrix of* **order** n*. The* **determinant** *of a square matrix* $A \quad^{\times}$ *is the real number* $\det(A)$ *defined as follows:*

$$\det(A) = \quad \operatorname{sign}(\nu \ ,\nu \ ,\ldots,\nu \ )a \quad a \quad \ldots a \quad .$$

The summation is over all $n!$ *permutations* $(\nu \ ,\nu \ ,\ldots,\nu \ )$ *of the integers* $1,2,\ldots,n$*, and* $\operatorname{sign}(\nu \ ,\nu \ ,\ldots,\nu \ ) = +1$ *or* $\quad 1$ *depending on whether the* n*-tuple* $(\nu \ ,\nu \ ,\ldots,\nu \ )$ *is an even or odd permutation of* $(1,2,\ldots,n)$*, respectively. An even (odd) permutation is obtained by an even (odd) number of exchanges of two adjacent elements in the array* $(1,2,\ldots,n)$*. A matrix* $A \quad^{\times}$ *is said to be* **nonsingular** *when its determinant* $\det(A)$ *is nonzero.*

The **inverse matrix** A^{-} of a nonsingular matrix $A \quad^{\times}$ is defined as the element of $\quad^{\times}$ such that $A^{-} \ A = AA^{-} \ = I$, where I is the $n \quad n$ identity matrix

$$I = \begin{matrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \ldots & \ldots & \ldots & \ldots \\ 0 & 0 & \ldots & 1 \end{matrix} \quad . \tag{2.3}$$

In order to find an explicit expression for A^{-} in terms of the elements of the matrix A, we recall from linear algebra that, for each $i = 1,2,\ldots,n$,

$$a \ A \quad + a \ A \quad + \quad + a \ A \quad = \begin{matrix} \det(A) & \text{if } i = k, \\ 0 & \text{if } i = k, \end{matrix} \tag{2.4}$$

where $A \quad = (\ 1) \quad \operatorname{Cof}(a \ )$ and $\operatorname{Cof}(a \ )$, called the **cofactor** of $a \ ,$ is the determinant of the $(n \quad 1) \quad (n \quad 1)$ matrix obtained by erasing from $A \quad^{\times}$ row i and column j. Then, it is a trivial matter to show using (2.4) that A^{-} has the form

$$A^{-} \ = \frac{1}{\det(A)} \begin{matrix} A & A & \ldots & A \\ A & A & \ldots & A \\ \ldots & \ldots & \ldots & \ldots \\ A & A & \ldots & A \end{matrix} \quad . \tag{2.5}$$

Having found an explicit formula for the matrix A^{-} , we now multiply both sides of the equation $A \quad =$ on the left by A^{-} to deduce that

$A^{-}(A\ ) = A^{-}\ $; finally, since $A^{-}(A\ ) = (A^{-}A)\ = I\ =\ $, it follows that

$$= A^{-}\ \ , \tag{2.6}$$

where the inverse A^{-} of the nonsingular matrix A is given in terms of the entries of A by (2.5).

An alternative approach to the solution of the linear system $A\ =\ $, called Cramer's rule, proceeds by expressing the ith entry of as

$$x\ = D\ /D\,, \qquad i = 1, 2, \ldots, n\,,$$

where $D = \det(A)$, and D is the $n\ \ n$ determinant obtained by replacing the ith column of D by the entries of . Evidently, we must require that A is nonsingular, *i.e.*, that $D = \det(A) = 0$. Thus, all we need to do to solve $A\ =\ $ is to evaluate the $n+1$ determinants $D, D\ , \ldots, D\ $, each of them $n\ \ n$, and check that $D = \det(A)$ is nonzero; the final calculation of the elements $x\ , i = 1, 2, \ldots, n$, is then trivial.

The purpose of our next example is to illustrate the application of Cramer's rule.

Example 2.1 *Suppose that we wish to solve the system of linear equations*

$$\begin{aligned} x\ +\ x\ +\ x\ &=\ 6\,, \\ 2x\ +4x\ +2x\ &=\ 16\,, \\ x\ +5x\ \ \ 4x\ &=\ 3\,. \end{aligned}$$

The solution of such a small system can easily be found in terms of determinants, by Cramer's rule. This gives

$$x\ = D\ /D\,,\quad x\ = D\ /D\,,\quad x\ = D\ /D\,,$$

1 6 6 , B 4?D B 4#D 6 , B 4#D ,
, 1 B 4?D A #4
2 F B*# K 6 #& F 5 + C K 6 #&> $
$ O+ 9 , D4) #&* ' , B9 6 #?@.
% 6 , C # K #& ? , D 1
, #& . 6 1 4)
, E 5 N 1
* * 6 , , "4
5 1 6 1 % 8
P) , 6 1 , , 5 N B#&> D 5 6
E G 1 , 5 1 5 4

where

$$D = \begin{matrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ 1 & 5 & 4 \end{matrix}, \quad D = \begin{matrix} 6 & 1 & 1 \\ 16 & 4 & 2 \\ 3 & 5 & 4 \end{matrix},$$

with similar expressions for D and D. To obtain the solution we therefore need to evaluate four determinants.

Now you may think that since, for A $\times$ nonsingular, we have expressed the solution to $A = $ in the 'closed form'

$$= A^{-}$$

and have even found a formula for A^{-} in terms of the coefficients of A, or may simply compute the entries of directly using Cramer's rule, the story about the simultaneous set of linear equations (2.2) has reached its happy ending. We are sorry to disappoint you: a disturbing tale is about to unfold.

Imagine the following example: let $n = 100$, say, and suppose that you have been given all 10000 entries of a 100 100 matrix A, together with the entries of a 100-component column vector . To avoid trivialities, let us suppose that none of the entries of A or is equal to 0. Question: *Does the linear system* $A = $ *have a solution? If it does, how would you find, say, the 53rd entry of the solution vector ?* Of course, you could calculate the determinant of A and check whether it is equal to zero; if not, you could then calculate the determinant D obtained by replacing the 53rd column of A by the vector , and the required result, by Cramer's rule, is then the ratio of these two determinants. How much time do you think you would need to accomplish this task? An hour? A day? A month?

I imagine that you do not have a large enough sheet of paper in front of you to write down this 100 100 matrix. Let us therefore start with a somewhat simpler setting. Assume that n is any integer, n 2, and denote by d the number of arithmetic operations that are required to calculate $\det(A)$ for A $\times$. For example, for a 2 2 matrix,

$$\det(A) = a \quad a \qquad a \quad a \quad ;$$

this evaluation requires 3 arithmetic operations – 2 multiplications and 1 subtraction – giving $d = 3$. In general, we can calculate $\det(A)$ by expanding it in the elements of its first row. This requires multiplying each of the n elements in the first row of A by a subdeterminant of size

that we cannot by this means reduce the total by more than a factor of about n, which hardly affects our conclusion.

Our other approach to solving $A = $, based on computing A^- from (2.5) and writing $= A^-$, is equally inefficient: in order to compute the inverse of an n n matrix A using determinants, one has to calculate the determinant of A as well as n determinants of size n 1 each of which then has to be divided by $\det(A)$, requiring a total of approximately

$$e\,n! + n\ e\,(n \quad 1)! + n \qquad e\,(n+1)!$$

arithmetic operations, just the same as before.

The aim of this chapter is to develop alternative methods for the solution of the system of linear equations $A = $. We begin by considering a classical technique, Gaussian elimination. We shall then explore its relationship to the factorisation $A = LU$ of the matrix A where L is lower triangular and U is upper triangular. It will be seen that by using the Gaussian elimination the number of arithmetic operations required to solve the linear system $A = $ with an n n matrix A is approximately $-n$ – a dramatic reduction from the $(e\,(n+1)!)$ operation count associated with matrix inversion using determinants.

We conclude the chapter with a discussion of another classical idea attributed to Gauss: the least squares method for the solution of the system of linear equations $A = $ where A $^{\times}$, is the column vector of unknowns of size n and a given column vector of size m.

2.2 Gaussian elimination

The technique for solving systems of linear algebraic equations that we shall describe in this section was developed by Carl Friedrich Gauss and was first published in his *Theoria motus corporum coelestium in sectionibus conicis solem ambientium* (1809), a major two-volume treatise on the motion of celestial bodies. Gauss was concerned with the study of

the asteroid Pallas, and derived a set of six linear equations with six unknowns, also giving a systematic method for its solution.

The method proceeds by successively eliminating the elements below the diagonal of the matrix of the linear system until the matrix becomes triangular, when the solution of the system is very easy. This technique is now known under the name **Gaussian elimination**.

Before we embark on the general description of Gaussian elimination, let us illustrate its basic steps through a simple example; this is the same as Example 2.1 above, written out again for convenience.

Example 2.2 *Consider the system of linear equations*

$$\begin{aligned} x + x + x &= 6, \\ 2x + 4x + 2x &= 16, \\ x + 5x \quad 4x &= 3. \end{aligned}$$

It is convenient to rewrite this in the form $A = $ where A $^{\times}$ and and are column vectors of size 3; thus,

$$\begin{matrix} 1 & 1 & 1 \\ 2 & 4 & 2 \\ 1 & 5 & 4 \end{matrix} \quad \begin{matrix} x \\ x \\ x \end{matrix} \quad = \quad \begin{matrix} 6 \\ 16 \\ 3 \end{matrix} \quad . \tag{2.9}$$

We begin by adding the first row, multiplied by 2, to the second row, and adding the first row to the third row, giving the new system

$$\begin{matrix} 1 & 1 & 1 \\ \textit{0} & 2 & 0 \\ \textit{0} & 6 & 3 \end{matrix} \quad \begin{matrix} x \\ x \\ x \end{matrix} \quad = \quad \begin{matrix} 6 \\ 4 \\ 3 \end{matrix} \quad . \tag{2.10}$$

The newly created 0 entries in the first column have been typeset in italics. Now adding the new second row, multiplied by 3, to the third row, we find

$$\begin{matrix} 1 & 1 & 1 \\ \textit{0} & 2 & 0 \\ \textit{0} & \textit{0} & 3 \end{matrix} \quad \begin{matrix} x \\ x \\ x \end{matrix} \quad = \quad \begin{matrix} 6 \\ 4 \\ 9 \end{matrix} \quad , \tag{2.11}$$

1 1 , 6 " $
6 4 " - B 6 4
2 (' H81 5 6
#@@@D , 8 1 1 6 1E5 G
E5 " 4 " 5 6 R 6 1 $
, , , " 5 5
64 % ' , ,
' " 0 6 6
E , 6 % / 4

which can easily be solved for the unknowns in the reverse order, beginning with $x_3 = 3$.

Each of these successive row operations can be expressed as a multiplication on the left of the matrix $A^{(n,2)}$ (in our example $n = 3$), of the system of linear equations by a transformation matrix. Writing $E^{(n,n)}$ for the $n \times n$ matrix whose only nonzero element is $e_{rs} = 1$, we see that the product

$$(I + \mu E^{(n,n)})A \tag{2.12}$$

is the same as the original matrix A, except that the elements of row s, multiplied by a real number μ, have been added to the corresponding elements of row r. Here I denotes the $n \times n$ identity matrix defined by (2.3). In the elimination process we always add a multiple of an earlier row to a later row in the matrix, so that $1 \leq s < r \leq n$ in (2.12); the transformation matrix $I + \mu E^{(n,n)}$ is therefore lower triangular in the following sense.

Definition 2.2 *Let n be an integer, $n \geq 2$. The matrix $L^{(n,n)}$ is said to be* **lower triangular** *if $l_{ij} = 0$ for every i and j with $1 \leq i < j \leq n$. The matrix $L^{(n,n)}$ is called* **unit lower triangular** *if it is lower triangular, and also the diagonal elements are all equal to unity, that is $l_{ii} = 1$ for $i = 1, 2, \ldots, n$.*

Thus the matrix $I + \mu E^{(n,n)}$ appearing in (2.12) is unit lower triangular if $1 \leq s < r \leq n$, and the above elimination process can be expressed by multiplying A on the left successively by the unit lower triangular matrices $I + \mu E^{(n,n)}$ for $r = s+1, \ldots, n$ and $s = 1, \ldots, n-1$, with $\mu = -a_{rs}/a_{ss}$; there are $\frac{1}{2}n(n-1)$ of these matrices, one for each element of A below the diagonal (since there are n elements on the diagonal and, therefore, $1 + 2 + \cdots + (n-1) = \frac{1}{2}(n-1)n$ elements below the diagonal). The next theorem lists the technical tools which are required for proving that the resulting product is a lower triangular matrix.

Theorem 2.1 *The following statements hold for any integer $n \geq 2$:*

(i) *the product of two lower triangular matrices of order n is lower triangular of order n;*

(ii) *the product of two unit lower triangular matrices of order n is unit lower triangular of order n;*

(iii) *a lower triangular matrix is nonsingular if, and only if, all the*

diagonal elements are nonzero; in particular, a unit lower triangular matrix is nonsingular;

(iv) *the inverse of a nonsingular lower triangular matrix of order* n *is lower triangular of order* n*;*

(v) *the inverse of a unit lower triangular matrix of order* n *is unit lower triangular of order* n*.*

Proof The proofs of parts (i), (ii), (iii) and (v) are very straightforward, and are left as an exercise.

Part (iv) is proved by induction; it is easily verified for a nonsingular lower triangular matrix of order 2, using (2.5). Let $n > 2$, suppose that (iv) is true for all nonsingular lower triangular matrices of order k, with $2 \leq k < n$, and let L be a nonsingular lower triangular matrix of order $k+1$. Both L and its inverse L^{-1} can be partitioned by their last row and column:

$$L = \begin{bmatrix} L_k & \mathbf{0} \\ \gamma & \alpha \end{bmatrix}, \quad L^{-1} = \begin{bmatrix} X & \\ \delta & \beta \end{bmatrix},$$

where L_k is a nonsingular lower triangular matrix of order k and $X \in \mathbb{R}^{k \times k}$; α and β are real numbers and γ, δ and $\mathbf{0}$ are column vectors of size k. Since the product LL^{-1} is the identity matrix of order $k+1$, we have

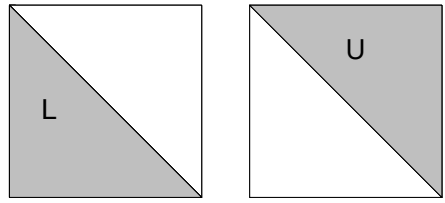
$$L_k X = I_k, \quad L_k \delta = \mathbf{0}, \quad \gamma^T X + \alpha \delta^T = \mathbf{0}^T, \quad \gamma^T \delta + \alpha\beta = 1;$$

here I_k signifies the identity matrix of order k. Thus $X = L_k^{-1}$, which is lower triangular of order k by the inductive hypothesis, and $\delta = \mathbf{0}$ given that L_k is nonsingular; the remaining two equations determine γ and β on noting that $\alpha = 0$ (given that L is nonsingular). This shows that L^{-1} is lower triangular of order $k+1$, and the inductive step is complete; consequently, (iv) is true for any $n \geq 2$. □

We shall also require the concept of upper triangular matrix.

Definition 2.3 *Let* n *be an integer,* $n \geq 2$*. The matrix* $U \in \mathbb{R}^{n \times n}$ *is said to be* **upper triangular** *if* $u_{ij} = 0$ *for every* i *and* j *with* $1 \leq j < i \leq n$.

We note that results analogous to those in the preceding theorem concerning lower triangular matrices are also valid for upper triangular matrices (replacing the words 'lower triangular' by 'upper triangular' throughout).



34 # 5 !

The elimination process for A × may now be written as follows:

$$L \quad L \quad \_ \quad \ldots L \quad A = U, \qquad N = -n(n \quad 1), \tag{2.13}$$

where U × is an upper triangular matrix and each of the matrices L × , j = 1, . . . , N, is unit lower triangular of order n and has the form I + μ E with 1 s < r n, where I is the identity matrix of order n. That is,

$$L \quad = I + \mu \quad E \quad , L \quad = I + \mu \quad E \quad , \ldots , L \quad = I + \mu \quad \_ \quad E \quad ^{-} .$$

It is easy to see that E E = δ E , where

$$\delta \quad = \quad \begin{array}{ll} 1 & \text{for } r = s, \\ 0 & \text{for } r = s \end{array}$$

is known as the **Kronecker delta**. Thus, for 1 s < r n, the inverse of the matrix I + μ E is the lower triangular matrix I μ E , which corresponds to the subtraction of row s, multiplied by μ , from row r. Hence

$$A = L^{-} \quad \ldots L^{-} \quad U = LU, \tag{2.14}$$

where L, as the product of a finite number of unit lower triangular matrices of order n, is itself unit lower triangular of order n by Theorem 2.1(ii); see Figure 2.1.

2.3 LU factorisation

Having seen that the Gaussian elimination process gives rise to the factorisation A = LU of the matrix A × , n 2, where L is unit

1 : ," B& / , #. * : + F 6 B : ,
D C @ / , #.@# F 6D E, ,
6 1 , 1 , 6 1 1 G ,
1 4

lower triangular and U is upper triangular, we shall now show how to calculate the elements of L and U directly. Equating the elements of A and LU

in the formula (2.18) is zero. To investigate this possibility we use the properties of certain submatrices of A.

Definition 2.4 *Suppose that* $A \in \mathbb{R}^{n \times n}$ *with* $n \geq 2$, *and let* $1 \leq k \leq n$. *The* **leading principal submatrix** *of order* k *of* A *is defined as the matrix* $A_k \in \mathbb{R}^{k \times k}$ *whose element in row* i *and column* j *is equal to the element of the matrix* A *in row* i *and column* j *for* $1 \leq i, j \leq k$.

Armed with this definition, we can now formulate the main result of this section. It provides a sufficient condition for ensuring that the algorithm (2.18), (2.19) for calculating the entries of the matrices L and U in the LU factorisation $A = LU$ of a matrix $A \in \mathbb{R}^{n \times n}$ does not break down due to division by zero in (2.18).

Theorem 2.2 *Let* $n \geq 2$, *and suppose that* $A \in \mathbb{R}^{n \times n}$ *is such that every leading principal submatrix* $A_k \in \mathbb{R}^{k \times k}$ *of* A *of order* k, *with* $1 \leq k < n$, *is nonsingular. (Note that* A *itself is not required to be nonsingular.) Then,* A *can be factorised in the form* $A = LU$, *where* $L \in \mathbb{R}^{n \times n}$ *is unit lower triangular and* $U \in \mathbb{R}^{n \times n}$ *is upper triangular.*

Proof The proof is by induction on the order n. Let us begin by verifying the statement of the theorem for $n = 2$. We intend to show that any 2×2 matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix},$$

with $a \neq 0$, is equal to the product of a unit lower triangular matrix L of order 2 and an upper triangular matrix U of order 2; that is, we wish to establish the existence of

$$L = \begin{bmatrix} 1 & 0 \\ m & 1 \end{bmatrix}, \qquad U = \begin{bmatrix} u & v \\ 0 & \eta \end{bmatrix},$$

such that $LU = A$, where m, u, v and η are four real numbers, to be determined. Equating the product LU with A, we deduce that

$$u = a, \quad v = b, \quad mu = c, \quad mv + \eta = d.$$

Since $a \neq 0$ by hypothesis, the first of these equalities implies that $u \neq 0$ also; hence $m = c/u$, $v = b$, and $\eta = d - mv$. Thus we have shown the existence of the required matrices L and U in $\mathbb{R}^{2 \times 2}$ and completed the proof for $n = 2$.

Now, suppose that the statement of the theorem has already been verified for matrices of order k, $2 \leq k < n$; suppose that A $\times$ and all leading principal submatrices of A of order k and less are nonsingular. We mimic the proof in the case of $n = 2$ by partitioning A into blocks by the last row and column:

$$A = \begin{matrix} A & \\ & d \end{matrix},$$

where A $\times$ is a nonsingular matrix (all of whose leading principal submatrices are themselves nonsingular), , are column vectors of size k, and d is a real number. According to our inductive hypothesis, there exist a unit lower triangular matrix L of order k and an upper triangular matrix U of order k such that $A = L U$. Thus we shall seek the desired unit lower triangular matrix L of order $k+1$ and the upper triangular matrix U of order $k+1$ in the form

$$L = \begin{matrix} L & \mathbf{0} \\ & 1 \end{matrix} \quad \text{and} \quad U = \begin{matrix} U & \\ \mathbf{0} & \eta \end{matrix}$$

where and are column vectors of size k and η is a real number, to be determined from the requirement that the product LU be equal to the matrix A. On equating LU with A, we obtain

$$L U = A, \quad L = , \quad U = , \quad + \eta = d.$$

The first of these four equalities provides no new information. However, we can use the remaining three to determine the column vectors and and the real number η. Since L is unit lower triangular, its determinant is equal to 1; therefore L is nonsingular. This means that the second equation uniquely determines the unknown column vector . Further, since by hypothesis A is nonsingular and $A = L U$, we conclude that

$$\det(A) = \det(L U) = \det(L)\det(U) = \det(U);$$

given that $\det(A) = 0$ by the inductive hypothesis, this implies that $\det(U) = 0$ also, and therefore the third equation uniquely determines . Having found and , the fourth equation yields $\eta = d$. Thus we have shown the existence of the desired matrices L and U of order $k+1$, and the inductive step is complete. □

[1]) 1 ! " # , 1 , A 5 B D A B D B D4 5 #.# 6 6 % $: , 6 B#&.@C#.>&D K ,G ' B#&.?C#.>?D4

2.4 Pivoting

The aim of this section is to show that even if the matrix A does not satisfy the conditions of Theorem 2.2, by permuting rows and columns it can be transformed into a new matrix $\tilde{A}$ of the same size so that $\tilde{A}$ admits an LU factorisation.

Example 2.3 *Consider, for example, the system obtained from (2.9) by replacing the coefficient of* x *in the first equation by zero. Then, the leading element in the matrix* A *is zero, the computation fails at the first step, and the LU factorisation of* A *does not exist. However if we interchange the first two equations we obtain a new matrix* $\tilde{A}$ *which is the same as* A *but with the first two rows interchanged,*

$$\tilde{A} = \begin{matrix} 2 & 4 & 2 \\ 0 & 1 & 1 \\ 1 & 5 & 4 \end{matrix} . \tag{2.20}$$

Since the leading principal submatrices of order 1 *and* 2 *of* $\tilde{A}$ *are non-singular, by Theorem 2.2 the matrix* $\tilde{A}$ *now has the required LU factorisation, which is easily computed.*

A computation which fails when an element is exactly zero is also likely to run into difficulties when that element is nonzero but of very small absolute value; the problem stems from the presence of rounding errors. The basic operation in the elimination process consists of multiplying the elements of one row of the matrix by a scalar μ, and adding to the elements of another row. The multiplication operation will always introduce a rounding error, so the elements which are multiplied by μ will already contain a rounding error from operations with earlier rows of the matrix; these errors will therefore themselves be multiplied by μ before adding to the new row. The errors will be magnified if $\mu > 1$, and will be greatly magnified if $\mu \quad 1$.

The accumulation of rounding errors alluded to in the previous paragraph can be alleviated by permuting the rows of the matrix. Thus, at each stage of the elimination process we interchange two rows, if necessary, so that the largest element in the current column lies on the diagonal. This process is known as **pivoting**. Clearly, when pivoting is performed none of the multipliers μ have absolute value greater than unity. The process is easily formalised by introducing permutation matrices. This leads us to our next definition.

Definition 2.5 *Suppose that* $n \geq 2$. *A matrix* $P \in \mathbb{R}^{n \times n}$ *in which every element is either* 0 *or* 1, *and whose every row and every column contain exactly one nonzero element, is called a* **permutation matrix**.

Example 2.4 *Here are three of the possible* 3! *permutation matrices in* $\mathbb{R}^{3 \times 3}$:

$$\begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 1 & 0 \\ 1 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}, \quad \begin{bmatrix} 0 & 0 & 1 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{bmatrix}.$$

The proof of our next result is elementary and is left to the reader.

Lemma 2.1 *Let* $n \geq 2$ *and suppose that* $P \in \mathbb{R}^{n \times n}$ *is a permutation matrix. Then, the following statements hold:*

(i) *given that* I *is the identity matrix of order* n, *the matrix* P *can be obtained from* I *by permuting rows;*

(ii) *if* $Q \in \mathbb{R}^{n \times n}$ *is another permutation matrix, then the products* PQ *and* QP *are also permutation matrices;*

(iii) *let* $P \in \mathbb{R}^{n \times n}$ *denote the* **interchange matrix**, *obtained from the identity matrix* $I \in \mathbb{R}^{n \times n}$ *by interchanging rows* r *and* s; *any interchange matrix is a permutation matrix; moreover, any permutation matrix of order* n *can be written as a product of interchange matrices of order* n;

(iv) *the determinant of a permutation matrix* $P \in \mathbb{R}^{n \times n}$ *is equal to* 1 *or* -1, *depending on whether* P *is obtained from the identity matrix of order* n *by an even or odd number of permutations of rows, respectively; in particular, a permutation matrix is nonsingular.*

Now we are ready to prove the next theorem.

Theorem 2.3 *Let* $n \geq 2$ *and* $A \in \mathbb{R}^{n \times n}$. *There exist a permutation matrix* P, *a unit lower triangular matrix* L, *and an upper triangular matrix* U, *all three in* $\mathbb{R}^{n \times n}$, *such that*

$$PA = LU. \tag{2.21}$$

Proof The proof is by induction on the order n. Let $n = 2$ and consider the matrix

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix}.$$

If $a = 0$, the proof follows from Theorem 2.2 with P taken as the 2×2 identity matrix. If $a = 0$ but $c = 0$, we take

$$P = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix}$$

and write

$$PA = \begin{matrix} c & d \\ 0 & b \end{matrix} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \begin{matrix} c & d \\ 0 & b \end{matrix} \quad LU .$$

If $a = 0$ and $c = 0$, the result trivially follows by writing

$$\begin{matrix} 0 & b \\ 0 & d \end{matrix} = \begin{matrix} 1 & 0 \\ 0 & 1 \end{matrix} \begin{matrix} 0 & b \\ 0 & d \end{matrix} \quad LU$$

and taking P as the 2×2 identity matrix. That completes the proof for $n = 2$.

Now, suppose that $A \quad \times$ and assume that the theorem holds for every matrix of order k with $2 \quad k < n$. We begin by locating the element in the first column of A which has the largest absolute value, or any one of them if there is more than one such element, and interchange rows if required; if the largest element is in row r we interchange rows 1 and r. We then partition the new matrix according to the first row and column, writing

$$P \quad A = \begin{matrix} \alpha & \\ B & \end{matrix} = \begin{matrix} 1 & \mathbf{0} \\ & I \end{matrix} \begin{matrix} \alpha & \\ \mathbf{0} & C \end{matrix} \tag{2.22}$$

where α is the element of largest absolute value in the first column, $B, C \quad \times$, and , , and are column vectors of size k, with , and C to be determined. Writing out the product we find that

$$\begin{aligned} &= , \\ \alpha &= , \\ C &= B \quad . \end{aligned} \tag{2.23}$$

If $\alpha = 0$, then the first column of A consists entirely of zeros ($= \mathbf{0}$); in this case we can evidently choose $= \mathbf{0}$, $=$ and $C = B$. Suppose now that $\alpha = 0$; then $= (1/\alpha)$, so that all the elements of have absolute value less than or equal to unity, since α is the largest in absolute value element in the first column. By the inductive hypothesis we can now write

$$P' C = L' U' , \tag{2.24}$$

where P', L', U' $\times$, P' is a permutation matrix, L' is unit lower triangular, and U' is upper triangular. Hence, by (2.23),

$$P \quad A = \begin{matrix} 1 & \mathbf{0} \\ \mathbf{0} & P' \end{matrix} \qquad \begin{matrix} 1 & \mathbf{0} \\ \mathbf{0} & P' \end{matrix} \qquad \begin{matrix} \alpha & \\ \mathbf{0} & U' \end{matrix} \tag{2.25}$$

since $P'P' = I$. Now, defining the permutation matrix P by

$$P = \begin{matrix} 1 & \mathbf{0} \\ \mathbf{0} & P' \end{matrix} \quad P \quad , \tag{2.26}$$

we obtain

$$PA = \begin{matrix} 1 & \mathbf{0} \\ P' & L' \end{matrix} \qquad \begin{matrix} \alpha & \\ \mathbf{0} & U' \end{matrix} \quad , \tag{2.27}$$

which is the required factorisation of A $\times$. This completes the inductive step. The theorem therefore holds for every matrix of order n 2. □

The proof of this theorem also contains an algorithm for constructing the permutation matrix P, and the matrices L and U. The permutation matrix is conveniently described by specifying the sequence of interchanges: given the n 1 integers p , p , . . . , p – , the matrix P is the product of the permutation matrices which interchange rows 1 and p , 2 and p , and so on.

2.5 Solution of systems of equations

Consider the linear system A = where A $\times$ and and are column vectors of size n. According to Theorem 2.3 there exist a permutation matrix P $\times$, a unit lower triangular matrix L $\times$ and an upper triangular matrix U $\times$ such that $PA = LU$. Having obtained the LU factorisation of the matrix PA, the solution of the system of linear equations A = is straightforward: multiplying both sides of A = on the left by the permutation matrix P, we obtain that

$$PA \quad = P \quad ; \tag{2.28}$$

equivalently, LU $= P$. On defining $= U$ we can rewrite (2.28) as the following coupled set of linear equations:

$$L \quad = P \quad , \qquad U \quad = \quad . \tag{2.29}$$

Assuming that the matrix P and the LU factorisation of PA are already known, there are three stages to the calculation of :

(First we apply the sequence of permutations to the vector , to produce P ;

( )* + We then solve the lower triangular system L = P , calculating the elements in the order y , y , . . . , y ;

(,)- + Finally the required solution is obtained from the upper triangular system U = , calculating the elements of in the reverse order, x , x – , . . . , x .

(, will break down if any of the diagonal elements of U are zero, but if this happens the matrix A is singular.

The next section is devoted to assessing the amount of computational work for this algorithm.

2.6 Computational work

In this section we shall show that the work involved in factorising an n n matrix in the form A = LU is proportional to n . An estimate of the amount of computational work of this kind is important in deciding in advance how long a calculation would take for a very large matrix, and is also useful in comparing different methods for the solution of a given problem. For example, in the next chapter we shall derive a method for solving a system of equations with a symmetric positive definite matrix; that method requires only half the amount of work involved in the standard LU factorisation algorithm which takes no account of symmetry.

Accurate estimates of the time taken by a computation are very complicated and require some detailed knowledge of the computer being used. The estimates which we shall give are simple but crude; they are normally good enough for the types of comparisons we have just mentioned.

We see from (2.18) that the calculation of l requires j 1 multiplications, j 2 additions, 1 subtraction and 1 division, a total of 2j 1 operations. In the same way, (2.19) shows that the calculation of u requires 2i 2 operations. Recalling that, for any integer k 2,

$$1 + \quad + k = -k(k+1) \quad \text{and} \quad 1 + \quad + k = -k(k+1)(2k+1)\,,$$

we then deduce that the total number of operations involved in the LU

1 , , 1 P N4

factorisation is

$$\sum_{j=1}^{i-1}(2j-1)+\sum_{i=1}^{n}2(i-1)=\tfrac{1}{3}n(n-1)(4n+1)\,.$$

It is enough to say that the number of multiplications required is about $\frac{1}{3}n^3-\frac{1}{2}n^2$, for moderately large values of n.

Having constructed the factorisation we can now count the number of operations required to compute the vectors y and x in (2.29). Given the vector P, the elements of y are obtained from

$$y_1=(P_1)\,,\qquad y_i=(P_i)-\sum_{j=1}^{i-1}l_{ij}y_j\,,\quad i=2,3,\ldots,n\,,\tag{2.30}$$

which requires $2i-2$ operations. Summing over i this gives a total of $n(n-1)$. The calculation of the elements of x is similar:

$$x_n=\frac{1}{u_{nn}}y_n\,,\qquad x_i=\frac{1}{u_{ii}}\Big(y_i-\sum_{j=i+1}^{n}u_{ij}x_j\Big)\,,\quad i=n-1,n-2,\ldots,1\,.\tag{2.31}$$

This requires $2(n-i)+1$ operations, giving a total of n^2.

The total number of operations involved in the solution of the system of equations is therefore approximately $\frac{1}{3}n^3-\frac{1}{2}n^2$ for the factorisation, followed by $n(n-1)+n^2=2n^2$ for the solution of the two triangular systems, that is, approximately $\frac{1}{3}n^3+\frac{3}{2}n^2$, ignoring terms of size $O(n)$.

We often need to solve a number of systems of this kind, all with different right-hand sides, but with the same matrix. We then need only factorise the matrix once, and the total number of multiplications required for k right-hand sides becomes approximately $\frac{1}{3}n^3+2kn^2-kn$. When k is fairly large it might appear that it would be more efficient to form the inverse matrix A^{-1}, and then multiply each right-hand side by the inverse; but we shall show that it is not so.

To form the inverse matrix we first factorise the matrix A, and then solve n systems, with the right-hand sides being the vectors which constitute the columns of the identity matrix. Because these right-hand sides have a special form, there is the possibility of saving some work; some careful counting shows that the total can be reduced from $\frac{1}{3}n^3+2n^2=\frac{7}{3}n^3$ to an approximate total of $2n^3$ operations. It is easy to see that the operation of multiplying a vector by the inverse matrix requires $n(2n-1)$ operations; hence the whole computation of first constructing the inverse matrix, and then multiplying each right-hand side by the inverse, requires a total of $2n^3+2kn^2$ multiplications (ignoring terms of size

(n)). This is always greater than the previous value $-n + 2k - n$, whether k is small or large. The most efficient way of solving this problem is to construct and save the L and U factors of A, rather than to form the inverse of A.

2.7 Norms and condition numbers

The analysis of the effects of rounding error on solutions of systems of linear equations requires an appropriate measure. This is provided by the concept of **norm** defined below. In order to motivate the axioms of norm stated in Definition 2.6, we note that the set of real numbers is a linear space, and that the **absolute value** function

$$v \qquad v = \begin{cases} v & \text{if } v \quad 0, \\ v & \text{if } v < 0 \end{cases}$$

has the following properties:

v 0 for any v, and $v = 0$ if, and only if, $v = 0$;
$\lambda v = \lambda v$ for all λ and all v;
$u + v$ $u + v$ for all u and v in .

The absolute value v of a real number v measures the distance between v and 0 (the zero element of the linear space). Our next definition aims to generalise this idea to an arbitrary linear space over the field of real numbers: even though the discussion in the present chapter is confined to finite-dimensional linear spaces of vectors (=) and square matrices (= $^{\times}$), norms over other linear spaces, including infinite-dimensional function spaces, will appear elsewhere in the text (see Chapters 8, 9, 11 and 14).

Definition 2.6 *Suppose that is a linear space over the field of real numbers. The nonnegative real-valued function is said to be a* **norm** *on the space provided that it satisfies the following axioms:*

$v = 0$ *if, and only if,* $v = 0$ *in ;*
$\lambda v = \lambda v$ *for all* λ *and all* v *in ;*
$u + v$ $u + v$ *for all* u *and* v *in (the triangle inequality).*

A linear space , equipped with a norm, is called a **normed linear space**.

Remark 2.1 *If is a linear space over the field of complex numbers, then in the second axiom of Definition 2.6 should be replaced by , with* λ *signifying the modulus of* λ .

Any norm on the linear space = will be called a **vector norm**. Three vector norms are in common use in numerical linear algebra: the 1-norm , the 2-norm (or Euclidean norm) , and the -norm ∞; these are defined below.

Definition 2.7 *The* 1**-norm** *of the vector* = (v , . . . , v) *is defined by*

$$= \quad \mathsf{v} \ . \tag{2.32}$$

Definition 2.8 *The* 2**-norm** *of the vector* = (v , . . . , v) *is defined by* = . *In other words,*

$$= \quad \mathsf{v} \quad . \tag{2.33}$$

Definition 2.9 *The* **-norm** *of the vector* = (v , . . . , v) *is defined by*

$$_\infty = \max \mathsf{v} \ . \tag{2.34}$$

When n = 1, each of these norms collapses to the absolute value, , the simplest example of a norm on = .

It is easy to show that and ∞ obey all axioms of a norm. For the 2-norm the first two axioms are still trivial to verify; to show that the triangle inequality is satisfied by the 2-norm requires use of the Cauchy –Schwarz inequality.

Lemma 2.2 (**Cauchy–Schwarz inequality**)

$$\mathsf{u}\,\mathsf{v} \qquad , \qquad . \tag{2.35}$$

1 % $: , 6 B # % #&.@ 9 , C * ' 6 #.>& , 8
B D 9 , D 5 6 E, ,
64 0 1 1 1 , 6 6
1 , 8 1 , , 1 , 6 64
2 0 % , + B > K 6 #. * 0 1 F 6
B D C * 5 #@ # F 6D ,,
1 1 ' , #.@ 4 H ,
, , 1 , (6 9 $
, 6 6 , 1 4

Theorem 2.4 (**Young's inequality**) *Let* $p, q > 1$, $(1/p) + (1/q) = 1$. *Then, for any two nonnegative real numbers* a *and* b,

$$ab \quad \frac{a}{p} + \frac{b}{q}.$$

Proof If either $a = 0$ or $b = 0$ the inequality holds trivially. Let us therefore suppose that $a > 0$ and $b > 0$. We recall that a function $x \quad f(x)$ is said to be **convex** if

$$f(\theta x + (1 \quad \theta)y) \quad \theta f(x) + (1 \quad \theta) f(y)$$

for all $\theta \quad [0, 1]$, and all x and y in ; *i.e.*, for any x and y in the graph of the function f between the points $(x, f(x))$ and $(y, f(y))$ lies below the chord that connects these two points. Note that the function $x \quad e$ is convex. Therefore, with $\theta = 1/p$ and $1 \quad \theta = 1/q$, we get that

$$ab = e \qquad = e \qquad \frac{1}{p} e \qquad + \frac{1}{q} e \qquad = \frac{a}{p} + \frac{b}{q},$$

and the proof is complete. (When $p = q = 2$ the proof is trivial: as $(a \quad b) \quad 0$ also $2ab \quad a + b$, and hence the required result.) □

The next step is to establish Hölder's inequality; it is a generalisation of the Cauchy–Schwarz inequality.

Theorem 2.5 (**Hölder's inequality**) *Let* $p, q > 1$, $(1/p) + (1/q) = 1$. *Then, for any and , we have*

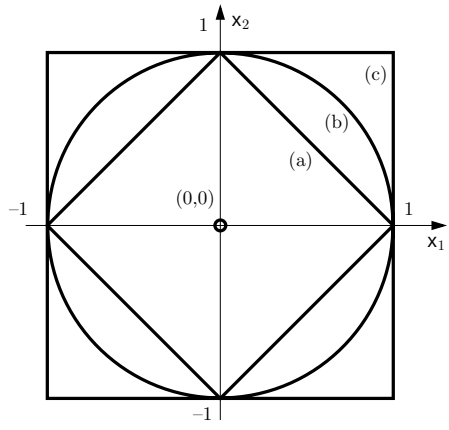
$$u\, v \qquad .$$

Proof If either $= \mathbf{0}$ or $= \mathbf{0}$ the inequality holds trivially. Let us therefore suppose that $= \mathbf{0}$ and $= \mathbf{0}$, and consider the vectors ˜ and ˜ in with components $\tilde{u} = u /$ and $\tilde{v} = v /$, respectively, $i = 1, 2, \ldots, n$. By Young's inequality,

$$\tilde{u}\, \tilde{v} \qquad \tilde{u}\, \tilde{v} \quad \frac{1}{p} \quad \tilde{u} \quad + \frac{1}{q} \quad \tilde{v} \quad = \frac{1}{p} + \frac{1}{q} = 1.$$

Inserting the defining expressions for $\tilde{u}$ and $\tilde{v}$ into the left-most expression in this chain, the result follows. □

[illegible]



64 0 # 2
* + " *,+ " * + "

the unit sphere in a normed linear space , with norm , is defined as the set : = 1 . It can be seen from Figure 2.2 that

: 1 : 1 : ∞ 1 .

We leave it to the reader as an exercise to show that analogous inclusions hold in for any n 1. (See Exercise 8.)

The unit sphere in a normed linear space with norm is the boundary of the closed unit ball B̄ (**0**) centred at **0** defined by

B̄ (**0**) = : 1 .

Analogously, the open unit ball centred at **0** is defined by

B (**0**) = : < 1 .

More generally, for ε > 0 and ,

B̄ ( ) = : ε

is the **closed ball** of radius ε centred at ; analogously,

B ( ) = : < ε

is the **open ball** of radius ε centred at .

Any norm on the linear space × of n n matrices with real entries will be referred to as a **matrix norm**. In particular, we shall now

consider matrix norms which are induced by vector norms in a sense that will be made precise in the next definition.

Definition 2.10 *Given any norm on the space of* n*-dimensional vectors with real entries, the* **subordinate matrix norm** *on the space* $\times$ *of* n n *matrices with real entries is defined by*

$$A = \max_{\square} \frac{A}{\quad}. \tag{2.38}$$

In (2.38) we used $\square$ to denote $\mathbf{0}$, where, for sets A and B, $A \quad B = x \quad A: x / B$.

Remark 2.4 *Let* $\times$ *denote the linear space of* n n *matrices with complex entries over the field of complex numbers. Given any norm on the linear space , the* **subordinate matrix norm** *on* $\times$ *is defined by*

$$A = \max_{\square} \frac{A}{\quad},$$

where $\square = \mathbf{0}$.

It is easy to show that a subordinate matrix norm satisfies the axioms of norm listed in Definition 2.6; the details are left as an exercise. Definition 2.10 implies that, for A $\times$,

$$A \qquad A \qquad , \qquad \text{for all} \qquad .$$

In a relation like this any vector norm may be used, but of course it is necessary to use the same norm throughout. It follows from Definition 2.10 that, in any subordinate matrix norm on $\times$,

$$I = 1$$

where I is the n n identity matrix.

Given any vector in , it is a trivial matter to evaluate each of the three norms , , ∞; however, it is not yet obvious how one can calculate the corresponding subordinate matrix norm of a given matrix A in $\times$. Definition 2.10 is unhelpful in this respect: calculating A *via* (2.38) would involve the unpleasant task of maximising the function A / over $\square$ (or, equivalently, maximising A over the unit sphere : $= 1$). This difficulty is resolved by the following three theorems.

Theorem 2.7 *The matrix norm subordinate to the vector norm* $\|\cdot\|_\infty$ *can be expressed, for an* $n \times n$ *matrix* $A = (a_{ij})_{1 \le i,j \le n}$*, as*

$$\|A\|_\infty = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}|. \tag{2.39}$$

This result is often loosely expressed by saying that the ∞*-norm of a matrix is its largest row-sum.*

Proof Given an arbitrary vector v in $\mathbb{C}^n$, write $K = \|v\|_\infty$, so that $|v_j| \le K$ for $j = 1, 2, \ldots, n$. Then,

$$|(Av)_i| = \left|\sum_{j=1}^{n} a_{ij} v_j\right| \le \sum_{j=1}^{n} |a_{ij}|\,|v_j| \le K \sum_{j=1}^{n} |a_{ij}|, \qquad i = 1, 2, \ldots, n.$$

Now we define

$$C = \max_{1 \le i \le n} \sum_{j=1}^{n} |a_{ij}| \tag{2.40}$$

and note that

$$\frac{\|Av\|_\infty}{\|v\|_\infty} = \frac{\max_i |(Av)_i|}{\|v\|_\infty} \le \frac{\max_i |(Av)_i|}{K} \le C \qquad \forall\, v \in \mathbb{C}^n.$$

Hence, $\|A\|_\infty \le C$.

Next we show that $\|A\|_\infty \ge C$. To do so, we take v to be a vector in $\mathbb{C}^n$ each of whose entries is 1, with the choice of sign to be made clear below. In the definition of C, equation (2.40), let m be the value of i for which the maximum is attained, or any one of the values if there is more than one. Then, in the vector v we give the element v_j the same sign as that of a_{mj}; if a_{mj} happens to be zero, the choice of the sign of v_j is irrelevant. With this definition of v we see at once that

$$\|A\|_\infty = \max_i |(Av)_i| \ge |(Av)_m| = \sum_{j=1}^{n} |a_{mj} v_j| = \sum_{j=1}^{n} |a_{mj}| = C.$$

As $\|v\|_\infty = 1$, it follows that

$$\|A\|_\infty \ge C = C\,\|v\|_\infty,$$

which means that $\|A\|_\infty \ge C$. Hence $\|A\|_\infty = C$, as required. □

Theorem 2.8 *The matrix norm subordinate to the vector norm can be expressed, for an* $n \times n$ *matrix* $A = (a_{ij})_{1 \le i,j \le n} \in \mathbb{C}^{n \times n}$, *as*

$$\|A\|_1 = \max_{1 \le j \le n} \sum_{i=1}^{n} |a_{ij}| .$$

This is often loosely expressed by saying that the 1-norm of a matrix is its largest column-sum. The proof of this theorem is very similar to that of the previous one, and is left as an exercise (see Exercise 7). Note that Theorems 2.7 and 2.8 mean that the 1-norm of a matrix $A = (a_{ij})_{1 \le i,j \le n}$ is the ∞-norm of the transpose $A^T = (a_{ji})_{1 \le i,j \le n}$ of the matrix.

Before we state a characterisation of the subordinate matrix 2-norm, we recall the following definition from linear algebra.

Definition 2.11 *Suppose that* $A \in \mathbb{C}^{n \times n}$. *A complex number* λ, *for which the set of linear equations*

$$(A - \lambda I)x = 0$$

has a nontrivial solution $x \neq \mathbf{0}$, *is called an* **eigenvalue** *of* A; *the associated solution* x *is called an* **eigenvector** *of* A *(corresponding to* λ*).*

Now we are ready to state our result.

Theorem 2.9 *Let* $A \in \mathbb{C}^{n \times n}$ *and denote the eigenvalues of the matrix* $B = A^H A$ *by* λ_i, $i = 1, 2, \ldots, n$. *Then,*

$$\|A\|_2 = \max_{1 \le i \le n} \sqrt{\lambda_i} .$$

Proof Note first that the matrix B is symmetric, *i.e.*, $B = B^H$; therefore all of its eigenvalues are real and the associated eigenvectors belong to $\mathbb{R}^n$. (You may wish to prove this: consult the proof of Theorem 3.1, part (ii), for a hint.) Moreover, all eigenvalues of B are nonnegative, since if x is an eigenvector of B and λ is the associated eigenvalue λ, then

$$x^H A^H A x = x^H B x = \lambda x^H x$$

and therefore

$$\lambda = \frac{x^H A^H A x}{x^H x} = \frac{\|Ax\|_2^2}{\|x\|_2^2} \ge 0 .$$

Suppose that the vectors $x_i \neq 0$, $i = 1, 2, \ldots, n$, are eigenvectors of B corresponding to the eigenvalues λ_i, $i = 1, 2, \ldots, n$. Since B is symmetric

we may assume that the vectors are orthogonal, *i.e.*, $= 0$ for $i = j$, and we can normalise them so that $= 1$ for $i = 1, 2, \ldots, n$. Now choose an arbitrary vector in and express it as a linear combination of the vectors , $i = 1, 2, \ldots, n$:

$$= c + + c .$$

Then,

$$B = c \lambda + + c \lambda .$$

We may assume, without loss of generality, that

$$(0)\lambda \lambda \lambda .$$

Using the orthonormality of the vectors , $i = 1, 2, \ldots, n$, we get that

$$\begin{aligned} A &= A\ A = B \\ &= c \lambda + + c \lambda \\ &\quad (c + + c)\lambda \\ &= \lambda , \end{aligned} \tag{2.41}$$

for any vector . Hence $A \lambda$. To prove equality we simply choose = in (2.41), so that $c = = c - = 0$ and $c = 1$. □

The square roots of the (nonnegative) eigenvalues of $A\ A$ are referred to as the **singular values** of A. Thus we have shown that the 2-norm of a matrix A is equal to the largest singular value of A.

If the matrix A is symmetric, then $B = A\ A = A$, and the eigenvalues of B are just the squares of the eigenvalues of A. In this special case the 2-norm of A is the largest of the absolute values of its eigenvalues.

Theorem 2.10 *Given that is a subordinate matrix norm on* $^{\times}$ *,*

$$AB \quad A \quad B$$

for any two matrices A *and* B *in* $^{\times}$ *.*

Proof From the definition of subordinate matrix norm,

$$AB = \max \frac{AB}{}.$$

As

$$AB \quad A \quad B$$

while if $D = [0,1]$, *then* $\mathrm{Cond}(f) = +\infty$. *Indeed, in the latter case, perturbing* $x = 0$ *to* $x = \varepsilon$, $0 < \varepsilon \ll 1$, *leads to a perturbation of the function value* $f(0) = 0$ *to* $f(\varepsilon) = \varepsilon = \varepsilon \cdot 10^{-1}$ *: a magnification by a factor* 10^{-1} *in comparison with the size of the perturbation in* x.

When $\|f(y) - f(x)\| / \|y - x\|$ exhibits large variation as (x, y) ranges through $D \times D$, it is more helpful to consider a finer, local measure of conditioning, the **absolute local condition number**, at $x \in D$, of the function f, defined by

$$\mathrm{Cond}\,(f) = \sup_{\substack{\delta x \in V \setminus \{0\} \\ x + \delta x \in D}} \frac{\|f(x + \delta x) - f(x)\|}{\|\delta x\|}. \tag{2.43}$$

Example 2.6 *Let us consider the function* $f : x \in D \to \sqrt{x}$, *defined on the interval* $D = (0, \infty)$. *The absolute local condition number of* f *at* $x \in D$ *is* $\mathrm{Cond}\,(f) = 1/(2\sqrt{x})$. *Clearly,* $\lim_{x \to \infty} \mathrm{Cond}\,(f) = +\infty$, $\lim_{x \to \infty} \mathrm{Cond}\,(f) = 0$.

Although the definitions (2.42) and (2.43) seem intuitive, they are not always satisfactory from the practical point of view since they depend on the magnitudes of $f(x)$ and x. A more convenient definition of conditioning is arrived at by rescaling (2.43) by the norms of $f(x)$ and x. This leads us to the notion of **relative local condition number**

$$\mathrm{cond}\,(f) = \sup_{\substack{\delta x \in V \setminus \{0\} \\ x + \delta x \in D}} \frac{\|f(x + \delta x) - f(x)\| / \|f(x)\|}{\|\delta x\| / \|x\|},$$

where it is implicitly assumed that $x \neq 0$ and $f(x) \neq 0$. The next example highlights the difference between the absolute local condition number and the relative local condition number of f.

Example 2.7 *Let us consider the function* $f : x \in D \to \sqrt{x}$, *defined on the interval* $D = (0, \infty)$. *Recall from the preceding example that the absolute local condition number of* f *at* $x \in D$ *approaches* $+\infty$ *as* x *tends to zero. In contrast with this, the relative local condition number of* f *is* $\mathrm{cond}\,(f) = 1/2$ *for all* $x \in D$.

You may also wish to ponder the following, seemingly paradoxical, observation: $\lim_{x \to \pi} \mathrm{cond}\,(\sin) = 1$ *and* $\lim_{x \to \pi} \mathrm{cond}_{-}\,(\sin) = \infty$, *even though* $\sin 0 = \sin \pi = 0$ *and* $\mathrm{Cond}\,(\sin) = \mathrm{Cond}\,(\sin) = 1$.

Since the present section is concerned with the solution of the linear system $Ax = b$, where $A \in \mathbb{R}^{n \times n}$ is nonsingular and $b \neq 0$, let us

consider the relative local condition number of the mapping

$$\mathsf{A}^{-} \quad : \qquad \mathsf{A}^{-}$$

at $\square = \mathbf{0}$. We suppose that has been equipped with a vector norm and, since there is no danger of confusion, we denote the associated subordinate matrix norm by

There is a condition number for each norm; for example, if we use the 2-norm, then $\kappa(A) = \|A\|\,\|A^{-}\|$, and so on. Indeed, the size of the condition number of a matrix A $^{\times}$ is strongly dependent on the choice of the norm in . In order to illustrate the last point, let us consider the unit lower triangular matrix A $^{\times}$ defined by

$$A = \begin{matrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{matrix}, \tag{2.46}$$

and note that its inverse is

$$A^{-} = \begin{matrix} 1 & 0 & 0 & 0 & \dots & 0 \\ 1 & 1 & 0 & 0 & \dots & 0 \\ 1 & 0 & 1 & 0 & \dots & 0 \\ 1 & 0 & 0 & 1 & \dots & 0 \\ \dots & \dots & \dots & \dots & \dots & \dots \\ 1 & 0 & 0 & 0 & \dots & 1 \end{matrix}.$$

Since

$$\|A\| = n \quad \text{and} \quad \|A^{-}\| = n,$$

it follows that $\kappa(A) = n$. On the other hand,

$$\|A\|_{\infty} = 2 \quad \text{and} \quad \|A^{-}\|_{\infty} = 2.$$

so that $\kappa_{\infty}(A) = 4$ n $= \kappa(A)$ when n 1. (A question for the curious: how does the condition number $\kappa(A)$ of the matrix A in (2.46) depend on the size n of A? See Exercise 11.)

It is left as an exercise to show that for a symmetric matrix A (*i.e.*, when $A = A$), the 2-norm condition number $\kappa(A)$ is the ratio of the largest of the absolute values of the eigenvalues of A to the smallest of the absolute values of the eigenvalues (see Exercise 9).

6 8 1

A 8 1 A 8 1

$ 1 2 G 5 1 G

1 5 1 8 T 7 4@4

We can now assess the sensitivity of the solution of the system $A = $ to changes in the right-hand side vector .

Theorem 2.11 *Suppose that* A $^{\times}$ *is a nonsingular matrix,* , $A = $ *and* $A(+) = + $ *, with* , . *Then,* and

$$\text{——} \quad \kappa(A) \text{——} .$$

Proof Evidently,

$$= A \quad \text{and} \quad = A^{-}(+) \quad = A^{-} \quad .$$

As $= \mathbf{0}$ and A is nonsingular, the first of these implies that $= \mathbf{0}$. Further,

$$A \qquad \text{and} \qquad A^{-} \qquad .$$

The result follows immediately by multiplying these inequalities. □

Owing to the effect of rounding errors during the calculation, the numerical solution of $A = $ will not be exact. The numerical solution may be written $ + $, and we shall usually find that this vector satisfies the equation $A(+) = + $, where the elements of are very small. If the matrix A has a large condition number, however, the elements of may not be so small. An example of this will be presented in the next section.

2.8 Hilbert matrix

We consider the Hilbert matrix H of order n, whose elements are

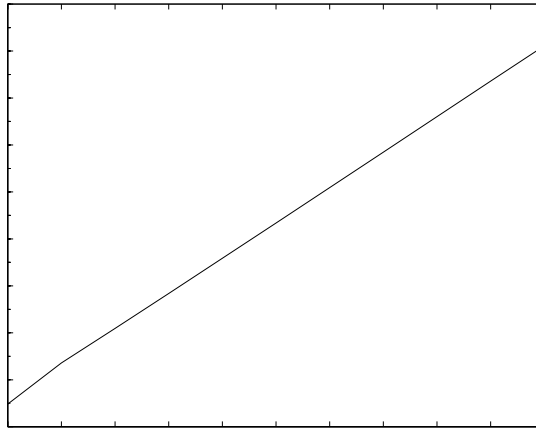
$$h = \frac{1}{i + j - 1}, \qquad i, j = 1, 2, \ldots, n .$$

This matrix is symmetric and positive definite (*i.e.*, $H = H$, and $H > 0$ for all), and therefore all of its eigenvalues are real and positive (cf. Theorem 3.1, part (ii)). However, H becomes very nearly singular as n increases. Table 2.1 shows the largest and smallest eigenvalues, and the 2-norm condition number $\kappa(H)$ of H, for various values of n.

1 / 5 0 B * K 6 #.? I B D C
9 6 #@ * FI F 6D 1
FI , 1 , 4 0 E, , 6
1 7 , , 6 6 , , 15
1 , 6 G 1 1 , 4

Table 2.1. *Eigenvalues and condition number of the Hilbert matrix* H .

	max	min	2*	+
(	%	&&	6 '	5
			13 %	13
(	&		21 %	20
	$	)	29 (	28
(		$	36	36



' 8 , 2* + 9 , ! : #
& " "

Figure 2.4 depicts the logarithm of the condition number κ (H) in the 2-norm of the Hilbert matrix H against its order, n; the straight line in our semilogarithmic-scale plot indicates that κ (H), as a function of n, exhibits exponential growth. Indeed, it can be shown that

$$\kappa\,(H\,) \quad \frac{\overline{2}+1}{2 \quad \overline{\pi n}} \quad \text{as } n \qquad .$$

We now define the vector with elements b = (j /(i + j 1)), i = 1, 2, . . . , n, chosen so that the solution of A = , with A = H , is the vector with elements x = i, i = 1, 2, . . . , n. We obtain a numerical solution of the system, using the method described in Section

2.5 to give the calculated vector + , and then compute the residual from A(+) = + . The calculation uses arithmetic operations correct to 15 decimal digits, which is roughly the accuracy used by many computer systems. The results are listed in Table 2.2.

Table 2.2. *Rounding errors in the solution of* H = , *where* H *is the Hilbert matrix of order* n *and* = (1, 2, . . . , n) .

	2	2		2	2
(		15	(		11
)		15	&		3
(		15	'		
% &		15	)		
($		13	((		2

The relative size of the residual is, in nearly every case, about the size of the basic rounding error, 10⁻ . The resulting errors in are smaller than the bound given by Theorem 2.11, as might be expected, since that bound corresponds to the worst possible case. In any case, for the Hilbert matrix of order greater than 14 the error is larger than the calculated solution itself, which renders the calculated solution meaningless. For matrices of this kind the condition number and the bound given by Theorem 2.11 are so large that they have little practical relevance, though they do indicate that, due to sensitivity to rounding errors, the numerical calculations are of unreliable accuracy.

The Hilbert matrix is, of course, a rather extreme example of an ill-conditioned matrix. However, we shall meet it in an important problem in Section 9.3 concerning the least squares approximation of a function by polynomials, where we shall see how a reformulation of the problem using an orthonormal basis avoids the disastrous loss of accuracy that would otherwise occur. In the next section, we introduce the idea of least squares approximation in the context of linear algebra and consider the solution of the resulting system of linear equations using the QR algorithm; this, too, relies on the notion of (ortho)normalisation.

2.9 Least squares method

Up to now, we have been dealing with systems of linear equations of the form A = where A × . However, it is frequently the case

in practical problems (typically, in problems of data-fitting) that the matrix A is not square but rectangular, and we have to solve a linear system of equations $A = $ with A $^{\times}$, , with $m > n$; since there are more equations than unknowns, in general such a system will have no solution. Consider, for example, the linear system (with $m = 3$, $n = 2$)

$$\begin{matrix} 3 & 1 \\ 1 & 1 \\ 4 & 2 \end{matrix} \quad \begin{matrix} x \\ x \end{matrix} \quad = \quad \begin{matrix} 1 \\ 0 \\ 2 \end{matrix} \quad ;$$

by adding the first two of the three equations and comparing the result with the third, it is easily seen that there is no solution. If, on the other hand, $m < n$, then the situation is reversed and there may be an infinite number of solutions. Consider, for example, the linear system (with $m = 1$, $n = 2$)

$$(3 \;\; 1) \quad \begin{matrix} x \\ x \end{matrix} \quad = 1 \, ;$$

any vector $= (\mu, 1 \quad 3\mu)$, with μ , is a solution to this system.

Suppose that m n; we may then need to find a vector which satisfies A $\mathbf{0}$ in as nearly as possible in some sense. This suggests that we define the residual vector $= A$ and require to minimise a certain norm of in . From the practical point of view, it is particularly convenient to minimise the residual vector in the 2-norm on ; this leads to the **least squares** problem:

$$\text{Minimise} \quad A \qquad .$$

This is clearly equivalent to minimising the square of the norm; so, on noting that

$$A \qquad = (A \qquad) \; (A \qquad) \, ,$$

the problem may be restated as

$$\text{Minimise} \quad (A \qquad) \; (A \qquad) \, .$$

Since

$$(A \qquad) \; (A \qquad) = \quad A \; A \quad 2 \quad A \quad + \quad ,$$

the quantity to be minimised is a nonnegative quadratic function of the n components of the vector ; the minimum therefore exists, and may

be found by equating to zero the partial derivatives with respect to the components. This leads to the system of equations

$$\mathsf{B} \quad = \mathsf{A} \quad , \qquad \text{where } \mathsf{B} = \mathsf{A} \ \mathsf{A} .$$

The matrix B is symmetric, and if A has full rank, n, then B is nonsingular; it is called the **normal** matrix, and the system $\mathsf{B} \quad = \mathsf{A} \quad$ is called the system of **normal equations**.

The normal equations have important theoretical properties, but do not lead to a satisfactory numerical algorithm, except for fairly small problems. The difficulty is that in a practical least squares problem the matrix A is likely to be quite ill-conditioned, and $\mathsf{B} = \mathsf{A} \ \mathsf{A}$ will then be extremely ill-conditioned. For example, if

$$\mathsf{A} = \begin{matrix} \varepsilon & 0 \\ 0 & 1 \end{matrix}$$

where $\varepsilon \quad (0, 1)$, then $\kappa \ (\mathsf{A}) = \varepsilon^{-} \ > 1$, while

$$\kappa \ (\mathsf{B}) = \kappa \ (\mathsf{A} \ \mathsf{A}) = \varepsilon^{-} \ = \varepsilon^{-} \ \kappa \ (\mathsf{A}) \quad \kappa \ (\mathsf{A})$$

when $0 < \varepsilon \quad 1$. If possible, one should avoid using a method which leads to such a dramatic deterioration of the condition number.

There are various alternative techniques which avoid the direct construction of the normal matrix $\mathsf{A} \ \mathsf{A}$, and so do not lead to this extreme ill-conditioning. Here we shall describe just one algorithm, which begins by factorising the matrix A, but using an orthogonal matrix rather than the lower triangular factor as in Section 2.3.

Theorem 2.12 *Suppose that* $\mathsf{A} \quad ^{\times}$ *where* $\mathsf{m} \quad \mathsf{n}$. *Then,* A *can be written in the form*

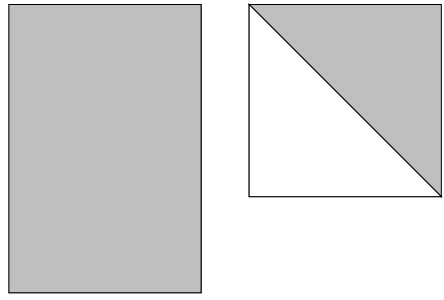
$$\mathsf{A} = \hat{\mathsf{Q}}\hat{\mathsf{R}} ,$$

where $\hat{\mathsf{R}}$ *is an upper triangular* $\mathsf{n} \quad \mathsf{n}$ *matrix, and* $\hat{\mathsf{Q}}$ *is an* $\mathsf{m} \quad \mathsf{n}$ *matrix which satisfies*

$$\hat{\mathsf{Q}} \ \hat{\mathsf{Q}} = \mathsf{I} \ , \tag{2.47}$$

where I *is the* $\mathsf{n} \quad \mathsf{n}$ *identity matrix; see Figure 2.5. If* $\operatorname{rank}(\mathsf{A}) = \mathsf{n}$, *then* $\hat{\mathsf{R}}$ *is nonsingular.*

Proof We use induction on n, the number of columns in A. The theorem clearly holds when $\mathsf{n} = 1$ so that A has only one column. Indeed, writing for this column vector and assuming that $= \mathbf{0}$, the matrix $\hat{\mathsf{Q}}$ has just



(;<
=T = # ! =

== =

one column, the vector / , and $\hat{R}$ has a single element, . In the special case where is the zero vector we can choose $\hat{R}$ to have the single element 0, and $\hat{Q}$ to have a single column which can be an arbitrary vector in whose 2-norm is equal to 1.

Suppose that the theorem is true when $n = k$, where $1 \quad k < m$. Consider a matrix A which has m rows and $k+1$ columns, partitioned as

$$A = (A \quad),$$

where is a column vector and A has k columns. To obtain the desired factorisation $\hat{Q}\hat{R}$ of A we seek $\hat{Q} = (\hat{Q} \quad)$ and

$$\hat{R} = \begin{matrix} \hat{R} \\ \mathbf{0} \quad \alpha \end{matrix}$$

such that

$$A = (A \quad) = (\hat{Q} \quad) \begin{matrix} \hat{R} \\ \mathbf{0} \quad \alpha \end{matrix} .$$

Multiplying this out and requiring that $\hat{Q}\ \hat{Q} = I$, the identity matrix of order $k+1$, we conclude that

$$\begin{aligned} A &= \hat{Q}\ \hat{R}\ , && (2.48)\\ &= \hat{Q} \quad + \alpha , && (2.49)\\ \hat{Q}\ \hat{Q} &= I\ , && (2.50)\\ \hat{Q} &= \mathbf{0}\ , && (2.51)\\ &= 1 . && (2.52) \end{aligned}$$

These equations show that $\hat{Q}\ \hat{R}$ is the factorisation of A, which exists by the inductive hypothesis, and then lead to

$$\begin{aligned} &= \hat{Q} \quad , \\ &= (1/\alpha)(\quad \hat{Q}\ \hat{Q} \quad), \end{aligned}$$

where $\alpha = \quad \hat{Q}\ \hat{Q} \quad$. The number α is the constant required to ensure that the vector is normalised.

The construction fails when $\quad \hat{Q}\ \hat{Q} \quad = \mathbf{0}$, for then the vector cannot be normalised. In this case we choose to be any normalised vector in which is orthogonal in to all the columns of $\hat{Q}$, for then $\quad \hat{Q} \quad = \mathbf{0}$ as required. The condition at the beginning of the proof, that $k < m$, is required by the fact that when $k = m$ the matrix $\hat{Q}$ is a square orthogonal matrix, and there is no vector in $\mathbf{0}$ such that $\quad \hat{Q} \quad = \mathbf{0}$.

With these definitions of , $, \alpha, \hat{Q}$ and $\hat{R}$ we have constructed the required factors of A, showing that the theorem is true when $n = k + 1$. Since it holds when $n = 1$ the induction is complete.

Now, for the final part, suppose that $\mathrm{rank}(A) = n$. If $\hat{R}$ were singular, there would exist a nonzero vector such that $\hat{R} \quad = \mathbf{0}$; then, $A \quad = \hat{Q}\hat{R} \quad = \mathbf{0}$, and hence $\mathrm{rank}(A) < n$, contradicting our hypothesis that $\mathrm{rank}(A) = n$. Therefore, if $\mathrm{rank}(A) = n$, then $\hat{R}$ is nonsingular. □

The matrix factorisation whose existence is asserted in Theorem 2.12 is called the **QR factorisation**. Here, we shall present its use in the solution of least squares problems. In Chapter 5 we shall revisit the idea in a different context which concerns the numerical solution of eigenvalue problems.

Theorem 2.13 *Suppose that* $A \quad ^{\times}$ *, with* $m \quad n$ *and* $\mathrm{rank}(A) = n$*, and let . Then, there exists a unique least squares solution of the system of equations* $A \ = \ :$ *a vector in which minimises the function* A *over all in . The vector can be obtained by finding the factors* $\hat{Q}$ *and* $\hat{R}$ *of* A *defined in Theorem 2.12, and then solving the nonsingular upper triangular system* $\hat{R} \quad = \hat{Q} \quad .$

Proof The matrix $\hat{Q}$ has m rows and n columns, with $m \quad n$, and it satisfies

$$\hat{Q}\ \hat{Q} = I \quad .$$

We shall suppose that $m > n$, the case $m = n$ being a trivial special case with

$$= A^- \quad = (\hat{Q}\hat{R})^- \quad = \hat{R}^- \hat{Q}^- \quad = \hat{R}^- \hat{Q} \quad ,$$

and hence $\hat{R} \quad = \hat{Q} \quad$, as required.

For $m > n$ now, the vector can be written as the sum of two vectors:

$$= \quad + \quad ,$$

where is in the linear space spanned by the n columns of the matrix $\hat{Q}$, and is in the orthogonal complement of this space in . The vector is a linear combination of the columns of $\hat{Q}$, and is orthogonal to every column of $\hat{Q}$; *i.e.*, there exists such that

$$= \quad + \quad , \qquad = \hat{Q} \quad , \quad \hat{Q} \qquad = \mathbf{0} \, . \tag{2.53}$$

Now, suppose that is the solution of $\hat{R} \quad = \hat{Q} \quad$, and that is any vector in . Then,

$$\begin{aligned} A \qquad &= \hat{Q}\hat{R} \\ &= \hat{Q}\hat{R}(\qquad) + \hat{Q}\hat{R} \\ &= \hat{Q}\hat{R}(\qquad) + \hat{Q}\hat{Q} \\ &= \hat{Q}\hat{R}(\qquad) + \hat{Q}\hat{Q} \qquad + \hat{Q}\hat{Q} \\ &= \hat{Q}\hat{R}(\qquad) + \hat{Q}\hat{Q} \quad \hat{Q} \\ &= \hat{Q}\hat{R}(\qquad) \qquad , \end{aligned}$$

where we have used (2.53) repeatedly; in particular, the last equality follows by noting that $\hat{Q} \quad \hat{Q} = I$. Hence

$$\begin{aligned} A \qquad &= (\qquad) \quad \hat{R} \quad \hat{Q} \quad \hat{Q}\hat{R}(\qquad) + \qquad 2(\qquad) \quad \hat{R} \quad \hat{Q} \\ &= \quad \hat{R}(\qquad) \quad + \end{aligned}$$

since $\hat{Q} \qquad = \mathbf{0}$. Thus $A \qquad$ is smallest when $\hat{R}(\qquad) = \mathbf{0}$, which implies that $= \quad$, since the matrix $\hat{R}$ is nonsingular. Hence , defined as the solution of $\hat{R} \quad = \hat{Q} \quad$, is the required least squares solution. □

2.10 Notes

There are many good books on the subject of numerical linear algebra which cover the topics discussed in this chapter in much greater detail,

and address questions which we have not touched on here. Without any attempt to be exhaustive, we single out four texts from the vast literature. The first two books on the list below are well-known monographs on the subject, while the last two are excellent textbooks.

, *Matrix Computations,* Third Edition, Johns Hopkins University Press, Baltimore, 1996.

, *Accuracy and Stability of Numerical Algorithms,* SIAM, Philadelphia, 1996.

! " ###, *Numerical Linear Algebra,* SIAM, Philadelphia, 1997.

$, *Introduction to Numerical Linear Algebra and Optimisation*, Cambridge University Press, Cambridge, 1989.

As we have already noted in Section 2.2, the invention of the elimination technique is attributed to Gauss who published the method in his *Theoria motus* (1809), although the idea was already known to the Chinese two thousand years ago. Gauss himself was concerned with positive definite systems. The method was extended to linear systems with general matrices by Jacobi. The interpretation of Gaussian elimination as matrix factorisation is due to P.S. Dwyer: A matrix presentation of least squares and correlation theory with matrix justification of improved methods of solutions, *Ann. Math. Stat.* **15**, 82–89, 1944.

The sensitivity of Gaussian elimination to rounding errors was studied by Wilkinson in Error analysis of direct methods of matrix inversion, *J. Assoc. Comput. Math.* **8**, 281–330, 1961. The idea of pivoting was used as early as 1947 by von Neumann and Goldstein. The concept of the condition number of a matrix was introduced by Turing in Rounding-off errors in matrix processes, *Quart. J. Mech. Appl. Math.* **1**, 287–308, 1948. Our treatment of condition numbers follows the textbook of Trefethen and Bau, cited above.

```
1      F    5K ,   K ,    B#  / ,       #.                    0  6          $
       B    F      6D C #. 9      6 #.>#         F      6D
  ,                      6 1       ,1  ,          J         G       4       $
                  6 F4  4          1K ,   N  F                    ,      5
  1        )
2 K      0    6     "       B &            #@#@                          C > H,
  #@.? :                 D4
3 K     5               B . / ,        #@ *             %      CO      6 B
   0      6D C . 9       6 #@>&               /       %D4
4  6    6 F         B* / ,       #@ * 0             C    K      6 #@.@
   '     ,           %D4
5 %     '               B * K     #@#   :                C & K     #@>
                  D4
```

Normed linear spaces play a key role in functional analysis (see, for example, K. Yosida, *Functional Analysis*, Third Edition, Springer, Berlin, 1971, page 30). Here, we have concentrated on finite-dimensional normed linear spaces over the field of real numbers.

The relevance of norms in numerical linear algebra was highlighted by Householder in his book *The Theory of Matrices in Numerical Analysis,* Blaisdell, New York, 1964.

The idea of least squares fitting is due to Gauss, who invented the method in the 1790s. However, it was the French mathematician Legendre who first published the method in 1806 in a book on determining the orbits of comets. Legendre's method involved a number of observations taken at equal intervals and he assumed that the comet followed a parabolic path, so he ended up with more equations than there were unknowns. Legendre then applied his methods to the data known for two comets. In an Appendix to the book Legendre described the least squares method of fitting a curve to the data available. Gauss published his version of the least squares method in 1809 and, although acknowledging that it had already appeared in Legendre's book, Gauss nevertheless claimed priority for himself. This greatly hurt Legendre, leading to one of the infamous priority disputes in the history of mathematics. A recent exhaustive monograph on numerical algorithms for least squares problems is due to Å. Björk: *Numerical Methods for Least Squares Problems,* SIAM, Philadelphia, 1996.

The version of the QR factorisation considered here is the *reduced version*, following the terminology in Chapter 7 of Trefethen and Bau. In the *full version* of the QR factorisation for a matrix A $^{\times}$, we have $\mathsf{A} = \mathsf{QR}$, where Q $^{\times}$, R $^{\times}$ (cf. Chapter 5).

In a footnote to Definition 2.12 we mentioned the Moore–Penrose generalised inverse A of a matrix A $^{\times}$. A can be defined through the singular value decomposition of A (cf. L.N. Trefethen and D. Bau, III: *Numerical Linear Algebra*, SIAM, Philadelphia, 1997). Recall that the singular values of A are the square roots of the (nonnegative) eigenvalues of the matrix A A.

1 % , 0 B> ' 6 #@ ,"1) % C K 6 #@@*
' 1 %D 1 1 , 4
0 N 6 6 F4 4 , ' (+ 5 1

2 % $' : B#. #&> 9 , C # K 6 #.**
9 , D4

Theorem 2.14 (Singular value decomposition) *Let* A $^{\times}$ *; then, there exist* U $^{\times}$, Σ $^{\times}$ *and* V $^{\times}$ *such that*

$$A = U\Sigma V \quad ,$$

where Σ *is a diagonal matrix whose diagonal entries,* σ , $i = 1, 2, \ldots, n$, *are the singular values of* A, U $U = I$ *and* V $V = I$ *, with* I *denoting the* n n *identity matrix.*

The Moore–Penrose generalised inverse of the diagonal matrix Σ $^{\times}$ is defined as the diagonal matrix Σ $^{\times}$ whose diagonal entries are

$$\sigma \quad = \quad \begin{array}{ll} \sigma^{-} & \text{if } \sigma \ = 0\,, \\ 0 & \text{if } \sigma \ = 0\,. \end{array}$$

The generalised inverse A $^{\times}$ of a matrix A $^{\times}$ with singular value decomposition $A = U\Sigma V$ is defined by

$$A \quad = V \Sigma \ U \quad .$$

In the special case when $m = n$ and A $^{\times}$ is nonsingular, the n singular values of A are all nonzero and therefore $\Sigma \ = \Sigma^{-}$. Hence, also, $A \ = A^{-}$, which then justifies the use of the terminology 'generalised inverse' for the matrix A defined above.

Exercises

2.1 Let n 2. Given the matrix $A = (a \ )$ $^{\times}$, the permutation matrix Q $^{\times}$ reverses the order of the rows of A, so that $(QA) \ = a \ _{-}$. If L $^{\times}$ is a lower triangular matrix, what is the structure of the matrix QLQ?

Show how to factorise A $^{\times}$ in the form $A = UL$, where U $^{\times}$ is unit upper triangular and L $^{\times}$ is lower triangular. What conditions on A will ensure that the factorisation exists? Give an example of a square matrix A which cannot be factorised in this way.

2.2 Let n 2. Consider a matrix A $^{\times}$ whose every leading principal submatrix of order less than n is nonsingular. Show that A can be factored in the form $A = LDU$, where L $^{\times}$ is unit lower triangular, D $^{\times}$ is diagonal and U $^{\times}$ is unit upper triangular.

If the factorisation $A = LU$ is known, where L is unit lower

triangular and U is upper triangular, show how to find the factors of the transpose A .

2.3 Let n 2 and suppose that the matrix A $^{\times}$ is nonsingular. Show by induction, as in Theorem 2.3, that there are a permutation matrix P $^{\times}$, a lower triangular matrix L $^{\times}$, and a unit upper triangular matrix U $^{\times}$ such that $PA = LU$.

By finding a suitable 2 2 matrix A, or otherwise, show that this may not be true if A is singular.

2.4 The lower triangular matrix L $^{\times}$, n 2, is nonsingular, and the vector is such that b $= 0,\ i = 1, 2, \ldots, k$, with 1 k n. The vector is the solution of L $=$. Show, by partitioning L, that y $= 0,\ j = 1, 2, \ldots, k$. Hence give an alternative proof of Theorem 2.1(iv), that the inverse of a nonsingular lower triangular matrix is itself lower triangular.

2.5 Given a matrix A $^{\times}$, define the matrix B $^{\times}$ in which the first n columns are the columns of A, and the last n columns are the columns of the identity matrix I . Consider the following computational scheme. Treat the rows of the matrix B in order, so that $j = 1, 2, \ldots, n$. Multiply every element in row j by the reciprocal of the diagonal element, $1/b$; then, replace every element b which is not in row j, so that $i = j$, by b b b .

Show that the result is equivalent to multiplying B on the left by a sequence of matrices. Explain why, at the end of the computation, the first n columns of B are the columns of the identity matrix I , and the last n columns are the columns of the inverse matrix A^{-} . Give a condition on the matrix A which will ensure that the computation does not break down.

Show that the process as described requires approximately $2n$ multiplications, but that, if the multiplications in which one of the factors is zero are not counted, the total is approximately n .

2.6 Use the method of Exercise 5 to find the inverse of the matrix

$$A = \begin{matrix} 2 & 4 & 2 \\ 1 & 0 & 3 \\ 3 & 1 & 2 \end{matrix} .$$

2.7 Suppose that for a matrix A $\times$,

$$a \quad C, \quad j = 1, 2, \ldots, n.$$

Show that, for any vector ,

$$(A \ ) \quad C \quad .$$

Find a nonzero vector for which equality can be achieved, and deduce that

$$A \quad = \max \quad a \quad .$$

2.8 (i) Show that, for any vector $= (v, \ldots, v)$,

$$\infty \quad \text{and} \quad \infty .$$

In each case give an example of a nonzero vector for which equality is attained. Deduce that ∞ . Show also that $\overline{n} \quad \infty$.
(ii) Show that, for any matrix A $\times$,

$$A \ \infty \quad \overline{n} \ A \quad \text{and} \quad A \quad \overline{m} \ A \ \infty .$$

In each case give an example of a matrix A for which equality is attained. (See the footnote following Definition 2.12 for the meaning of A , A and $A \ \infty$ when A $\times$.)

2.9 Prove that, for any nonsingular matrix A $\times$,

$$\kappa \ (A) = \frac{\lambda}{\lambda} \quad ,$$

where λ is the smallest and λ is the largest eigenvalue of the matrix $A \ A$.

Show that the condition number $\kappa \ (Q)$ of an orthogonal matrix Q is equal to 1. Conversely, if $\kappa \ (A) = 1$ for the matrix A, show that all the eigenvalues of $A \ A$ are equal; deduce that A is a scalar multiple of an orthogonal matrix.

2.10 Let A $\times$. Show that if λ is an eigenvalue of $A \ A$, then

$$0 \quad \lambda \quad A \quad A \ ,$$

provided that the same subordinate matrix norm is used for

both A and A . Hence show that, for any nonsingular n n matrix A,

$$\kappa\ (A) \qquad \kappa\ (A)\,\kappa_\infty(A) \qquad .$$

2.11 For the matrix defined by (2.46) write down the matrix A A. Show that any vector $= \mathbf{0}$ is an eigenvector of A A with eigenvalue $\lambda = 1$, provided that x $= 0$ and x $+$ $+ x$ $= 0$. Show also that there are two eigenvectors with x $=$ $= x$ and find the corresponding eigenvalues. Deduce that

$$\kappa\ (A) = -(n+1)\ \ 1 + \sqrt{1 \quad \text{———}} \quad .$$

2.12 Let B $^{\times}$ and denote by I the identity matrix of order n. Show that if the matrix I B is singular, then there exists a nonzero vector such that $(I \quad B)$ $= \mathbf{0}$; deduce that B 1, and hence that, if A < 1, then the matrix I A is nonsingular.

Now suppose that A $^{\times}$ with A < 1. Show that

$$(I \quad A)^{-} = I + A(I \quad A)^{-} \ ,$$

and hence that

$$(I \quad A)^{-} \qquad 1 + A \quad (I \quad A)^{-} \quad .$$

Deduce that

$$(I \quad A)^{-} \qquad \frac{1}{1 \quad A} \, .$$

2.13 Let A $^{\times}$ be a nonsingular matrix and $\square$. Suppose that A $=$ and $(A+\delta A)(\ +\ ) =$, and that A^{-} δA < 1. Use the result of Exercise 12 to show that

$$\text{———} \quad \frac{A^{-}\ \delta A}{1 \quad A^{-}\ \delta A} \, .$$

2.14 Suppose that A $^{\times}$ is a nonsingular matrix, and $\square$. Given that A $=$ and $A(\ +\ ) =$ $+$, Theorem 2.11 states that

$$\text{———} \quad \kappa(A)\text{———} \, .$$

By considering the eigenvectors of A A, show how to find vectors and for which equality is attained, when using the 2-norm.

2.15 Find the QR factorisation of the matrix

$$\mathsf{A} = \begin{matrix} 9 & 6 \\ 12 & 8 \\ 0 & 20 \end{matrix},$$

and hence find the least squares solution of the system of linear equations

$$\begin{aligned} 9\mathsf{x} \quad 6\mathsf{y} &= 300\,, \\ 12\mathsf{x} \quad 8\mathsf{y} &= 600\,, \\ 20\mathsf{y} &= 900\,. \end{aligned}$$

3

Special matrices

3.1 Introduction

In this chapter we show how one can modify the elimination method for the solution of $A = $ when the matrix A has certain special properties. In particular when $A \quad ^{\times}$ is symmetric and positive definite the amount of computational work can be halved. For matrices with a band structure, having nonzero elements only in positions close to the diagonal, the efficiency can be improved even more dramatically.

3.2 Symmetric positive definite matrices

Definition 3.1 *The matrix* $A = (a \quad) \quad ^{\times}$ *is said to be* **symmetric** *if* $a \quad = a \quad$ *for all* i *and* j *in the set* $1, 2, \ldots, n$ *; i.e., if* $A = A \quad .$ *The set of all symmetric matrices* $A \quad ^{\times}$ *will be denoted by* $^{\times} \quad .$ *A matrix* $A \quad ^{\times}$ *is called* **positive definite** *if*

$$A \quad > 0$$

for every vector $\square = \quad \mathbf{0} \quad .$

Example 3.1 *Consider the matrix* $A \quad ^{\times}$ *,*

$$A = \begin{matrix} a & b \\ c & d \end{matrix}$$

and a vector $= (x \ , x \ ) \quad \square = \quad \mathbf{0} \quad .$

Clearly, $A \quad = ax \ + (b + c)x \ x \ + dx \ .$ The quadratic form on the right-hand side is positive for all real numbers $x \ , \ x \ $ such that

$= (x\ , x\ ) = (0,0) = \mathbf{0}$ if, and only if,

$$a > 0\,,\quad d > 0 \text{ and } (b+c) < 4ad\,.$$

We see that if A $^{\times}$ is positive definite, then the diagonal elements of A are positive. Further, noting that the third inequality can be rewritten as

$$(b\quad c) < 4(ad\quad bc) = 4\,\det(A)\,,$$

we deduce that the determinant of a positive definite matrix A $^{\times}$ is positive. This, of course, is still true in the special case when A $^{\times}$, *i.e.*, when $b = c$.

The next theorem extends the observations of the last example to any symmetric positive definite matrix A $^{\times}$.

Theorem 3.1 *Suppose that* $n \quad 2$ *and* $A = (a\ )$ $^{\times}$ *is positive definite; then:*

(i) *all the diagonal elements of* A *are positive, that is,* $a > 0$*, for* $i = 1, 2, \ldots, n$*;*

(ii) *all the eigenvalues of* A *are real and positive, and the eigenvectors of* A *belong to* $\square$*;*

(iii) *the determinant of* A *is positive;*

(iv) *every submatrix* B *of* A *obtained by deleting any set of rows and the corresponding set of columns from* A *is symmetric and positive definite; in particular, every leading principal submatrix is positive definite;*

(v) $a < a\ a$ *for all* i *and* j *in* $1, 2, \ldots, n$ *such that* $i = j$*;*

(vi) *the element of* A *with largest absolute value lies on the diagonal;*

(vii) *if* α *is the largest of the diagonal elements of* A*, then*

$$a \quad \alpha \quad i, j \quad 1, 2, \ldots, n\ .$$

Proof (i) Consider the vector with only one nonzero element, in position $i \quad 1, 2, \ldots, n$. Since A is positive definite and $\square$, it follows that $x\ a\ x = A > 0$, and therefore $a > 0$.

(ii) Suppose that λ is an eigenvalue of A and let $\square = \mathbf{0}$ denote the associated eigenvector. Further, let $^{-}$ denote the vector in $\square$ whose ith element is the complex conjugate of the ith element of

, $i = 1, 2, \ldots, n$. As $A \quad = \lambda \quad$, it follows that $^{-} \quad A \quad = \lambda(^{-} \quad)$, and therefore, using the symmetry of A,

$$A^{-} = \quad A \quad ^{-} = (^{-} \quad A \quad) \quad = (\lambda(^{-} \quad)) \quad = \lambda(\quad ^{-}).$$

Complex conjugation then yields $^{-} \quad A \quad = \bar{\lambda}(^{-} \quad)$, and hence $\lambda(^{-} \quad) = \bar{\lambda}(^{-} \quad)$. As $\quad = \mathbf{0}$, it follows that $\lambda = \bar{\lambda}$; i.e., λ is a real number.

The fact that the eigenvector associated with λ has real elements follows by noting that all elements of the singular matrix $A \quad \lambda I$ are real numbers. Therefore, the column vectors of $A \quad \lambda I$ are linearly dependent in $\quad$. Hence there exist n real numbers $x \ , \ldots , x \quad$ such that $(A \quad \lambda I) \quad = \mathbf{0}$, where $\quad = (x \ , \ldots , x \ ) \ .$

Finally, as $A \quad = \lambda \quad$ with $\lambda \quad$ and $\quad \square$, we have that $\quad A \quad = \lambda \quad$. Since $\lambda = \quad A \ / \quad$ and A is positive definite, λ is the ratio of two positive real numbers and therefore also real and positive.

(iii) This follows from the fact that the determinant of A is equal to the product of its eigenvalues, and the previous result. Indeed, since A is symmetric, there exist an orthogonal matrix X and a diagonal matrix Λ, whose diagonal elements are the eigenvalues $\lambda \ , \ i = 1, 2, \ldots, n$, of A, such that $A = X \ \Lambda X = X^{-} \ \Lambda X$. By the Binet–Cauchy Theorem (see Chapter 2, end of Section 2.3),

$$\begin{aligned} \det(A) \quad &= \quad \det(X^{-} \ ) \ \det(\Lambda) \ \det(X) \\ &= \quad \frac{1}{\det(X)} \ \det(\Lambda) \ \det(X) \\ &= \quad \det(\Lambda) = \lambda \ \ldots \lambda \quad > 0 \, . \end{aligned}$$

(iv) Consider the vector $\quad \square$ with zeros in the positions corresponding to the rows which have been deleted. Then,

$$A \quad = \quad B$$

where B is the submatrix of A containing the rows and columns which remain after deletion, and $\quad$ is the vector consisting of the elements of $\quad$ which were not deleted. Since the expression on the left is positive, the same is true of the expression on the right, for all vectors $\quad$ except the zero vector. Therefore B is positive definite.

(v) By the previous result the $2 \quad 2$ submatrix consisting of rows and columns r and s of A is positive definite, and its determinant is therefore positive.

(vi) This follows from the previous result, since it shows that $a \quad$ cannot exceed the greater of $a \quad$ and $a \quad$.

(vii) This follows at once from the previous result. $\square$

The converses of two of these results are also true:

(i) If all the eigenvalues of the symmetric matrix $A \quad ^{\times}$ are positive, then A is positive definite;
(ii) If the determinant of each leading principal submatrix of a matrix $A \quad ^{\times}$ is positive, then A is positive definite.

The proof of the second result is involved and will not be given here; see, however, Example 3.1 for the case of $n = 2$. The proof of the first statement, on the other hand, is quite simple and proceeds as follows.

Since $A \quad ^{\times}$ is symmetric, it has a complete set of orthonormal eigenvectors $\quad, \ldots, \quad$ in $\square$, and the corresponding eigenvalues $\lambda \ , \ldots, \lambda$ are all real. Given any vector $\square$, it can be expressed as

$$= \quad \alpha$$

where $\alpha \quad , \ i = 1, 2, \ldots, n$, and $\alpha \ + \quad + \alpha \ = \quad > 0$. Since $A \quad = \lambda \quad , \ i = 1, 2, \ldots, n$, it follows that

$$A \ = \quad \alpha \ \lambda \quad .$$

As $\quad = 0$ for $i = j$ and $\quad = 1$, we deduce that

$$A \quad = \quad \lambda \ \alpha$$

$$\min \lambda \quad \alpha \ > 0 \, ,$$

since min $\lambda \ > 0$; therefore A is positive definite.

For a symmetric positive definite matrix A we can now obtain an LU factorisation $A = LU$ in which $U = L \ .$

Theorem 3.2 *Suppose that* $n \quad 2$ *and* $A \quad ^{\times}$ *is a positive definite matrix; then, there exists a lower triangular matrix* $L \quad ^{\times}$ *such that*

$$A = LL \ .$$

This is known as the **Cholesky factorisation** *of* A.

1 9 4%4 0 4 4 K ' %
5 6 #@@ &4 4>4
2 P% -$: "6 B#.&>C#@#.D 9 , 6 S, 5 5
6 5 6 %1 , 7 1)4 0

Proof Since A is symmetric and positive definite, all the leading principal submatrices of A are positive definite, and hence by Theorem 2.2 the usual LU factorisation exists, with

$$A = L \quad U \quad ,$$

L × a unit lower triangular and U × an upper triangular matrix. In this factorisation the product of the leading principal submatrices of L and U of order k is the leading principal submatrix of A of order k, 1 k n. Since the determinant of this submatrix is positive and all the diagonal elements of L are unity, it follows that

$$u \quad u \quad \ldots u \quad > 0 , \quad k = 1, 2, \ldots , n .$$

Thus all the diagonal elements of U are positive. If we now define D to be the diagonal matrix with elements d = u , i = 1, 2, . . . , n, we can write

$$A = L \quad U \quad = (L \quad D)(D^{-} U \quad) = LU ,$$

where now l = u = u . The symmetry of the matrix A shows that

$$LU = A = A \quad = U \quad L \quad ,$$

so that

$$U(L \quad)^{-} \quad = L^{-} U \quad .$$

In this equality the left-hand side is upper triangular, and the right-hand side is lower triangular, and hence both sides must be diagonal. Therefore, U = D□L , where D□ is a diagonal matrix; but U and L have the same diagonal elements, so D□ = I and U = L .

The same argument shows that L and L are unique, except for the arbitrary choice of the signs of the square roots in the definition of the diagonal matrix D. If we make the natural choice, taking all the square roots to be positive, then the diagonal elements of L are positive, and the factorisation is unique. □

5 1 ,
G 1 G E 64 0 "
6 1 #@ 6 1 S,
23 3 / 4N C 5 ' (1) 9 6 #. #@@ (
@) &

In practice we construct the elements of L directly, rather than forming L and U first. This is done in a similar way to the LU factorisation. Suppose that $i \geq j$; we then require that

$$a = \quad l \, l \, , \qquad 1 \quad i \quad j \quad n . \tag{3.1}$$

Note that we have used the fact that $(L^{T})_{kj} = l_{jk}$; the sum only extends up to $k = i$ since L is lower triangular. The same equation will also hold for $i > j$, since A is symmetric. For $i = j$, equation (3.1) gives

$$l \quad = a \quad , \qquad l \quad = \quad a \qquad \bar{} \quad l \qquad , \qquad 1 < i \quad n . \tag{3.2}$$

As A is a positive definite matrix, $a > 0$ and therefore l is a positive real number. Further, as we have seen in the proof of the preceding theorem, $l > 0$, $i = 2, 3, \ldots, n$. We find similarly that

$$l \quad = \frac{1}{l} \quad a \qquad \bar{} \quad l \, l \qquad , \qquad 1 \quad i < j \quad n . \tag{3.3}$$

These equations now enable us to calculate the elements of L in succession. For each $i \quad 1, 2, \ldots, n \quad 1$, we first calculate l from (3.2), and then calculate $l \quad , \; l \quad , \ldots, l$ from (3.3). Finally, we compute l using (3.2).

As, by hypothesis, the matrix $A \qquad ^{\times}$ is positive definite, the required factorisation exists, so we can be sure that the divisor l in (3.3), and the expression in the curly brackets in (3.2) whose square root is taken, will be positive. Thus, (3.2) implies that

$$l \quad = a \quad , \qquad \bar{\max} \, l \qquad a \quad , \qquad i = 2, 3, \ldots, n .$$

The elements of the factor L cannot therefore grow very large, and no pivoting is necessary.

The evaluation of l from (3.2) requires $i \quad 1$ multiplications, $i \quad 1$ subtractions and one square root operation, a total of $2i \quad 1$ operations. The calculation of each l from (3.3) also requires $2i \quad 1$ operations. The total number of operations required to construct L is therefore

$$(2i \quad 1) = \qquad (2i \quad 1)(1 + n \quad i) = \; -n(n+1)(2n+1) .$$

For large n the number of operations required is approximately $-n$, which, as might be expected, is half the number given in Section 2.6 for the LU factorisation of a nonsymmetric matrix.

3.3 Tridiagonal and band matrices

As we shall see in the final chapters, in the numerical solution of boundary value problems for second-order differential equations one encounters a particular kind of matrix whose elements are mostly zeros, except for those along its main diagonal and the two adjacent diagonals. Matrices of this kind are referred to as tridiagonal. In order to motivate the definition of tridiagonal matrix stated in Definition 3.2 below, we begin with an example which is discussed in more detail in Chapter 13.

Example 3.2 *Consider the two-point boundary value problem*

$$\frac{\mathrm{d}\,y}{\mathrm{d}x} + r(x)y = f(x), \qquad x \in (0,1),$$

$$y(0) = 0, \quad y(1) = 0.$$

where r *and* f *are continuous functions of* x *defined on the interval* $[0,1]$.

The numerical solution of the boundary value problem proceeds by selecting an integer $n \geq 4$, choosing a step size $h = 1/n$, and subdividing the interval $[0,1]$ by the points $x_k = kh$, $k = 0, 1, \ldots, n$. The numerical approximation to $y(x_k)$, the value of the analytical solution y at the point $x = x_k$, is denoted by Y_k. The values Y_k are obtained by solving the set of linear equations

$$\frac{Y_{k-1} - 2Y_k + Y_{k+1}}{h} + r(x_k)Y_k = f(x_k)$$

for $k = 1, 2, \ldots, n-1$, together with the boundary conditions

$$Y_0 = 0, \quad Y_n = 0.$$

Equivalently,

$$\begin{aligned} a_k Y_{k-1} + c_k Y_k + b_k Y_{k+1} &= d_k, \qquad k = 1, 2, \ldots, n-1, \\ Y_0 &= 0, \quad Y_n = 0, \end{aligned}$$

where

$$a = b = 1/h, \quad c = 2/h + r(x), \quad d = f(x),$$

for $k = 1, 2, \ldots, n-1$.

Clearly, for $1 < k < n-1$, the kth equation in the linear system above involves only three of the $n-1$ unknowns: Y_{k-1}, Y_k and Y_{k+1}.

The example motivates the following definition of a tridiagonal (or triple diagonal) matrix.

Definition 3.2 *Suppose that* $n \geq 3$. *A matrix* $T = (t_{ij}) \in \mathbb{R}^{n \times n}$ *is said to be* **tridiagonal** *if it has nonzero elements only on the main diagonal and the two adjacent diagonals; i.e.,*

$$t_{ij} = 0 \quad \text{if} \quad |i - j| > 1, \qquad i, j = 1, 2, \ldots, n.$$

Such matrices are also sometimes called **triple diagonal**.

It is easy to see that in the LU factorisation process of a tridiagonal matrix $T \in \mathbb{R}^{n \times n}$, without row interchanges, the unit lower triangular matrix $L \in \mathbb{R}^{n \times n}$ and the upper triangular matrix $U \in \mathbb{R}^{n \times n}$ each have only two elements in each row. Writing T in the compact notation

$$T = \begin{bmatrix} b & c & & & & & \\ a & b & c & & & & \\ & a & b & c & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \ldots & \\ & & & & a & b \end{bmatrix}, \tag{3.4}$$

the factorisation may be written $T = LU$ where

$$L = \begin{bmatrix} 1 & & & & & \\ l & 1 & & & & \\ & l & 1 & & & \\ \ldots & \ldots & \ldots & \ldots & \ldots & \\ & & & l & 1 \end{bmatrix} \tag{3.5}$$

and

$$U = \begin{bmatrix} u & v & & & \\ & u & v & & \\ & & u & v & \\ \ldots & \ldots & \ldots & \ldots & \ldots \\ & & & & u \end{bmatrix}, \tag{3.6}$$

with the convention that the missing elements in these matrices are all equal to zero. It is often convenient to define $a = 0$ and $c = 0$. Multiplying L and U shows that $v = c$, and that the elements l and u can be calculated from

$$l = a/u_{-}, \quad u = b - lc_{-}, \qquad j = 2, 3, \ldots, n, \tag{3.7}$$

starting from $u = b$.

Let us suppose that our aim is to solve the system of linear equations $T = $, where the matrix T $^{\times}$ is tridiagonal and nonsingular, and . Having calculated the elements of the matrices L and U in the LU factorisation $T = LU$ using (3.7), the forward and backsubstitution are then also very simple. Letting $= U$, the equation $L = $ gives

$$y = r, \tag{3.8}$$

$$y = r - ly_{-}, \qquad j = 2, 3, \ldots, n, \tag{3.9}$$

and finally from $U = $ we get

$$x = y/u, \tag{3.10}$$

$$x = (y - vx)/u, \qquad j = n-1, n-2, \ldots, 1. \tag{3.11}$$

The LU factorisation of a tridiagonal matrix requires approximately $3n$ operations. The forward and backsubstitution together involve approximately $5n$ operations. Thus, the whole solution process requires approximately $8n$ operations. The total amount of work is therefore far less than for a full matrix, being of order n for large n, compared with $-n$ for a full matrix. The method we have described is a minor variation on what is often known as the *Thomas algorithm*.

So far we have assumed that pivoting was not necessary; clearly any interchange of rows will destroy the tridiagonal structure of T. However, it is easy to see that the only interchanges required will be between two adjacent rows.

Theorem 3.3 *Suppose that* $n \quad 3$ *and* T $^{\times}$ *is a tridiagonal matrix; then, there exists a permutation matrix* P $^{\times}$ *such that*

$$PA = L \quad U \tag{3.12}$$

1 %1 : 6 04 6 , #@> $
5 6) 'N , : 64 0
6 " , , C9 , 4
6 P N , 1 / 5 ! 4 :404
.) # ./ (+ , 4
4 : 4 5 6 ! " #@ @4 (1) (4@?
@

by the condition of strict diagonal dominance (3.13), which then shows that . holds. That completes the inductive step.

We have thus proved that $u > c$ for all $j = 1, 2, \ldots, n$. In particular, we deduce that $u = 0$ for all $j \quad 1, 2, \ldots, n$; hence the LU factorisation $T = LU$ defined by (3.7) exists. Further,

$$\det(T) = \det(L)\det(U) = \det(U) = u\ u\ \ldots u\ = 0\,,$$

so T is nonsingular.

The formula (3.7) and the inequalities $u > c$, $j = 1, 2, \ldots, n$, now imply that

$$\begin{aligned} u \quad & b + l\ c_- \\ = \quad & b + a\ c_-\ / u_- \\ & b + a\ , \qquad j = 1, 2, \ldots, n\,, \end{aligned} \tag{3.15}$$

so the elements u cannot grow large, and rounding errors are kept under control without pivoting. □

It is easy to see that the same result holds under the weaker assumption that the matrix is diagonally dominant, but not necessarily strictly diagonally dominant, provided that we also require that all the elements c , $j = 1, 2, \ldots, n$ 1, are nonzero (see Exercise 5).

Note also that the matrix constructed in Example 3.2 satisfies this condition, provided that the function r is nonnegative; this often holds in practical boundary value problems.

If the matrix T $^{\times}$ is symmetric and positive definite, as well as tridiagonal, it can be factorised in the form $T = LL$, where L $^{\times}$ is lower triangular with nonzero elements only on and immediately below the diagonal. If we use the notation $d = l$, $e = l_-$ we easily find from (3.2) and (3.3) that the elements can be calculated in succession from the following formulae:

$$\begin{aligned} d \quad &= \quad b \quad , \\ e \quad &= \quad c_-\ /d_- \ , \quad d = b \quad e \qquad , \qquad i = 2, 3, \ldots, n\,. \end{aligned}$$

This calculation involves about $4n$ operations. Including also the work required by the forward and backsubstitution stages, the complete solution of $T \quad =$ will be found to involve about $10n$ operations. For the tridiagonal matrix the Cholesky factorisation method thus requires more work for the complete solution than the Thomas algorithm; in this case there is no particular advantage in exploiting the symmetry of the matrix in this way.

```
   &   5       >            &%   :
2  *  +    !
```

More generally, a system of equations may often involve a matrix of band type.

Definition 3.4 B × *is a* **band matrix** *if there exist nonnegative integers* p < n *and* q < n *such that* b = 0 *for all* i, j 1, 2, . . . , n *such that* p < i j *or* q < j i. *The band is of width* p + q + 1, *with* p *elements to the left of the diagonal and* q *elements to the right of the diagonal, in each row. Such a matrix is said to be* Band(p, q).

Thus, for example, a tridiagonal matrix is Band(1,1), and an n n lower triangular matrix is Band(n 1,0).

An example of a Band(1,2) matrix A × is shown in Figure 3.1, where each nonzero element in the matrix is identified by an asterisk. In addition to its main diagonal, the matrix has nonzero elements on its lower subdiagonal and two of its superdiagonals.

It is easy to see that, provided that no interchanges are necessary, such a band matrix can be written in the form B = LU, where L is Band(p,0) and U is Band(0,q) (see Exercise 7). It is also fairly simple to count the operations required in this calculation; the result is approximately proportional to np(p + 2q) when n is moderately large. The most common situation has q = p, and then the number of operations is approximately proportional to np . As in the tridiagonal case, this is much smaller than n when p and q are fairly small compared with n.

3.4 Monotone matrices

If a positive real number a is increased by ε > 0 to a + ε, then its reciprocal a⁻ decreases to (a + ε)⁻ . It is not usually true, however,

that if we increase some or all of the elements of a nonsingular matrix $A \quad ^{\times}$, then the elements of the inverse $A^{-} \quad ^{\times}$ will decrease. This useful property holds for the class of monotone matrices defined below.

The discussion in this section is not related to Gaussian elimination and LU factorisation, but it is of relevance in the iterative solution of systems of linear equations with monotone matrices which arise in the course of numerical approximation of boundary value problems for certain ordinary and partial differential equations.

Definition 3.5 *The nonsingular matrix* $A \quad ^{\times}$ *is said to be* **monotone** *if all the elements of the inverse* A^{-} *are nonnegative.*

Example 3.3 *Suppose that* a *and* d *are positive real numbers, and* b *and* c *are nonnegative real numbers such that* $ad > bc$. *Then,*

$$A = \begin{matrix} a & b \\ c & d \end{matrix}$$

is a monotone matrix. This is easily seen by considering the inverse of the matrix A,

$$A^{-} = \frac{1}{ad \quad bc} \begin{matrix} d & b \\ c & a \end{matrix} ,$$

and noting that all elements of A^{-} *are nonnegative.*

Next we introduce the concept of ordering in and $^{\times}$.

Definition 3.6 *For vectors and in we use the notation*

to mean that

$$x \quad y , \quad i = 1, 2, \ldots, n .$$

In the same way, for matrices A *and* B *in* $^{\times}$ *we write*

$$A \quad B$$

to mean that

$$a \quad b , \quad i, j = 1, 2, \ldots, n .$$

The sign is read 'succeeds or is equal to' or, simply, 'is greater than or equal to'.

Note that, given two arbitrary matrices A and B in $\times$, in general none of $A \quad B$, $A = B$ and $B \quad A$ will be true. Therefore the relation is a partial, rather than a total, ordering on $\times$; the same is true of the ordering on .

Theorem 3.5 *(i) Suppose that the nonsingular matrix* $A \quad \times$ *is monotone,* , , *and the vectors and in are the solutions of*

$$A = , \quad A = ,$$

respectively. If , then .

(ii) Suppose that A *and* B *are nonsingular matrices in* $\times$ *and that both are monotone. If* $A \quad B$*, then* $B^{-} \quad A^{-}$.

Proof (i) Since the elements of A^{-} are nonnegative and

$$= A^{-} (\quad),$$

the result follows from the fact that all elements of the vector $A^{-} (\quad)$ appearing on the right-hand side of this equality are nonnegative.

(ii) Since $A \quad B$ and all the elements of B^{-} are nonnegative, it follows that

$$B^{-} A \quad B^{-} B = I .$$

In the same way, since all the elements of A^{-} are nonnegative, it follows that

$$B^{-} = B^{-} A A^{-} \quad A^{-} ,$$

as required. □

The following theorem will be useful in Chapter 13.

Theorem 3.6 *Suppose that* $n \quad 3$ *and* $T \quad \times$ *is a tridiagonal matrix of the form (3.4) with the properties*

$$a < 0, \quad i = 2, 3, \ldots, n, \qquad c < 0, \quad i = 1, 2, \ldots, n \quad 1,$$

and

$$a + b + c \quad 0, \qquad i = 1, 2, \ldots, n,$$

where we have followed the convention that $a = 0$, $c = 0$*; then, the matrix* T *is monotone.*

Proof Let $k = 1, 2, \ldots, n$. Column k of the inverse T^{-1} is the solution of the linear system $Tx = e$, where e is column k of the identity matrix of size n, having a single nonzero element, 1, in row k. By applying the Thomas algorithm to this linear system, it is easy to deduce by induction from (3.7) that $l_j \neq 0$, $u_j \neq 0$ and $v_j \neq 0$ for all j; the argument is very similar to the proof of Theorem 3.4. It then follows from (3.8) and (3.9) that, in the notation of the Thomas algorithm, the vectors x and y have nonnegative elements. Hence column k of the inverse T^{-1} has nonnegative elements. Since the same is true for each $k = 1, 2, \ldots, n$, it follows that T is monotone. □

3.5 Notes

Symmetric systems of linear algebraic equations arise in the numerical solution of self-adjoint boundary value problems for differential equations with real-valued coefficients.

For further details on the Cholesky factorisation, the reader may consult any of the books listed in the Notes at the end of Chapter 2, particularly Chapter 10 of N.J. Higham, *Accuracy and Stability of Numerical Algorithms*, SIAM, Philadelphia, 1996.

Classical iterative methods for the solution of systems of linear equations with monotone matrices are discussed, for example, in

% &, *Matrix Iterative Analysis,* Prentice–Hall, Englewood Cliffs, NJ, 1962.

A more recent reference on iterative algorithms for linear systems is

' ( ) * ", *Iterative Solution Methods*, Cambridge University Press, Cambridge, 1996.

In particular, Chapter 6 of Axelson's book considers the relevance of monotone matrices in the context of iterative solution of systems of linear equations.

Theorem 3.6 is a slight variation on the following general result.

Theorem 3.7 *A sufficient condition for* $A^{\times}$ *to be a monotone matrix is that* A *is an* , *that is, (a)* $a \neq 0$ *for all* $i, j = 1, 2, \ldots, n$ *such that* $i = j$*, and (b) there exists a vector with positive elements such that all elements of* A *are positive.*

Exercises

3.1 Find the Cholesky factorisation of the matrix

$$A = \begin{matrix} 4 & 6 & 2 \\ 6 & 10 & 3 \\ 2 & 3 & 5 \end{matrix} .$$

3.2 Use the method of Cholesky factorisation to solve the system of equations

$$\begin{aligned} x \quad 2x \quad + 2x \quad &= \quad 4 , \\ 2x \quad + 5x \quad 3x \quad &= \quad 7 , \\ 2x \quad 3x \quad + 6x \quad &= \quad 10 . \end{aligned}$$

3.3 Let $n \quad 3$. The $n \quad n$ tridiagonal matrix T has the diagonal elements

$$T \quad = 2 , \quad i = 1, 2, \ldots , n ,$$

and the off-diagonal elements

$$T \quad = T \quad = \quad 1 , \quad i = 1, 2, \ldots , n \quad 1 .$$

In the factorisation $T = LU$, where $L \quad ^{\times}$ is unit lower triangular and $U \quad ^{\times}$ is upper triangular, show that

$$L \quad = \quad i/(i+1) , \qquad i = 1, 2, \ldots , n ,$$

and find expressions for the elements of U. What is the determinant of T?

3.4 Let $n \quad 3$ and $1 \quad k \quad n$. Define the vector with elements given by

$$v \quad = \quad \begin{matrix} i(n+1 \quad k) , & i = 1, \ldots , k , \\ k(n+1 \quad i) , & i = k+1, \ldots , n . \end{matrix}$$

Evaluate M , the inner product of the vector with column j of the matrix T defined in Exercise 3. (The inner product , ! of two vectors and in is defined as the real number .) Hence give expressions for the elements of the inverse matrix T^{-} , and verify that this inverse is symmetric. Find the -norm of the inverse, $T^{-} \quad _{\infty}$, and show that the condition number of T is

$$\kappa_{\infty}(T) = \frac{1}{2}(n+1) \quad , \quad n \text{ odd} .$$

What is the condition number $\kappa_{\infty}(T)$ when n is even?

3.5 Given that $n \geq 3$, in the notation of Theorem 3.4 suppose that

$$b_j = a_j + c_j, \quad j = 1, 2, \ldots, n,$$

and

$$c_j > 0, \quad j = 1, 2, \ldots, n-1,$$

with the convention that $a_1 = 0$ and $c_n = 0$. Show that the factorisation $T = LU$ exists without pivoting, and can be constructed by the Thomas algorithm. Give an example of a matrix T which satisfies these conditions, except that $c_k = 0$ for some $k \in 1, 2, \ldots, n-1$ and such that T is singular and cannot be written in the form $T = LU$ without pivoting.

3.6 Let $n \geq 3$ and suppose that the matrix $T \in \mathbb{R}^{n \times n}$ is tridiagonal. Show that there exists a permutation matrix $P \in \mathbb{R}^{n \times n}$ such that

$$PA = LU$$

where $L \in \mathbb{R}^{n \times n}$ is unit lower triangular with at most two nonzero elements in each row, and $U \in \mathbb{R}^{n \times n}$ is upper triangular with at most three nonzero elements in each row.

3.7 Suppose that the matrix B is Band(p,q), and that there exists a factorisation $B = LU$ without row interchanges. Show that L is Band(p,0) and U is Band(0,q).

3.8 Suppose that $n \geq 4$, that the matrix $A \in \mathbb{R}^{n \times n}$ is Band(3,3), and has the LU factorisation $A = LU$, so that $L \in \mathbb{R}^{n \times n}$ is Band(3,0) and $U \in \mathbb{R}^{n \times n}$ is Band(0,3). Suppose also that $a_{11} = 0$, $a_{i,i-1} = 0$ for $i = 1, 2, \ldots, n-2$. By considering u_{11} and l_{21}, or otherwise, show that in general the elements $l_{n-1,n-2}$ and $u_{n-1,n-1}$ are not zero.

4

Simultaneous nonlinear equations

4.1 Introduction

In Chapter 1 we discussed iterative methods for the solution of a single nonlinear equation of the form $f(x) = 0$ where f is a continuous real-valued function of a single real variable. In Chapters 2 and 3, on the other hand, we were concerned with direct (as opposed to iterative) methods for systems of linear equations. The purpose of the present chapter is to extend the techniques developed in Chapter 1 to systems of simultaneous nonlinear equations for functions of several real variables. We shall concentrate on two methods: the generalisation of simple iteration, usually referred to as simultaneous iteration, and Newton's method.

Given that $= (x , \ldots , x)$, as in Chapters 2 and 3 we denote by $_\infty$ the -norm of defined by

$$_\infty = \max x \quad .$$

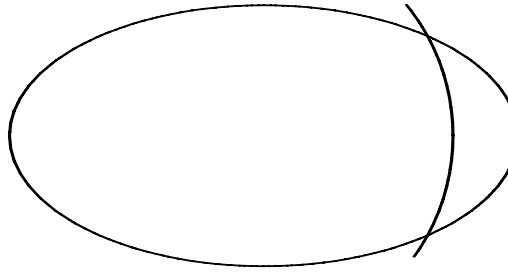
Throughout the chapter, will be thought of as a linear space equipped with the -norm; with only minor alterations all of our results can be restated in the p-norm with p [1,) on replacing $_\infty$ by throughout. We begin with some basic definitions which involve the concept of *open ball* defined in Section 2.7.

Let ; the open ball in (with respect to the -norm) of radius $\varepsilon > 0$ and centre is defined as the set

$$B ( ) = : _\infty < \varepsilon \quad .$$

A set D is said to be an **open set** in if for every D there exists $\varepsilon = \varepsilon( ) > 0$ such that $B ( ) D$ (see Figure 4.1). For example, any open ball in is an open set in . Given , any open set

$\varepsilon > 0$ such that $B(\quad) \quad D$. As $(\quad) \quad D$, no member of the sequence $(\quad)$ can enter $B(\quad)$. This, however, contradicts the fact that $(\quad)$ converges to . The contradiction implies that D. $\square$



Example 4.2 *Let us suppose that* A × *and* . *On letting* ( ) = A *we deduce that the problem of solving the system of simultaneous linear equations considered in Chapters 2 and 3 can be restated in the form: find such that* ( ) = **0**.

Let us assume that we have transformed the equation ( ) = **0** into an equivalent form ( ) = , where : is a continuous function, defined on the closed subset D , such that (D) D. For example, one can choose ( ) = α ( ), with α a suitable parameter. By 'equivalent' we mean that D satisfies ( ) = **0** if, and only if, ( ) = . Any D such that ( ) = is called a **fixed point** of the function in D. Thus the problem of finding a solution D to the equation ( ) = **0** has been converted into one of finding a fixed point in D of the function . We embark on the latter task by considering the natural extension to of the simple iteration discussed in Section 1.2 for the solution of the scalar nonlinear equation g(x) = x.

Definition 4.1 *Suppose that* : *is a function, defined and continuous on a closed subset* D *of* , *such that* (D) D. *Given that* D, *the recursion defined by*

$$= ( ), \qquad k = 0, 1, 2, \ldots , \tag{4.3}$$

is called a **simultaneous iteration**. *For* n = 1 *the recursion (4.3) is just the simple iteration considered in (1.3).*

Note that here we use the superscript k as the sequence index; following the convention adopted in Chapters 2 and 3, we reserve subscripts for labelling the entries of vectors. Thus x is entry i of the vector , the kth member of the sequence ( ). The motivation behind the definition of the simultaneous iteration (4.3) is, of course, our hope that, under suitable conditions on and D, the sequence ( ) will converge to a fixed point of .

Two remarks are in order at this point. First, it is easy to show that if a sequence of vectors ( ) converges in to in the norm ∞, then it also converges to this same limit in the norm for any p [1,). To see this, note that

$$\infty \qquad n \qquad \infty \qquad , \tag{4.4}$$

for 1 p < , and take = to deduce that, as k , convergence in the -norm implies convergence in the p-norm for any p [1,), and *vice versa*

Any function that satisfies a Lipschitz condition on a set D is continuous on D. For let D and $\varepsilon > 0$; then, on defining $\delta = \varepsilon/\mathsf{L}$, we deduce from (4.5) that if $_\infty < \delta$ for some D, then

$$( ) \quad ( ) _\infty \quad \mathsf{L} \quad _\infty < \varepsilon .$$

It follows from (4.4) that if satisfies a Lipschitz condition on D in the -norm then it also does so in the p-norm for any p $[1, \ )$, and *vice versa*. However, in general, the size of the constant L may depend on the choice of norm. Specifically, if is a contraction on a set D in the -norm (*i.e.*, (4.5) holds with $\mathsf{L} < 1$), then *need not* be a contraction in the p-norm, unless $\mathsf{L} < \mathsf{n}^{-}$. (See Exercise 1.) Conversely, if is a contraction on D in the p-norm for some p $[1, \ )$, it does *not* follow that is a contraction on D in the -norm.

For example, suppose that : is the linear function defined by $( ) = \mathsf{A}$, where A is the 2 2 matrix

$$\mathsf{A} = \begin{matrix} 3/4 & 1/3 \\ 0 & 3/4 \end{matrix} .$$

This function satisfies a Lipschitz condition on in for any p $[1, \ ]$, and if L is a Lipschitz constant for in the p-norm, then L A , in the subordinate matrix norm. It is easy to see that A $= \mathsf{A}_\infty = 13/12$, and a small calculation gives A $= 0.935$ to three decimal digits. Hence the function is a contraction in the 2-norm, but not in the 1- or -norm.

Our next result is a direct generalisation of Theorem 1.3 formulated in Chapter 1.

Theorem 4.1 (**Contraction Mapping Theorem**) *Suppose that* D *is a closed subset of* , : *is defined on* D, *and* (D) D. *Suppose further that is a contraction on* D *in the -norm. Then, has a unique fixed point in* D, *and the sequence* ( ) *defined by (4.3) converges to for any starting value* D.

Proof Assuming that has a fixed point in D, the *uniqueness* of the fixed point is easy to show: for suppose that is also a fixed point of in D. Then, by (4.5),

$$_\infty = \quad ( ) \quad ( ) _\infty \quad \mathsf{L} \quad _\infty ,$$

i.e., $(1 \quad \mathsf{L}) \quad _{\infty} \quad 0$. Since $\mathsf{L} \quad (0,1)$, and $_{\infty}$ is a norm, it follows that $= \mathbf{0}$, and hence $= \ $. Consequently, if $\ $ has a fixed point in D, then this is the unique fixed point of $\ $ in D.

Now, still *assuming* that $\ $ possesses a fixed point $\ \mathsf{D}$, we shall show that the sequence $(\ )$ defined by (4.3) converges to $\ $ for any starting value $\ \mathsf{D}$

for all $k \quad 2$. We then deduce by induction that

$$\quad \infty \quad L \quad \infty , \qquad k \quad 1 . \tag{4.8}$$

Suppose that m and k are positive integers and $m \quad k+1$. Then, by repeated application of the triangle inequality in the -norm and using (4.8), we have that

$$\begin{aligned} \infty &= (\quad) + \quad + (\quad) \infty \\ & \quad \infty + \quad + \quad \infty \\ & \quad (L \quad + \quad + L \quad) \quad \infty \\ &= L \ (L \quad + \quad + 1) \quad \infty \\ & \quad L \ \frac{1}{1 \quad L} \quad \infty , \end{aligned} \tag{4.9}$$

where, in the transition to the last line, we made use of the fact that the geometric series $1 + L + L + \quad$, with $L \quad (0,1)$, sums to $1/(1 \quad L)$.

As $\lim \to \infty L = 0$, it follows from (4.9) that ( ) is a Cauchy sequence in ; that is, for each > 0 there exists $k = k \ ()$ (defined by (4.6) above) such that

$$\infty < \qquad m,k \quad k = k \ (). \tag{4.10}$$

Any Cauchy sequence in is convergent in ; consequently, there exists such that $= \lim \to \infty$. Further, since satisfies a Lipschitz condition on D, the discussion in the paragraph following Definition 4.2 shows that is continuous on D. Hence, by Lemma 4.2,

$$= \lim_{\to \infty} \qquad = \lim_{\to \infty} (\quad) = \quad \lim_{\to \infty} \qquad = ( ),$$

which proves that is a fixed point of .

It remains to show that D. This follows from Lemma 4.1 since ( ) D, $= \lim \to \infty$ and D is closed. □

As a byproduct of the proof, we deduce from (4.7) that, given a positive tolerance , one can compute an approximation to the unknown solution using (4.3) in no more than $k = k \ ()$ iterations so that the approximation error , measured in the -norm, is less than ; the integer $k \ ()$ is defined by (4.6).

The next theorem relates the constant L from the Lipschitz condition (4.5) to the partial derivatives of , giving a more practically useful sufficient condition for convergence.

Definition 4.3 *Let* $= (g, \ldots, g) :$ *be a function defined and continuous in an (open) neighbourhood* $N( )$ *of* *. Suppose further that the first partial derivatives* $\frac{}{}$, $j = 1, \ldots, n$, *of* g *exist at* *for* $i = 1, \ldots, n$. *The* **Jacobian matrix** $J( )$ *of* *at* *is the* $n \quad n$ *matrix with elements*

$$J( ) = \frac{\partial g}{}$$

for $i = 1, \ldots, n$. Now $x \quad y \quad \infty$ for all $j \quad 1, \ldots, n$, and so (4.12) gives

$$g(\quad) \quad g(\quad) \quad \infty \quad \frac{\partial g}{\partial x}(\eta \quad + (1 \quad \eta) \quad)$$

$$\infty \quad J \ (\eta \quad + (1 \quad \eta) \quad) \quad \infty ,$$

for all $i = 1, \ldots, n$. Consequently, for any $\quad , \quad \bar{B} \ (\ )$,

$$(\quad) \quad (\quad) \ \infty \quad \max_{\square} \ J \ (t \quad + (1 \quad t) \quad) \ \infty \quad \infty$$

$$-(1 + K) \quad \infty , \tag{4.13}$$

due to (4.11), given that $t \quad + (1 \quad t) \quad \bar{B} \ (\ )$ for all $t \quad [0, 1]$. It follows that satisfies a Lipschitz condition (4.5), in the -norm, on the closed ball $\bar{B} \ (\ )$ with $L = -(1 + K) < 1$. Furthermore, on selecting $\quad =$ in (4.13) we get that

$$(\quad) \quad \infty = \quad (\quad) \quad (\quad) \ \infty < \quad \infty \quad \varepsilon$$

for all $\quad \bar{B} \ (\ )$. Hence, $(\bar{B} \ (\ )) \quad \bar{B} \ (\ )$. The convergence of the iteration (4.3) to , for an arbitrary starting value $\quad \bar{B} \ (\ )$, now follows from Theorem 4.1. □

We close this section with an example which illustrates the application of the method of simultaneous iteration to the solution of a system of nonlinear equations.

Example 4.4 *Let us consider, as in Example 4.1, the system of two simultaneous nonlinear equations in the unknowns* x *and* x *, defined by*

$$x \quad + x \quad 1 \quad = \quad 0 ,$$

$$5x \quad + 21x \quad 9 \quad = \quad 0 .$$

Here $\quad = (x \ , x \ )$ *and* $\quad = (f \ , f \ )$ *with*

$$f \ (x \ , x \ ) \quad = \quad x \quad + x \quad 1 ,$$

$$f \ (x \ , x \ ) \quad = \quad 5x \quad + 21x \quad 9 .$$

Let us suppose that we need to find the solution of the system $(\ ) = \mathbf{0}$ *in the first quadrant of the* $(x \ , x \ )$*-coordinate system.*

Of course, the example is a little artificial, since we already know from Example 4.1 that $\quad = (\ \overline{3}/2, 1/2)$ is the required solution. In what follows, however, we proceed as if we knew nothing about the location

of . Our aim here is to illustrate the construction of the function from and the verification of the hypotheses of Theorem 4.1.

Let us rewrite the two equations as

$$x = 1 \quad x \qquad , \qquad x = \frac{1}{\overline{21}} \; 9 \quad 5x \qquad ,$$

and define $g(x, x)$ and $g(x, x)$ as the right-hand sides of these, respectively. We consider the simultaneous iteration

for all and in D. Therefore, also,

$$( ) \quad ( ) \infty \quad L \quad \infty$$

with

$$L = \max_{\square} \; J \; ( ) \; \infty < 1 . \tag{4.15}$$

With our choice of D, (4.15) holds with $L = \max\ 0.75, 0.55\ = 0.75 < 1$. Furthermore, it is easy to check that (D) D. Thus we deduce from Theorem 4.1 that has a unique fixed point in D – we call this fixed point , for the sake of consistency with the notation in Example 4.1; moreover, the sequence ( ) defined by (4.14) converges to .

After all these preparations you are now probably curious to see what the successive iterates look like: Table 4.1 gives a flavour of the behaviour of the sequence ( ), with the starting value chosen as $= (0.5, 0.3)$. You can see that after 15 iterations the first 5 decimal digits have settled to their correct values.

4.3 Relaxation and Newton's method

We now go on to apply the ideas developed in the previous section to the construction of an iteration which converges to a solution of the equation $( ) = \mathbf{0}$, where : . One way of constructing such a sequence is by relaxation.

Definition 4.4 *The recursion*

$$= \quad \lambda \; (\quad), \qquad k = 0, 1, 2, \ldots , \tag{4.16}$$

where is given and where $\lambda = 0$ *is a constant, is called* **simultaneous relaxation**.

Suppose that the sequence ( ) converges to a limit and is continuous in a neighbourhood of ; then, on passing to the limit k in (4.16), we deduce that is a solution of the equation $( ) = \mathbf{0}$.

Simultaneous relaxation is evidently a simultaneous iteration defined by taking $( ) = \quad \lambda \; ( )$.

1 ! 6 , 1 G 2 +
! 4
G 6 B 4?D 1 B 8 D 1 " 5 5 Q

Table 4.1. *The first 15 iterates in the sequence* $= (x \quad , x \quad)$ *defined by (4.14), with starting value* $(0.5, 0.3)$ *. The exact solution is* $= (\ \sqrt{3}/2, 1/2) \ = (0.866025403784439, 0.500000000000000)$ *to 15 decimal digits.*

	$^{(\ )}_{1}$	$_{2}{}^{*}$ +
	(&	
	$(&$&$ $)%%)$)	%)'$ $% $&$(%
	)$'& (&% ' $	'% && '(($
&	))'))()(( )(& '$ (	
'	'$($ $ ' $	'$ '($ '%%
(	) '%')$ %&(	(%) &)$ '& &%%
%	%) %))'	'$) &() $(%
)	%)) &'$%&% $(	(% ' %) $ (%
	%($) $%' (%('	'$$' (('%& &%'%
$	%%& $	(& '&')$(&'
	%("$ % (	'$$))(($ ( )%
	%% $% &(% &$	($ '(%')
	%($) $ &)	'$$$) (%(%
&	%% ' & (%'(	(%) %(&
'	%% $%&%'$	'$$$$& ($) '))
(	%% $') %)'	((%& ') %

Theorem 4.3 *Suppose that* $(\ ) = \mathbf{0}$*, and that all the first partial derivatives of* $= (f \ , \ldots , f \ )$ *are defined and continuous in some (open) neighbourhood of* *, and satisfy a condition of strict diagonal dominance at* *; i.e.,*

$$\frac{\partial f}{\partial x}(\ ) > \quad \frac{\partial f}{\partial x}(\ ) , \quad i = 1, 2, \ldots , n . \tag{4.17}$$

□

Then, there exist $\varepsilon > 0$ *and a positive constant* λ *such that the relaxation iteration (4.16) converges to* *for any* *in the closed ball* $\bar{B} \ (\ )$ *of radius* ε*, centre* *.*

Proof The elements of the Jacobian matrix $J \ (\ ) = (\gamma \quad)^{\times}$ of the function $(\ ) = \quad \lambda \ (\ )$ at $=$ are

$$\gamma \ (\ ) = 1 \ \ \lambda \frac{\partial f}{\partial x}(\ ), \quad \gamma \ (\ ) = \ \lambda \frac{\partial f}{\partial x}(\ ), \quad j = i , \quad i, j \quad 1, \ldots , n \ .$$

We now define

$$m = \max \frac{\partial f}{\partial x}(\ )$$

and then choose $\lambda = 1/m$. Under hypothesis (4.17), $m > 0$ and therefore $\lambda > 0$. This choice of λ ensures that all the diagonal elements γ ( ), $i = 1, \ldots, n$, of J ( ) are nonnegative. Moreover, for any i $1, \ldots, n$,

γ

the system of equations $(\ ) = \mathbf{0}$. *It is implicitly assumed that the matrix* $\mathsf{J}(\ )$ *exists and is nonsingular for each* $\mathsf{k} = 0, 1, 2, \ldots$.

The next theorem is concerned with the convergence of Newton's method. As in the scalar case, for a starting value that is sufficiently close to the solution of $(\ ) = \mathbf{0}$, Newton's method converges quadratically. The precise definition of quadratic convergence is given below: it resembles Definition 1.7 of Chapter 1.

Definition 4.6 *Suppose that* $(\ )$ *is a convergent sequence in* *and* $= \lim_{\to\infty}$. *We say that* $(\ )$ *converges to* **with at least order** $\mathsf{q} > 1$, *if there exist a sequence* (ε) *of positive real numbers converging to* 0, *and* $\mu > 0$, *such that*

$$\varepsilon, \quad \mathsf{k} = 0, 1, 2, \ldots, \qquad \text{and} \qquad \lim_{\to\infty} \frac{\varepsilon}{\varepsilon} = \mu. \tag{4.19}$$

If (4.19) holds with $\varepsilon =$, $\mathsf{k} = 0, 1, 2, \ldots$, *then the sequence* $(\ )$ *is said to converge to* **with order** q. *In particular, if* $\mathsf{q} = 2$, *then we say that the sequence* $(\ )$ *converges to* **quadratically**.

Again, due to (4.4), if a sequence $(\ )$ converges quadratically in the -norm, then it also does so in the p-norm for any p $[1, \ )$, though the constant μ may be different.

Theorem 4.4 *Suppose that* $(\ ) = \mathbf{0}$, *that in some (open) neighbourhood* $\mathsf{N}(\ )$ *of* , *where* *is defined and continuous, all the second-order partial derivatives of* *are defined and continuous, and that the Jacobian matrix* $\mathsf{J}(\ )$ *of* *at the point* *is nonsingular. Then, the sequence* $(\ )$ *defined by Newton's method (4.18) converges to the solution* *provided that* *is sufficiently close to* ; *the convergence of the sequence* $(\ )$ *to* *is at least quadratic.*

Proof Let us begin by writing Newton's method as a simultaneous iteration $= (\ )$, $\mathsf{k} = 0, 1, 2, \ldots$, as in (4.3), with given and

$$(\ ) = \quad [\mathsf{J}(\ )]^{-} \quad (\ ).$$

The idea of the proof is to verify that the function satisfies all the conditions of Theorem 4.2 in a certain closed ball centred at , the fixed point of , and thus deduce that the sequence $(\ )$ converges to .

As the function $\det\mathsf{J}(\ )$ is continuous in $\mathsf{N}(\ )$ and $\det\mathsf{J}(\ ) = 0$, there exists $\varepsilon > 0$ such that $\det\mathsf{J}(\ ) = 0$ for all $\bar{\mathsf{B}}(\ )$ $\mathsf{N}(\ )$.

Further, as the entries of $[\mathsf{J}(\ )]^{-}$ depend continuously on the entries of $\mathsf{J}(\ )$ and since the entries of $\mathsf{J}(\ )$ are continuous functions of $$ in $\mathsf{N}(\ )$, we deduce that $[\mathsf{J}(\ )]^{-}(\ )$ is a continuous function on $\bar{\mathsf{B}}(\ )$; therefore,

$$(\ ) = [\mathsf{J}(\ )]^{-}(\ )$$

is also a continuous function on $\bar{\mathsf{B}}(\ )$. For later reference, we note that $[\mathsf{J}(\ )]^{-} \infty$

and so

$$\| \quad -n \; A \; C \| \quad .$$

On writing $M = -n \; A \; C$, we then deduce by induction that

$$\| \quad \frac{1}{M} \quad M \quad \| \quad , \qquad k = 0, 1, 2, \ldots .$$

Suppose that $\bar{B}$ ( ) where $\varepsilon \quad -\min \; 1, 1/M$. Then,

$$M \quad \| \quad \frac{1}{2}, \qquad k = 0, 1, 2, \ldots ,$$

and hence

$$\| \quad \frac{1}{M} \quad \frac{1}{2}$$

This implies that convergence is at least quadratic (on choosing $\varepsilon = M^{-} \; 2^{-}$ and $q = 2$ in Definition 4.6). □

Newton's method is defined in (4.18) by using the inverse of the Jacobian matrix. As we saw in Chapter 2 it is more efficient to avoid inverting a matrix, if possible. In practice the method is therefore implemented by writing (4.18) in the form

$$J \; (\quad)[\quad] = \quad (\quad). \tag{4.23}$$

Given the vector , we calculate ( ) and the Jacobian matrix J ( ) $^{\times}$, and then solve the system of linear equations (4.23) by Gaussian elimination; this gives the increment vector , which is added to to obtain the new iterate .

Example 4.5 *We close this section with an example which illustrates the application of Newton's method. Consider the simultaneous nonlinear equations*

$$\begin{aligned} f \; (x, y, z) \quad & x \; + y \; + z \quad 1 = 0 , \\ f \; (x, y, z) \quad & 2x \; + y \quad 4z = 0 , \\ f \; (x, y, z) \quad & 3x \quad 4y + z \; = 0 . \end{aligned}$$

Letting $= (f \; , f \; , f \;)$ *and* $= (x, y, z)$ *, the aim of the exercise is to determine the solution to the equation* ( ) $= \mathbf{0}$ *contained in the first octant* $(x, y, z) \quad : x > 0 , \; y > 0 , \; z > 0$ *in* .

P are, very roughly, y 0.5 and z 0.5, it is reasonable to choose as starting value for the Newton iteration the point

Chapter 1, and also an infinite number of complex solutions. It is easy to see from the periodic character of $e^{\#}$ that the equation has a solution near $w = (2m + -)\imath\pi$, $\imath = \overline{1}$, for integer values of m; a better estimate is given in Exercise 9. It is a good deal more difficult to prove that there are no other solutions.

The behaviour of Newton's method for this problem may be illustrated by showing a picture of the complex plane, with the sets S depicted in different colours. In our example we cannot, of course, show more than a small number of the solutions, and cannot use an infinite number of colours. We have therefore coloured the sets with six colours cyclically, so that, for example, the sets S , S , S , ... have the same colour. The background colour, white, represents the set S of points from which the iteration converges to the real negative root. It includes most of the negative half-plane. Successive pictures in the series from Figure 4.5 to Figure 4.9 show a magnified view of a small region of the previous picture, the region being outlined in black. In Figure 4.4 the black crosses mark the positions of solutions of $f(z) = 0$. The pictures show in a striking way the fractal behaviour of the boundary of a set. Figure 4.9 is very similar to Figure 4.5; the former is a magnified view of a small part of Figure 4.5, with a magnification of about 50000 in each direction. The same sort of behaviour is repeated when the picture is magnified indefinitely.

4.5 Notes

For an introduction to the topology of , including the definitions of open set, closed set, continuity, convergence and Cauchy sequence, the reader is referred to any standard textbook on the subject; see, *e.g.*,

% , *Principles of Mathematical Analysis*, Third Edition, International Series in Pure and Applied Mathematics, McGraw–Hill, New York, Auckland, Düsseldorf, 1976,

&) , *Introduction to Mathematical Analysis*, Addison–Wesley, Reading, MA, 1996.

Our first remark concerns the Contraction Mapping Theorem, Theorem 4.1, which is a direct generalisation of Theorem 1.3 from Chapter 1. Comparing the proofs of Theorems 1.3 and 4.1, we see that the proof of Theorem 1.3 is much simpler. This is not accidental: in the case of a single equation $x = g(x)$, involving a real-valued function g of a single real variable x, the existence of a fixed point follows directly from

Theorem 1.2, Brouwer's Fixed Point Theorem on a bounded closed interval of the real line. On the other hand, for the simultaneous system of equations = ( ) in considered in Theorem 4.1 we had to invoke the completeness of (i.e., the property that every Cauchy sequence in is a convergent sequence) to show the existence of a fixed point. An alternative, shorter proof of Theorem 4.1 could have been devised by applying Brouwer's Fixed Point Theorem in .

Theorem 4.5 (**Brouwer's Fixed Point Theorem**) *Let us assume that* D *is a nonempty, closed, bounded and convex subset of . Suppose further that* : *is a continuous function defined on* D *such that* (D) D. *Then, there exists* D *such that* ( ) = .

A set D is said to be convex if, whenever and belong to D, also

$$\theta + (1 \quad \theta) \quad D \quad \theta \quad [0,1].$$

For example, any nonempty interval of the real line = is a convex set, as is a nonempty (open or closed) ball in , n 2. Unfortunately, when n 2 the proof of Theorem 4.5 is nontrivial and is well beyond the scope of this book.

Benoit Mandelbrot (1924–) has been largely responsible for the present interest in fractal geometry and its connections with iterative methods. Mandelbrot highlighted in his book

! + , *Fractals: Form, Chance, and Dimension*, W.H. Freeman, San Francisco, 1977,

and, more fully, in

! + " *The Fractal Geometry of Nature*, W.H. Freeman, New York, 1983,

the omnipresence of fractals both in mathematics and elsewhere in nature. In relation with the subject of this chapter, we note that the **Mandelbrot set** is a connected set of points in the complex plane defined as follows. Choose a point z in the complex plane, and consider the iteration z = z +z , n = 0, 1, 2, ... If the sequence z , z , z , ... remains within a distance of 2 from the origin for ever, then the point z

1 9 1 1 4> , , K
4 ') # 5 + , : "
' , #@@&4

is said to be in the Mandelbrot set. If the sequence diverges from the origin, then the point z is not in the set.

A standard reference for theoretical results concerning the convergence of Newton's method in complete normed linear spaces is

, $) - " *Functional Analysis*, Second edition, Pergamon Press, Oxford, New York, 1982.

A further significant book in the area of iterative solution of systems of nonlinear equations is the text by

+ ' % " *Iterative Solution of Nonlinear Equations in Several Variables*, Reprint of the 1970 original, Classics in Applied Mathematics, 30, SIAM, Philadelphia, 2000.

It gives a comprehensive treatment of the numerical solution of n nonlinear equations in n unknowns, covering asymptotic convergence results for a number of algorithms, including Newton's method, as well as existence theorems for solutions of nonlinear equations based on the use of topological degree theory and Brouwer's Fixed Point Theorem.

Exercises

4.1 Suppose that the function is a contraction in the -norm, as in (4.5). Use the fact that

$$( ) \quad ( ) \quad n \quad ( ) \quad ( ) \ \infty$$

to show that is a contraction in the p-norm if $L < n^{-}$.

4.2 Show that the simultaneous equations $(x , x) = \mathbf{0}$, where $= (f , f)$, with

$$f (x , x) = x + x \quad 25 , \qquad f (x , x) = x \quad 7x \quad 25 ,$$

have two solutions, one of which is $x = 4$, $x = 3$, and find the other. Show that the function does not satisfy the conditions of Theorem 4.3 at either of these solutions, but that if the sign of f is changed the conditions are satisfied at one solution, and that if is replaced by $= (f \quad f , f)$, then the conditions are satisfied at the other. In each case, give a value of the relaxation parameter λ which will lead to convergence.

4.3 The complex-valued function $z \quad g(z)$ of the complex variable z is holomorphic in a convex region Ω containing the point ζ, at which $g(\zeta) = \zeta$. By applying the Mean Value Theorem (Theorem A.3) to the function ϕ of the real variable t defined by $\phi(t) = g((1 \quad t)u + tv)$ show that if u and v lie in Ω, then there is a complex number η in Ω such that

$$g(u) \quad g(v) = (u \quad v)g'(\eta)\,.$$

Hence show that if $g'(\zeta) < 1$, then the complex iteration defined by $z \quad = g(z \ )$, $k = 0, 1, 2, \ldots$, converges to ζ provided that z is sufficiently close to ζ.

4.4 Suppose that in Exercise 3 the real and imaginary parts of g are u and v, so that $g(x + \imath y) = u(x, y) + \imath v(x, y)$, $\imath = \sqrt{\quad 1}$. Show that the iteration defined by $\quad = \ '(\quad)$, $k = 0, 1, 2, \ldots$, where $'(\ ) = (u(x \ , x \ ), v(x \ , x \ ))$, generates the real and imaginary parts of the sequence defined in Exercise 3. Compare the condition for convergence given in that exercise with the sufficient condition given by Theorem 4.2.

4.5 Verify that the iteration $\quad = (\quad)$, $k = 0, 1, 2, \ldots$, where $= (g \ , g \ )$ and g and g are functions of two variables defined by

$$g \ (x \ , x \ ) = -(x \quad x \ + 3)\,, \quad g \ (x \ , x \ ) = -(2x \ x \ + 1)\,,$$

has the fixed point $= (1, 1)$. Show that the function does not satisfy the conditions of Theorem 4.3. By applying the results of Exercises 3 and 4 to the complex function g defined by

$$g(z) = -(z \ + 3 + \imath)\,, \qquad z \qquad , \quad \imath = \sqrt{\quad 1}\,,$$

show that the iteration, nevertheless, converges.

4.6 Suppose that all the second-order partial derivatives of the function : are defined and continuous in a neighbourhood of the point in , at which $(\ ) = \mathbf{0}$. Assume also that the Jacobian matrix, $J(\ )$, of is nonsingular at $=$, and denote its inverse by $K(\ )$ at all for which it exists. Defining the Newton iteration by $\quad = (\quad)$, $k = 0, 1, 2, \ldots$, with given, where $(\ ) = \quad K(\ )(\ )$, show that the (i, j)-entry

4.9 Suppose that the equation $e^{\zeta} = \zeta + 2$, ζ , has a solution

$$\zeta = (2m + -)\imath\pi + \ln[(2m + -)\pi] + \eta\,,$$

where m is a positive integer and $\imath = \sqrt{1}$. Show that

$$\eta = \ln[1 \quad \imath(\ln(2m + -)\pi + \eta + 2)/(2m + -\pi)]$$

and deduce that $\eta = \quad (\ln m/m)$ for large m.
(Note that $\ln(1 + \imath t) < t$ for all $t \qquad 0$.)

▫

' 5 '$& '$% &
!

' $ 5 '$&%'('$&%%( )&
)(!

5

Eigenvalues and eigenvectors of a symmetric matrix

5.1 Introduction

Eigenvalue problems for symmetric matrices arise in all areas of applied science. The terminology *eigenvalue* comes from the German word *Eigenwert* which means proper or characteristic value. The concept of eigenvalue first appeared in an article on systems of linear differential equations by the French mathematician d'Alembert in the course of studying the motion of a string with masses attached to it at various points.

Let us recall from Chapter 2 the definition of eigenvalue and eigenvector.

Definition 5.1 *Suppose that* A × . *A complex number* λ *for which the set of linear equations*

$$\mathsf{A} \quad = \lambda \tag{5.1}$$

has a nontrivial solution = **0** *is called an* **eigenvalue** *of* A*; the associated solution is called an* **eigenvector** *of* A *(corresponding to* λ*).*

1 K N% B#& 5 #&#& 9 , C @ H, #&.*
9 , D 6 , 1 , , 1
K 6 1 1 , 4
N% , , E 9 , ,6,
#&>#C#&& L ,6, 6 K / , .
5 4 / N% 1 E, , 6 ,
1 6 6 , 6 , $ 6
6 1 5 4

In order to motivate the discussion that will follow, we begin with two familiar elementary examples.

In considering the rotation of a rigid body Ω, the *inertia matrix* is the 3×3 symmetric matrix

$$J = \begin{matrix} I & I & I \\ I & I & I \\ I & I & I \end{matrix}$$

whose diagonal elements are the moments of inertia about the axes,

$$I = \int (y + z)\, d\Omega, \quad I = \int (z + x)\, d\Omega, \quad I = \int (x + y)\, d\Omega,$$

and whose off-diagonal elements are defined by the corresponding products of inertia

$$\begin{aligned} I &= I = \int xy\, d\Omega, \\ I &= I = \int yz\, d\Omega, \\ I &= I = \int zx\, d\Omega. \end{aligned}$$

Then, the eigenvectors of the inertia matrix are the directions of the *principal axes of inertia* of the body, about which free steady rotation is possible, and the eigenvalues are the *principal moments of inertia* about these axes.

A second example, which involves matrices of any order, arises in the solution of systems of linear ordinary differential equations of the form

$$\frac{d}{dt} = A \quad ,$$

where is a vector of n elements, each of which is a function of the independent variable t, and A is an $n \times n$ matrix whose elements are constants. If A were a diagonal matrix, with diagonal elements $a = \lambda$, $i = 1, 2, \ldots, n$, the solution of this system would be straightforward, as each of the equations could be solved separately, giving

$$x(t) = x(0) \exp(\lambda t), \qquad i = 1, 2, \ldots, n.$$

When A is not a diagonal matrix, suppose that we can find a nonsingular matrix M such that

$$M^{-} AM = D,$$

where D is a diagonal matrix. Then, on letting

$$= M^{-} \quad ,$$

we easily see that

$$\frac{d}{dt} = M^{-} AM \quad = D \quad .$$

The solution of this system of differential equations is straightforward, as we have just seen, and we then find that

$$x \; = (My) \; = \qquad M \;\; y \; (0) \exp(\lambda \; t) \, ,$$

where $\lambda \; = d \;$ is one of the diagonal elements of D. The numbers λ , $j = 1, 2, \ldots, n$, are the eigenvalues of the matrix $A \qquad ^{\times}$, and the columns of M are the eigenvectors of A, so the solution of this system of differential equations requires the calculation of the eigenvalues and eigenvectors of the matrix A.

In systems of differential equations of this kind the matrix A is not necessarily symmetric. In that case, the problem is more difficult; if the eigenvalues of A are not distinct there may not exist a complete set of linearly independent eigenvectors, and then the matrix M will not exist.

In this chapter, we shall develop numerical algorithms for the solution of the algebraic eigenvalue problem (5.1), assuming throughout that A $^{\times}$ is a symmetric matrix. As has been noted above, the analogous problem for a nonsymmetric matrix is more involved, and will not be considered here.

Throughout this chapter, the set of all real-valued symmetric matrices of order n will be denoted by $^{\times}$; thus, given a matrix $A = (a \;)$,

$$A \qquad ^{\times} \qquad \# \qquad A \qquad ^{\times} \qquad \& \quad a \;\; = a \;\; , \quad i, j = 1, 2, \ldots, n \, .$$

We begin with a reminder of some fundamental properties.

1 1 8

A #

#

8 5 1 , 6 1 2 A # 6 B 6

D 5 , B# D^{T}4

2 1 1 , 1 K4O4 " N

. $ H81 5 6

! " #@..4

Theorem 5.1 *Suppose that* A × *; then, the following statements are valid.*

(i)

5.2 The characteristic polynomial

Given that $A \quad ^{\times}$ and $n \quad 4$, it is quite easy to write down the characteristic polynomial $\det(A \quad \lambda I)$ by expanding the determinant, and then find the roots of this polynomial of degree n in order to determine the eigenvalues of A. If $n > 4$ there is no general closed formula for the roots of a polynomial in terms of its coefficients, and therefore we have to resort to a numerical technique. A further difficulty is that the roots may be very sensitive to small changes in the coefficients of the polynomial, and we find that the effect of rounding errors in the construction of the characteristic polynomial is usually catastrophic.

Example 5.1 *Consider, for example, the diagonal matrix of order* 16 *whose diagonal elements are* $j + -$, $j = 1, 2, \ldots, 16$*; the eigenvalues are, of course, just the diagonal elements. Constructing the characteristic polynomial, working with 10 significant digits throughout, gives the result*

$$\lambda \quad 141.3333333\lambda \quad + 9193.333333\lambda \qquad .$$

Using a standard numerical algorithm (such as Newton's method) for computing the roots of the polynomial and working with 10 significant digits gives the smallest root as 1.333333331, which is nearly correct to 10 significant digits. The three largest roots, however, are computed as, approximately, $15.5 \quad 1.3\imath$ and 16.7, which are very different from their true values $14.\dot{3}$, $15.\dot{3}$, $16.\dot{3}$, respectively, even though the matrix in this example is of quite modest size, and the eigenvalues are well spaced. Thus we conclude from this example that the numerical method which constructs the characteristic polynomial and finds its roots is completely unsatisfactory for general use, except for matrices of very small size.

The fact that in general the roots of the characteristic polynomial cannot be given in closed form shows that any method must proceed by successive approximation. Although one cannot expect to produce the required eigenvalues exactly in a finite number of steps, we shall see that there exist rapidly convergent iterative methods for computing the eigenvalues and eigenvectors numerically.

5.3 Jacobi's method

This method uses a succession of orthogonal transformations to produce a sequence of matrices which approaches a diagonal matrix in the limit.

Each step in the process involves a matrix representing a plane rotation. We begin with a simple example.

Example 5.2 (The plane rotation matrix in) *Let us suppose that* $\phi \quad [\ \pi, \pi]$ *and consider the matrix* $R(\phi) \quad ^{\times}$ *defined by*

$$R(\phi) = \begin{matrix} \cos\phi & \sin\phi \\ \sin\phi & \cos\phi \end{matrix} .$$

For a vector , $R(\phi)$ is the plane rotation of around the origin by an angle ϕ (in the clockwise direction when $\phi > 0$ and in the anticlockwise direction when $\phi < 0$).

We note in passing that since $\cos(\ \phi) = \cos\phi$, $\sin(\ \phi) = \ \sin\phi$ *and* $\cos\ \phi + \sin\ \phi = 1$, *we have that*

$$(R(\phi))\ \ = R(\ \ \phi) \qquad \textit{and} \qquad R(\phi)\, R(\ \ \phi) = I .$$

Hence $R(\phi)$ *is an orthogonal matrix; i.e.,*

$$R(\phi)R(\phi)\ \ = R(\phi)\ \ R(\phi) = I ,$$

where I *is the* $2 \quad 2$ *identity matrix.*

The next definition extends the notion of plane rotation matrix to .

Definition 5.2 (The plane rotation matrix in) *Suppose that* $n \quad 2,\ 1 \quad p < q \quad n$ *and* $\phi \quad [\ \pi, \pi]$. *We consider the matrix* $R \quad (\phi) \quad ^{\times}$ *whose elements are the same as those of the identity matrix* $I \quad ^{\times}$ *, except for the four elements*

$$r \quad = c, \qquad r \quad = s,$$
$$r \quad = \ \ s, \qquad r \quad = c,$$

where $c = \cos\phi$, $s = \sin\phi$.

As in Example 5.2, it is a straightforward matter to show that

$$(R \quad (\phi))\ \ = R \quad (\ \ \phi), \qquad R \quad (\phi)\, R \quad (\ \ \phi) = I ,$$

and that, therefore,

$$R \quad (\phi)(R \quad (\phi))\ \ = (R \quad (\phi))\ \ R \quad (\phi) = I .$$

Hence $R \quad (\phi) \quad ^{\times}$ is an orthogonal matrix for any p, q such that $1 \quad p < q \quad n$, and any $\phi \quad [\ \pi, \pi]$.

The basic result underlying Jacobi's method is encapsulated in the next theorem.

Theorem 5.2 *Suppose that* $A \quad ^{\times}$. *For each pair of integers* (p, q) *with* $1 \quad p < q \quad n$, *there exists* $\phi \quad [\ \pi/4, \pi/4]$ *such that the* (p, q) *entry of the symmetric matrix* $R \quad (\phi) \quad AR \quad (\phi)$ *is equal to* 0.

Proof For the sake of notational simplicity, we shall write R instead of $R \quad (\phi)$ throughout the proof, and abbreviate $c = \cos \phi$ and $s = \sin \phi$.

Consider the product $A' = AR$. Evidently the only difference between A' and A is in columns p and q; these columns of A' are linear combinations of the same two columns of A:

$$\begin{array}{rcl} a'_{ip} & = & a_{ip}\, c \quad a_{iq}\, s \\ a'_{iq} & = & a_{ip}\, s + a_{iq}\, c \end{array} \quad , \quad i = 1, 2, \ldots, n \,. \tag{5.3}$$

Multiplication of A' by R on the left gives a similar result, but affects rows p and q, rather than columns p and q. Writing $B = R \ A'$ gives

$$\begin{array}{rcl} b_{pj} & = & a'_{pj}\, c \quad a'_{qj}\, s \\ b_{qj} & = & a'_{pj}\, s + a'_{qj}\, c \end{array} \quad , \quad j = 1, 2, \ldots, n \,. \tag{5.4}$$

Combining these equations shows that $B = R \ A R$, where

$$\begin{array}{rcl} b_{pp} & = & a_{pp}\, c \quad 2a_{pq}\, sc + a_{qq}\, s \,, \\ b_{qq} & = & a_{pp}\, s \ + 2a_{pq}\, sc + a_{qq}\, c \,, \\ b_{pq} & = & (a_{pp} \quad a_{qq})sc + a_{pq}\, (c^2 \quad s^2) = b_{qp} \,. \end{array} \tag{5.5}$$

The remaining elements of $B = R \ A R$ in columns p and q are given by the expressions

$$\begin{array}{rcl} b_{ip} & = & a_{ip}\, c \quad a_{iq}\, s \\ b_{iq} & = & a_{ip}\, s + a_{iq}\, c \end{array} \quad , \quad i = 1, 2, \ldots, n, \quad i = p, q \,.$$

The matrix $B = R \ A R$ is evidently symmetric, so the nondiagonal elements of B in rows p and q are also given by the same expressions.

Finally, we note that all the elements of B which do not lie either in row p or q or in column p or q are the same as the corresponding elements of A, that is,

$$b_{ij} = a_{ij} \,, \qquad \text{if } i = p, q \text{ and } j = p, q \,.$$

We see from (5.5) that in order to ensure that b_{pq}, the (p, q)-entry of the matrix $B = R \ A R$, is equal to 0, it suffices to choose ϕ such that

$$\tan 2\phi = \frac{2a_{pq}}{a_{pp} \quad a_{qq}} \,; \tag{5.6}$$

thus we select

$$\phi$$

and $B = R^T A R$ *where* R $^{\times}$ *is an orthogonal matrix, then*

$$b = a . \tag{5.9}$$

The quantity

$$A_{\$} = a$$

is called the **Frobenius norm** of A $^{\times}$. The Frobenius norm of A $^{\times}$ is the 2-norm of A, with A regarded as an element of a linear space of dimension n over the field of real numbers; however, it is *not* a subordinate norm in the sense of Definition 2.10. In particular, the Frobenius norm on $^{\times}$ is not subordinate to the 2-norm on .

Now, one can express (5.9) equivalently by saying that the Frobenius norm of a symmetric matrix A is invariant under an orthogonal transformation: $R^T AR_{\$} = A_{\$}$.

Proof of lemma The sum of squares of the elements of A is the same as the trace of A , for

$$\text{Trace}(A) = (A) = a\,a = a , \tag{5.10}$$

since A is symmetric. Analogously, as $B = R^T AR$ is symmetric, we have that

$$\text{Trace}(B) = b .$$

Thus, it remains to show that $\text{Trace}(B) = \text{Trace}(A)$. Now,

$$B = (R^T AR)(R^T AR) = R^T A R, \tag{5.11}$$

since R is orthogonal. Hence B is an orthogonal transformation of A which, by virtue of Theorem 5.1 (vi), means that B and A have the same eigenvalues, and therefore the same trace, since the trace is the sum of the eigenvalues (see Theorem 5.1 (viii)). □

1 9 F 9 B ? H, #. @ $
F 6 C * % #@#& F 6D , 6 1 $
6 , 1 , 6 1 J G 6
6 1 , 1 , 4

Now we are ready to embark on the convergence analysis of the classical Jacobi method.

Theorem 5.3 *Suppose that* $A \in \mathbb{R}^{n \times n}$, $n \geq 2$. *In the classical Jacobi method the off-diagonal entries in the sequence of matrices* $(A^{(k)})$, *generated from* $A^{(0)} = A$ *according to Definition 5.3, converge to* 0 *in the sense that*

$$\lim_{k \to \infty} [a_{ij}^{(k)}]^2 = 0. \tag{5.12}$$

Furthermore,

$$\lim_{k \to \infty} [a_{ii}^{(k)}]^2 = \text{Trace}(A). \tag{5.13}$$

Proof Let a_{pq} be the off-diagonal element of A with largest absolute value, and let $B = (R^{(k)})^T A R^{(k)}$, where ϕ is defined by (5.7). Then, letting $c = \cos\phi$ and $s = \sin\phi$, we have that

$$\begin{bmatrix} b_{pp} & b_{pq} \\ b_{qp} & b_{qq} \end{bmatrix} = \begin{bmatrix} c & -s \\ s & c \end{bmatrix} \begin{bmatrix} a_{pp} & a_{pq} \\ a_{qp} & a_{qq} \end{bmatrix} \begin{bmatrix} c & s \\ -s & c \end{bmatrix},$$

and Lemma 5.1 implies that

$$b_{pp}^2 + 2b_{pq}^2 + b_{qq}^2 = a_{pp}^2 + 2a_{pq}^2 + a_{qq}^2.$$

Writing

$$S(A) = \sum_{i \neq j} a_{ij}^2, \quad D(A) = \sum_{i=1}^{n} a_{ii}^2, \quad L(A) = \sum_{i<j} a_{ij}^2,$$

it follows that $S(A) = D(A) + L(A)$. Now $S(B) = S(A)$ by Lemma 5.1, and so $D(B) + L(B) = D(A) + L(A)$. The diagonal entries of B are the same as those of A, except the ones in rows p and q, $1 \leq p < q \leq n$. Further, as $b_{pq} = 0$, it follows that $b_{pp}^2 + b_{qq}^2 = a_{pp}^2 + a_{qq}^2 + 2a_{pq}^2$. Therefore,

$$D(B) = D(A) + 2a_{pq}^2.$$

Consequently,

$$L(B) = L(A) - 2a_{pq}^2.$$

Now a_{pq}^2 is the largest off-diagonal element of A; hence $L(A) \leq N a_{pq}^2$ where $N = n(n-1)$ is the number of off-diagonal elements, and therefore

$$L(B) \leq (1 - 2/N) L(A). \tag{5.14}$$

On writing $A^{(0)} = A$, $A^{(1)} = B$, and generating subsequent members of the sequence $(A^{(k)})$ in a similar manner, as indicated in the algorithm in Definition 5.3, we deduce from (5.14) that

$$0 \leq L(A^{(k+1)}) \leq (1 - 2/N)\, L(A^{(k)}), \qquad k = 1, 2, 3, \ldots, \tag{5.15}$$

where $N \geq 2$. Thus we conclude that $\lim_{k \to \infty} L(A^{(k)}) = 0$.

Now, (5.13) follows from (5.10) and (5.12) on noting that

$$\mathrm{Trace}(A^{(k)}) = S(A^{(k)}) = S(A^{(k+1)}) = D(A^{(k+1)}) + L(A^{(k+1)}) \qquad k \geq 0,$$

and passing to the limit $k \to \infty$: $\mathrm{Trace}(A^{(k)}) = \lim_{k \to \infty} D(A^{(k)})$. □

According to Theorem 5.1 (viii) the trace of $A^{(k)}$ is the sum of the eigenvalues of $A^{(k)}$, and the eigenvalues of $A^{(k)}$ are the squares of the eigenvalues of A. Thus, we have shown that the sum of the squares of the diagonal elements in the sequence of matrices $(A^{(k)})$ generated by the classical Jacobi method converges to the sum of the squares of the eigenvalues of A. More work is required to show that for each $i = 1, 2, \ldots, n$ the sequence of diagonal elements $(a_{ii}^{(k)})$ converges to an eigenvalue of A as $k \to \infty$. We shall further discuss this question in the final paragraphs of Section 5.4. First, however, we describe another variant of Jacobi's method.

Definition 5.4 (The serial Jacobi method) *This version of Jacobi's method proceeds in a systematic order, using transformations* $R(\phi)$ *to reduce to zero the elements* $(1,2)$, $(1,3)$, …, $(1,n)$, $(2,3)$, $(2,4)$, …, $(2,n)$, …, $(n-1,n)$ *in this order. The complete step is then repeated iteratively.*

It is not difficult to prove that this method also converges. Both these variants of the Jacobi method converge quite rapidly; the rate of convergence is in practice much faster than is suggested by (5.15), and in fact it can be shown that convergence is ultimately quadratic.

It is time for an example!

Example 5.3 *Let us consider the* 5×5 *matrix*

$$A = \begin{matrix} 4 & 1 & 2 & 1 & 2 \\ 1 & 3 & 0 & 3 & 4 \\ 2 & 0 & 1 & 2 & 2 \\ 1 & 3 & 2 & 4 & 1 \\ 2 & 4 & 2 & 1 & 1 \end{matrix} . \tag{5.16}$$

The values of $D(A^{(r)})$ *and* $L(A^{(r)})$ *after each iteration of the serial Jacobi method, with* $A^{(0)} = A$*, are shown in Table 5.1. The off-diagonal elements of the third iterate,* $A^{(3)}$*, are zero to 10 decimal digits. The diagonal elements of* $A^{(3)}$*, which give the eigenvalues, are*

$$8.094,\ 1.690,\quad 0.671,\ 7.170,\quad 3.282\,.$$

Note that the eigenvalues do not appear in any particular order.

Table 5.1. *Convergence of the serial Jacobi iteration.*

r	$\|D(A^{(r)})\|_F$	$\|L(A^{(r)})\|_F$
0	12.124356	9.165152
1	8.631713	3.544458
2	8.095460	1.691211
3	8.094285	0.000000

This concludes the discussion about the use of Jacobi's method for computing the eigenvalues of a symmetric matrix A. 'Fine,' you might say, 'but how do we determine the *eigenvectors* of A?'

It turns out that by collecting the information accumulated in the course of the Jacobi iteration, it is fairly easy to calculate the eigenvectors of A. We begin by noting that if M is an orthogonal matrix such that $M^{T}AM = D$, where D is diagonal, then the diagonal elements of D are the eigenvalues of A, and the columns of M are the corresponding eigenvectors of A.

In the course of the Jacobi iteration (be it classical or serial), we have constructed the plane rotations $R^{(r)}(\phi)$, $j = 1, 2, \ldots, k$. Thus, an approximation $M^{(r)}$ to the orthogonal matrix M can be obtained by considering the product of these rotation matrices: initially, we put $M^{(0)} = I$ and then we apply the column transformation $R^{(r)}(\phi)$ at each step $j = 1, 2, \ldots, k$. This corresponds to multiplying $M^{(r-1)}$ on the right by $R^{(r)}(\phi)$ for $j = 1, 2, \ldots, k$, and leads to the orthogonal matrix

$$M^{(r)} = R^{(r)}(\phi)\ldots R^{(r)}(\phi)$$

which represents the required approximation to the orthogonal matrix M. The columns of $M^{(r)}$ will be the desired approximate eigenvectors

of A corresponding to the approximate eigenvalues which appear along the diagonal of A .

The Jacobi method usually converges in a reasonable number of iterations, and is a satisfactory method for small or moderate-sized matrices. However, there are many problems, particularly in the area of numerical solution of partial differential equations, which give rise to very large matrices that are sparse, with most of the elements being zero. A further consideration is that in many practical situations one does not need to compute all the eigenvalues. It is much more common to require a few of the largest eigenvalues and corresponding eigenvectors, or perhaps a few of the smallest. Jacobi's method is not suitable for such problems, as it always produces all the eigenvalues, and will not preserve the sparse structure of a matrix during the course of the iteration. For example, it is easy to see that if Jacobi's method is applied to a symmetric tridiagonal matrix, then at the end of one sweep all (but two) of the elements of the matrix will in general be nonzero and, although still symmetric, the transformed matrix is no longer tridiagonal. Later on in this chapter we shall consider numerical algorithms for computing selected eigenvalues of a matrix. Thus, as an overture to what will follow, we now outline a 'rough and ready' technique for locating the eigenvalues.

5.4 The Gerschgorin theorems

Gerschgorin's Theorem provides a very simple way of determining a region that contains the eigenvalues of a matrix. It is very general, and does not assume that the matrix is symmetric; in fact we shall allow the elements of a square matrix of order n to be complex and write A × to express this fact.

Definition 5.5 *Suppose that* n 2 *and* A × . *The* **Gerschgorin discs** D , i = 1, 2, . . . , n, *of the matrix* A *are defined as the closed circular regions*

$$D = z : z \quad a \quad R \tag{5.17}$$

in the complex plane, where

$$R = a \tag{5.18}$$

is the radius of D .

1 %1 4%4 F , L , 5 6 1 7 9 7 K ,
9 , I 4%4 F , B#@ #C#@**D <F 2 % 4%4 F
B#@ #C#@**D= 2 ' ' ' 2 + ' ' 4 # #>C#@ #@@.4

Theorem 5.4 (Gerschgorin's Theorem) *Let* $n \geq 2$ *and* $A \in \mathbb{C}^{n \times n}$. *All eigenvalues of the matrix* A *lie in the region* $D = \bigcup_{i=1}^{n} D_i$, *where* D_i, $i = 1, 2, \ldots, n$, *are the Gerschgorin discs of* A *defined by (5.17), (5.18).*

Proof Suppose that λ and $x \neq \mathbf{0}$ are an eigenvalue and the corresponding eigenvector of A, so that

$$a_i x = \lambda x_i, \quad i = 1, 2, \ldots, n. \tag{5.19}$$

Suppose that x_k, with $k \in 1, 2, \ldots, n$, is the component of x which has largest modulus, or one of those components if more than one have the same modulus. We note in passing that $x_k \neq 0$, given that $x \neq \mathbf{0}$; also,

$$|x_j| \leq |x_k|, \quad j = 1, 2, \ldots, n. \tag{5.20}$$

This means that

$$\begin{aligned} |\lambda - a_{kk}||x_k| &= |\lambda x_k - a_{kk} x_k| \\ &= |\sum_{j \neq k} a_{kj} x_j| \\ &\leq \sum_{j \neq k} |a_{kj}||x_j| \\ &\leq \sum_{j \neq k} |a_{kj}||x_k| \\ &= R_k |x_k|, \end{aligned} \tag{5.21}$$

which, on division by $|x_k|$, shows that λ lies in the Gerschgorin disc D_k of radius R_k centred at a_{kk}. Hence, $\lambda \in D = \bigcup_{i=1}^{n} D_i$. □

Theorem 5.5 (Gerschgorin's Second Theorem) *Let* $n \geq 2$. *Suppose that* $1 \leq p < n$ *and that the Gerschgorin discs of the matrix* $A \in \mathbb{C}^{n \times n}$ *can be divided into two disjoint subsets* D_1 *and* D_2, *containing* p *and* $q = n - p$ *discs respectively. Then, the union of the discs in* D_1 *contains* p *of the eigenvalues, and the union of the discs in* D_2 *contains* $n - p$ *eigenvalues. In particular, if one disc is disjoint from all the others, it contains exactly one eigenvalue, and if all the discs are disjoint then each disc contains exactly one eigenvalue.*

Proof We shall use a so-called *homotopy* (or continuation) argument.

For $0 \quad \varepsilon \quad 1$, we consider the matrix $B(\varepsilon) = (b \ (\varepsilon))^{\times}$, where

$$b \ (\varepsilon) = \begin{cases} a & \text{if } i = j, \\ \varepsilon a & \text{if } i = j. \end{cases} \tag{5.22}$$

Then, $B(1) = A$, and $B(0)$ is the diagonal matrix whose diagonal elements coincide with those of A. Each of the eigenvalues of $B(0)$ is therefore the centre of one of the Gerschgorin discs of A; thus exactly p of the eigenvalues of $B(0)$ lie in the union of the discs in D. Now, the eigenvalues of $B(\varepsilon)$ are the zeros of its characteristic polynomial, which is a polynomial whose coefficients are continuous functions of ε; hence the zeros of this polynomial are also continuous functions of ε. Thus as ε increases from 0 to 1 the eigenvalues of $B(\varepsilon)$ move along continuous paths in the complex plane, and at the same time the radii of the Gerschgorin discs increase from 0 to the radii of the Gerschgorin discs of A. Since p of the eigenvalues lie in the union of the discs in D when $\varepsilon = 0$, and these discs are disjoint from all of the discs in $D(q)$, these p eigenvalues must still lie in the union of the discs in D when $\varepsilon = 1$, and the theorem is proved.

The same proof evidently still applies when the discs can be divided into any number of disjoint subsets. □

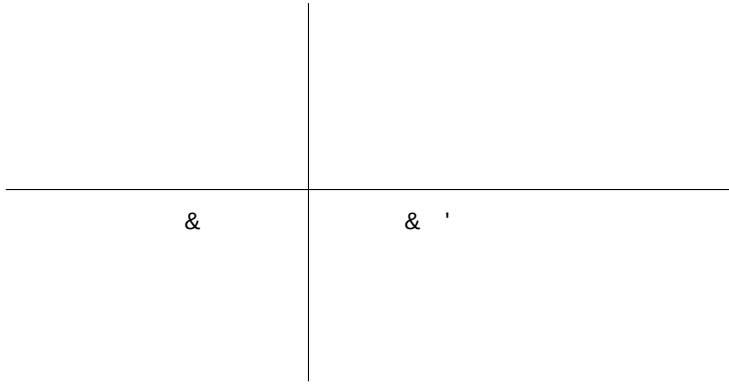
Example 5.4 *Consider the matrix*

$$A = \begin{matrix} 4.00 & 0.20 & 0.10 & 0.10 \\ 0.20 & 1.00 & 0.10 & 0.05 \\ 0.10 & 0.10 & 3.00 & 0.10 \\ 0.10 & 0.05 & 0.10 & 3.00 \end{matrix} . \tag{5.23}$$

Figure 5.1 shows, as solid circles, the Gerschgorin discs for this matrix; for instance, one of the discs has centre at 4.00 *and radius* 0.40. *The discs are clearly disjoint, so that each disc contains one eigenvalue of the matrix. The significance of the dotted circles will be explained in our next example.*

Example 5.5 *Let us consider the matrix* A *defined by (5.23), and then transform it into* $B = KAK^{-}$, *where* $K^{\times}$ *is the same as the identity matrix except that* $k = \kappa > 0$.

This transformation has the effect of multiplying the elements in row 2 by κ, and multiplying the elements in column 2 by $1/\kappa$; the diagonal element a thus remains unaltered. A small value of κ then means that the second disc of B is smaller than the second disc of A, but the other



discs grow larger. The dotted discs in Figure 5.1 are for the matrix B with $\kappa = 1/23$. For this value the other discs are still just disjoint from the disc centred at 1.00; the disc with centre at 4.00 almost touches the disc with centre at 1.00. The disc with centre 1.00 has radius 0.014, and is too small to be visible in the figure. The eigenvalue in this disc is 1.009 to three decimal digits. The same procedure can be used to reduce the size of each of the discs in turn.

This idea is formalised in the next theorem.

Theorem 5.6 *Let* $n \quad 2$*, and suppose that in the matrix* $A \quad ^{\times}$ *all the off-diagonal elements are smaller in absolute value than* ε*, so that* $a \quad < \varepsilon$*, for all* $i, j \quad 1, 2, \ldots, n$ *with* $i = j$*. Suppose also that for a particular integer* $r \quad 1, 2, \ldots, n$ *the diagonal element* a *is distant* δ *from all the other diagonal elements, so that* $a \quad a \quad > \delta$*, for all* i *such that* $i = r$*. Then, provided that*

$$\varepsilon < \frac{\delta}{2(n \quad 1)}, \tag{5.24}$$

there is an eigenvalue λ *of* A *such that*

$$|\lambda - a_{rr}| < 2(n-1)\varepsilon/\delta. \tag{5.25}$$

Proof We apply the **similarity transformation**

$$A \to A' = K A K^{-1},$$

where K is the same as the identity matrix, except that the diagonal element in row r is chosen to be $k_{rr} = \kappa > 0$. This has the effect of multiplying the off-diagonal elements of row r by κ, and the element in column r of row i, where $i \neq r$, by $1/\kappa$. The Gerschgorin disc from row r then has centre a_{rr} and radius not exceeding $\kappa(n-1)\varepsilon$, and the disc corresponding to row $i = r$ has centre a_{rr} and radius not exceeding $(n-2)\varepsilon + \varepsilon/\kappa$.

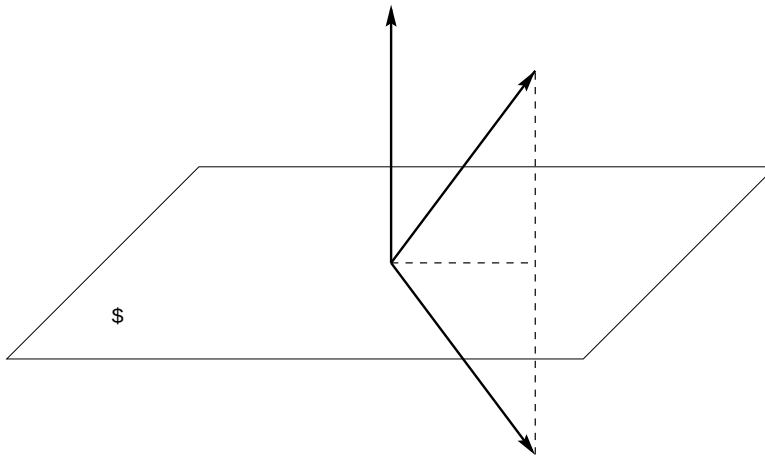
We now want to reduce the size of disc r by choosing a small value of κ, while keeping it disjoint from the rest. This is easily done by choosing $\kappa = 2\varepsilon/\delta$. The radius of disc r does not exceed $2(n-1)\varepsilon/\delta$, and the radius of disc $i = r$ does not exceed $(n-2)\varepsilon + \varepsilon/2\delta$. The sum of these radii therefore satisfies

$$\begin{aligned} R_r + R_i &\leq 2(n-1)\varepsilon/\delta + (n-2)\varepsilon + \varepsilon/2\delta \\ &< \varepsilon + (n-2)\varepsilon + \varepsilon/2\delta \\ &< \delta, \end{aligned} \tag{5.26}$$

where we have used the given condition (5.24) twice. As the centres a_{rr} and a_{ii} of these discs are distant more than δ from each other, (5.26) shows that the two discs are disjoint, and the required result is proved. □

Theorem 5.6 is sufficient to show that for a matrix satisfying its hypotheses we can find a Gerschgorin disc whose radius is of order ε provided that ε is sufficiently small. It also indicates that the spacing between the diagonal elements is important.

In particular, Theorem 5.6 applies to the matrix $A^{(k)}$ which results after k iterations of the Jacobi method. If at that stage all the off-diagonal elements have magnitude less than ε then there is one eigenvalue in each of the intervals $[a_{rr} - (n-1)\varepsilon, a_{rr} + (n-1)\varepsilon]$, provided that these intervals are disjoint; this follows from Theorem 5.5. If ε is sufficiently small compared with the distances between the diagonal elements of $A^{(k)}$, Theorem 5.6 may be used to give closer bounds on the eigenvalues.



hyperplane $ consisting of all vectors that are perpendicular to in is invariant under the mapping H . Finally, for any ,

$$\mathsf{H} \quad = \quad .$$

Hence, if the angle between and is denoted by ϕ, then the angle between and H is equal to $\pi+\phi$. We conclude from these observations that the vector H is the reflection of in the hyperplane $. For this reason, the mapping H is frequently referred to as **Householder reflector**, corresponding to the vector (see Figure 5.2).

Lemma 5.2 *Every Householder matrix is symmetric and orthogonal.*

Proof As $\mathsf{I} = \mathsf{I}$, ( ) = ( ) = , and is a (positive real) number, the symmetry of H follows. The orthogonality of H is a consequence of the identity

$$\mathsf{H}\ \mathsf{H} = \mathsf{H}\mathsf{H} = \mathsf{H} = \mathsf{I} \quad \frac{4}{} \quad + \frac{4}{(\quad)}(\quad)(\quad) = \mathsf{I},$$

since ( )( ) = ( ) = ( ) by the associativity of matrix multiplication. □

Lemma 5.3 *Let* 1 k < n *and suppose that* H *is a* k k *Householder matrix. Then, the matrix* H × *, written in partitioned form as*

$$H = \begin{matrix} I_{-} & 0 \\ 0 & H \end{matrix}$$

where I _ *is the identity matrix of order* n k *and* 0 *is the* (n k) k *zero matrix, is also a Householder matrix.*

The proof of this lemma is straightforward and is left as an exercise. (See Exercise 1.)

Lemma 5.4 *Given any vector* □*, there exists a Householder matrix* H × *such that all elements of the vector* H *are zero, except the first; i.e.,* H *is a nonzero multiple of* e *, the first column of the identity matrix.*

In geometrical terms this result can be rephrased by saying that for any vector □ there exists an (n 1)-dimensional hyperplane $ passing through the origin in such that the reflection H of in $ is equal to a nonzero multiple of e . To find $ it suffices to identify a vector □ normal to $. Since $ is unaffected by rescaling (see Definition 5.6), the length of is immaterial. As noted in the discussion following Definition 5.6, the vectors H , and are coplanar. Therefore, we shall seek □ as a suitable linear combination of and .

Proof of lemma We seek H = I [2/( )] with = + ce , where c is a nonzero real number to be determined. Hence,

$$\begin{aligned} &= + c\beta, \\ &= + 2c\beta + c , \end{aligned}$$

where β = e is the first entry of . A simple manipulation then shows that

$$H = \frac{2}{} ( ) = \frac{(c) 2c(+ c\beta)e}{+ 2c\beta + c}.$$

Thus, H will be a multiple of e provided that we choose c so that c = . Also, to avoid division by 0, we need to ensure that + 2cβ + c = 0. To do so, note that c β ; therefore

$$+ 2c\beta + c (\beta + c) = 0,$$

provided that $\beta + c = 0$, which can be ensured by selecting the appropriate sign for c, that is, by defining

$$c = \begin{cases} (\underline{\text{sign}}\,\beta)\sqrt{\quad} & \text{when } \beta = 0\,, \\ & \text{when } \beta = 0\,. \end{cases}$$

With this choice of c, we have $H = ce$, as required. □

We now show how Householder matrices can be used to reduce a given matrix to tridiagonal form.

Theorem 5.7 *Given that* A $\times$ *and* n 3, *there exists a matrix* Q $\times$, *a product of* n 2 *Householder matrices* H $\times$, $k = 2, \ldots, n$ 1, *given by*

$$Q = H \; - \; H \; - \; \ldots H$$

such that Q $AQ = T$ *is tridiagonal; the matrix* Q *is orthogonal.*

Proof The proof of the theorem will proceed by induction. Before embarking on this, we make some preparatory observations which highlight the key ideas in the proof.

Consider the matrix A $\times$, partitioned by its first row and column in the form

$$A = \begin{matrix} \alpha & \\ & C \end{matrix},$$

where α , $^{-}$ and C $^{-} \times ^{-}$, and define

$$\% = \quad : = (\lambda, 0, \ldots, 0) \quad \text{for some } \lambda \quad .$$

If happens to belong to $^{-}$, then, by Lemma 5.4, there exists an $(n \quad 1) \quad (n \quad 1)$ Householder matrix H_{-} such that each element of H_{-}, except the first, is equal to 0. If, on the other hand, $= \mathbf{0}$, then $H_{-} = \mathbf{0}$, trivially. Either way, $H_{-} \quad \% \; ^{-}$.

Let us extend the Householder matrix H_{-} $^{-} \times ^{-}$, using Lemma 5.3 with $k = n \quad 1$, to a Householder matrix H_{-} $\times$ by defining the $(1,1)$-entry of H_{-} as 1 and choosing the remaining entries in the first row and first column of H_{-} as 0. Then,

$$\begin{aligned} H_{-} AH_{-} &= \begin{matrix} 1 & \mathbf{0} \\ \mathbf{0} & H_{-} \end{matrix} \begin{matrix} \alpha & \\ & C \end{matrix} \begin{matrix} 1 & \mathbf{0} \\ \mathbf{0} & H_{-} \end{matrix} \\ &= \begin{matrix} \alpha & \\ & D \end{matrix}, \end{aligned} \tag{5.27}$$

where

=

orthogonal, Q $^{\times}$ is itself orthogonal. Moreover, for any $\mathbf{f}$ % we have Q $\mathbf{f}$ %, since the $(1,1)$-entry of Q is 1 and the remaining entries in the first column of Q are 0. This concludes the inductive step, and completes the proof. □

The recursive transformation of a symmetric matrix to tridiagonal form outlined in the proof of Theorem 5.7 is called **Householder's method**. In implementing this method in practice it is important to carry out the transformations efficiently. Counting the arithmetic operations involved is straightforward but tedious, and shows that the complete reduction requires approximately $-\mathsf{n}$ multiplications, for a moderately large value of n.

Example 5.6 *In order to illustrate Householder's method, we return to the matrix* A *defined in (5.16). The first stage uses the Householder matrix defined by the vector*

$$= (0.000,\ 4.162,\ 2.000,\ 1.000,\ 2.000) \quad . \tag{5.28}$$

The result of the transformation is the matrix

$$\begin{array}{ccccc} 4.000 & 3.162 & 0.000 & 0.000 & 0.000 \\ 3.162 & 5.300 & 1.232 & 0.332 & 0.284 \\ 0.000 & 1.232 & 1.653 & 3.312 & 0.275 \\ 0.000 & 0.332 & 3.312 & 5.149 & 1.123 \\ 0.000 & 0.284 & 0.275 & 1.123 & 3.102 \end{array} \quad .$$

The leading element of the matrix is unchanged, and the first row and column have tridiagonal structure.

The second stage uses the Householder matrix with the vector

$$= (0.000,\ 0.000,\ 2.540,\quad 0.332,\ 0.284) \tag{5.29}$$

and gives the new matrix

$$\begin{array}{ccccc} 4.000 & 3.162 & 0.000 & 0.000 & 0.000 \\ 3.162 & 5.300 & 1.308 & 0.000 & 0.000 \\ 0.000 & 1.308 & 0.057 & 2.166 & 0.792 \\ 0.000 & 0.000 & 2.166 & 6.610 & 0.420 \\ 0.000 & 0.000 & 0.792 & 0.420 & 2.967 \end{array} \quad .$$

This time the leading 2 2 minor is unaltered, and the first two rows and columns have tridiagonal structure.

The final stage uses the Householder matrix with vector

$$= (0.000,\ 0.000,\ 0.000,\quad 4.471,\ 0.792) \tag{5.30}$$

and gives the tridiagonal matrix

$$\begin{array}{ccccc} 4.000 & 3.162 & 0.000 & 0.000 & 0.000 \\ 3.162 & 5.300 & 1.308 & 0.000 & 0.000 \\ 0.000 & 1.308 & 0.057 & 2.306 & 0.000 \\ 0.000 & 0.000 & 2.306 & 5.208 & 3.411 \\ 0.000 & 0.000 & 0.000 & 3.411 & 1.565 \end{array} \quad . \qquad (5.31)$$

The numerical values are quoted here to three decimal digits, for simplicity.

Having shown how to transform a symmetric matrix into tridiagonal form, we can now consider the problem of determining the eigenvalues of a tridiagonal matrix.

5.6 Eigenvalues of a tridiagonal matrix

Before developing a numerical algorithm for calculating the eigenvalues and the eigenvectors of a symmetric tridiagonal matrix, let us spend some time exploring the location of the eigenvalues. The main result of this section is the so-called Sturm sequence property, stated in Theorem 5.9, which enables us to specify the number of eigenvalues of a symmetric tridiagonal matrix which exceed a given real number ϑ. The proof of the Sturm sequence property is based on Cauchy's Interlace Theorem which is of independent interest, and proving the latter is our first task.

To simplify the notation we now write the symmetric tridiagonal matrix in the form

$$T = \begin{pmatrix} a & b & & & & & \\ b & a & b & & & & \\ & b & a & b & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & \\ & & & & \ddots & \ddots & \ddots \\ & & & & & b & a \end{pmatrix} \quad .$$

[illegible]

The determinants of the successive principal minors of a matrix of this form can easily be calculated by recurrence. Defining $p(\lambda)$ to be the determinant of the leading principal minor of order r of $T - \lambda I$, we see that

$$\begin{aligned} p_1(\lambda) &= a - \lambda, \\ p_2(\lambda) &= (a - \lambda)(a - \lambda) - b^2. \end{aligned}$$

Expanding $p_r(\lambda)$ in terms of the elements of the last row, and then in terms of the last column, we obtain the relation

$$p_r(\lambda) = (a - \lambda)p_{r-1}(\lambda) - b^2 p_{r-2}(\lambda), \qquad r = 2, 3, \ldots, n,$$

with the convention that

$$p_0(\lambda) = 1.$$

In the rest of this section we shall assume that all the off-diagonal elements b are nonzero. For suppose that $b_k = 0$ for some k in the set $2, 3, \ldots, n$; then, the eigenvalues of the matrix T comprise the eigenvalues of the matrix consisting of the first $k - 1$ rows and columns, together with the eigenvalues of the matrix consisting of the last $n - k + 1$ rows and columns. These two problems become separated and can be treated independently; if several of the off-diagonal elements are zero, the matrix can be partitioned into a number of smaller matrices which can then be dealt with independently.

Theorem 5.8 (Cauchy's Interlace Theorem) *Let $n \geq 3$. The roots of p_r separate those of p_{r-1}, for $r = 1, 2, \ldots, n - 1$; i.e., between two consecutive roots of p_r there is exactly one root of the polynomial p_{r-1}, $r = 1, 2, \ldots, n - 1$.*

Proof The proof is by induction. It is trivial to show that the property holds for $r = 1$: the two roots

$$-a_1 + a_2 \pm \sqrt{(a_1 - a_2)^2 + 4b_2^2}$$

of p_2 are separated by a_1, the only root of the linear polynomial p_1.

Suppose that the statement is true when $r = i - 1, 2 \leq i \leq n - 1$, so that the roots of p_{i-1} separate those of p_{i-2}. On denoting by α and β two consecutive roots of p_{i-2}, the inductive hypothesis implies that p_{i-1} has exactly one root between α and β, which means that $p_{i-1}(\alpha)$ and

$p_-(\beta)$ have opposite signs. Now,

$$p(\lambda) = (a - \lambda)p(\lambda) - b\,p_-(\lambda),$$

so that, as α and β are roots of p, it follows that $p(\alpha)$ and $p(\beta)$ also have opposite signs. Hence p has at least one root between α and β. Choosing α and β to be each pair of consecutive roots of p in turn we have therefore located $i - 1$ roots of p.

Next choose α to be the algebraically smallest root of p

β of p_{r-1}. Then, there is exactly one root of p_r between α and β; denote this root by ϕ. As we saw in the proof of the previous theorem $p_r(\lambda)$ is positive when λ is large and negative, and the sign of $p_r(\lambda)$ is determined by the number of roots of p_r which are less than λ. Hence if $\vartheta < \phi$ both p_r and p_{r-1} have the same number of roots less than ϑ, so that $p_r(\vartheta)$ and $p_{r-1}(\vartheta)$ have the same sign, and $s(\vartheta) = s_{-1}(\vartheta) + 1$. Also if p_r and p_{r-1} have the same number of roots less than ϑ, then p_r must have one more root which is greater than ϑ; this means that $g(\vartheta) = g_{-1}(\vartheta) + 1$. Hence $s(\vartheta) = g(\vartheta)$. A similar argument shows that $s(\vartheta) = g(\vartheta)$ in the alternative situation where $\vartheta > \phi$. It is also a simple matter to modify the argument slightly for the cases where ϑ is less than the smallest root of p_{r-1}, or greater than the largest root of p_{r-1}, and so the inductive step is complete. □

The theorem and proof do not allow for any of the members of the sequence being zero, in which case the sign becomes undefined. A more careful analysis is tedious but not difficult; it shows that the theorem still holds if we adopt the convention that when $p_r(\vartheta)$ is zero it is given the same sign as $p_{r-1}(\vartheta)$. As we have already seen, two consecutive members of the sequence cannot both be zero.

Our next example will illustrate the application of the Sturm sequence property.

Example 5.7 *Determine the second largest eigenvalue of the matrix*

$$A = \begin{matrix} 3 & 1 & 0 & 0 \\ 1 & 1 & 2 & 0 \\ 0 & 2 & 1 & 1 \\ 0 & 0 & 1 & 1 \end{matrix} . \tag{5.32}$$

If the eigenvalues are λ_r, where $\lambda_1 > \lambda_2 > \lambda_3 > \lambda_4$, we wish to find λ_2. Now, it is easy to see from Theorem 5.5 that all the eigenvalues lie in the interval $[-4, 4]$. We take the midpoint of this interval, and evaluate the Sturm sequence with $\vartheta = 0$, giving

$$p_0(0) = 1, \quad p_1(0) = 3, \quad p_2(0) = -4, \quad p_3(0) = -16, \quad p_4(0) = -12.$$

In this sequence there are three agreements of sign:

$$(1, 3), \quad (-4, -16) \quad \text{and} \quad (-16, -12).$$

Hence $s(0) = 3$, and the matrix has three eigenvalues greater than 0; this means that λ_2 must lie in the right-hand half of the interval

$[\ -4, 4]$, that is, in $[0, 4]$. We construct the Sturm sequence for $\vartheta = 2$, the midpoint of the interval, giving

$$p\ (2) = 1\,,\ \ p\ (2) = 1\,,\ \ p\ (2) = \ \ 4\,,\ \ p\ (2) = \ \ 0\,,\ \ p\ (2) = 4\,.$$

Notice that here $p\ (2)$ is zero, and is given the negative sign to agree with $p\ (2)$. The number of agreements in sign here is two, so two of the eigenvalues are greater than 2, and λ must lie in $[2, 4]$, the right-hand half of the interval $[0, 4]$. For $\vartheta = 3$ we obtain the sequence

$$1\,,\ \ +0\,,\ \ \ 1\,,\ \ 2\,,\ \ \ 3\,,$$

with only one agreement of sign, so this time λ must lie in the left-hand half $[2, 3]$ of the interval $[2, 4]$, and we repeat the process, taking $\vartheta = -$, the midpoint of $[2, 3]$. This time the sequence is

$$1\,,\ \ \frac{1}{2}\,,\ \ \ \frac{11}{4}\,,\ \ \frac{17}{8}\,,\ \ \ \frac{7}{16}\,,$$

with one agreement in sign, showing that $\lambda\ < 2.5$.

The process of bisection can be repeated as many times as required to locate the eigenvalue to a given accuracy. After 13 stages we find that $\lambda\ = 2.450$ correct to three decimal digits.

This method is very similar to the usual bisection process for finding a solution of $f(x) = 0$, beginning with an interval $[a, b]$ such that $f(a)$ and $f(b)$ have opposite signs. A great advantage of the Sturm sequence method is that it not only determines the eigenvalue, but also indicates which eigenvalue it is. If we used the Jacobi method of Section 5.3 we would have to determine *all* the eigenvalues, sort them into order, and then choose the second largest eigenvalue as λ .

The Sturm sequence method will also determine how many eigenvalues of a matrix lie in a given interval (α, β); all that we need is to construct the Sturm sequences $(p\ (\alpha))$ and $(p\ (\beta))$; then, the required number of eigenvalues is $s\ (\alpha)\ \ s\ (\beta)$.

It is very important to calculate the sequence $p\ (\vartheta)$ directly from the recurrence relation. For instance, in Example 5.7, with $\vartheta = 2.445$ we obtain

$$\begin{aligned}
p\ (2.445) &= 1\,,\\
p\ (2.445) &= 3\ \ 2.445 = 0.555\,,\\
p\ (2.445) &= (\ \ 1\ \ 2.445)\ \ 0.555\ \ \ 1\ \ \ 1 = \ \ 2.9120\,,\\
p\ (2.445) &= (1\ \ \ 2.445)\ \ \ 2.9120\ \ \ 4\ \ \ 0.555 = 1.9878\,,\\
p\ (2.445) &= (1\ \ \ 2.445)\ \ \ 1.9878\ \ \ 1\ \ \ \ 2.9120 = 0.0396\,.
\end{aligned}$$

The alternative, to construct explicit forms for the polynomials p (λ), j = 0, 1, . . . , n, and *then* evaluate p (ϑ) by inserting the value of λ = ϑ into each of the polynomials p (λ), will lead to the construction of the explicit form of the characteristic polynomial of the matrix, which is p (λ), and we have already seen that this is affected disastrously by rounding errors. The calculation by direct use of the recurrence relation is perfectly satisfactory.

Example 5.8 *As a second example, we return to the matrix* A *in (5.16), which has been transformed to the tridiagonal form (5.31), to determine the largest eigenvalue.*

Table 5.2. *Bisection process for the largest eigenvalue. In the table* k *denotes the iteration number,* ϑ *the* k*th iterate approximating the unknown eigenvalue* λ *, and* s (ϑ) *signifies the number of sign agreements in the Sturm sequence* p (ϑ), . . . , p (ϑ).

	"	#4*" +
		&
	('%&	
&	$'	
'	% $	
(	) (	
%	) (&	
)	'	
	$	
$	%%	
	$	
	$&	
&	$%	
'	$'	
(	$'	

Table 5.2 shows the result of the bisection process, using the Sturm sequence. The -norm of the tridiagonal matrix is 10.926, so the process begins with the interval [10.926, 10.926]. The largest eigenvalue

1 8 , , 1 5 1
, 5 , A A
L 1 6 8 6 5
1 4

is 8.094, to three decimal digits, agreeing with the result of Jacobi's method, in Section 5.3. This table also shows how some savings are possible when all the eigenvalues are required. We see from the table that use of $\vartheta = 7.511$ gives 1 agreement in sign, while $\vartheta = 6.829$ gives 2 agreements in sign. The bisection process for the second largest eigenvalue can therefore begin with the interval [6.829, 7.511].

The method of bisection may appear rather crude, but it has the great advantage of guaranteed success, and is very little affected by rounding errors. Moreover, the amount of work involved is not large. If we have calculated the squares of the off-diagonal entries, b , of the matrix T in advance, each computation of all members of the sequence requires about $2n$ multiplications. If the bisection process is continued for 40 stages, the eigenvalue will be determined to about nine significant digits, and if we require to calculate m of the eigenvalues to this accuracy, we shall need about $80mn$ multiplications. If m is a good deal smaller than n, the order of the matrix, this is likely to be a great deal smaller than the work involved in the process of reduction to tridiagonal form, which, as we have seen, is about $-n$ multiplications. In most practical problems it is the initial Householder reduction to tridiagonal form which accounts for most of the computational work.

5.7 The QR algorithm

In this section we discuss briefly the QR algorithm, an alternative method for determining the eigenvalues of a tridiagonal matrix. In principle it could be applied to a full matrix, but it is more efficient to use the Householder method to reduce the matrix to tridiagonal form first. The basis of the method is the QR factorisation of the matrix which we have already encountered in Chapter 2, in the solution of least squares problems. In contrast with Section 2.9, however, where we were concerned with the solution of least squares problems for rectangular matrices A $\times$, here the focus is on eigenvalue problems for symmetric tridiagonal matrices A $\times$; we shall therefore revisit the derivation of the QR factorisation by adopting a slightly different approach from the one proposed in Section 2.9.

5.7.1 The QR factorisation revisited

Suppose that n 3 and A $\times$ is a symmetric tridiagonal matrix. We first show how to construct an orthogonal matrix Q $\times$ and

an upper triangular matrix R $^{\times}$ such that $\mathsf{A} =$

which are already equal to zero, remain zero upon multiplication by the next rotation matrix Q in the sequence, we deduce that, after successive multiplications of A on the left by Q , Q ,..., Q , the matrix

$$Q \; Q \; \ldots Q \; A = R, \tag{5.35}$$

is upper triangular. In fact, since A is tridiagonal, R is tridiagonal and upper triangular; consequently, R is **bidiagonal** in the sense that R $= 0$ if $i = j, j$ 1.

As the matrices Q $=$ R $(\phi$), $p = 1, 2, \ldots, n$ 1, are orthogonal, and therefore Q Q $= I$, on multiplying (5.35) on the left by Q Q $\ldots Q$, we find that

$$A = QR,$$

where

$$Q = Q \;\; Q \;\; \ldots Q$$

is an orthogonal matrix (as it is a product of orthogonal matrices). The next subsection describes the QR algorithm, based on the QR factorisation, for the numerical solution of the eigenvalue problem (5.1) where the matrix A $^{\times}$ is symmetric and tridiagonal.

5.7.2 The definition of the QR algorithm

Suppose that A $^{\times}$ is symmetric and tridiagonal. The QR algorithm defines a sequence of symmetric tridiagonal matrices A $^{\times}$, $k = 0, 1, 2, \ldots$, starting with A $= A$, as follows.

Suppose that k 0. The kth step of the QR algorithm takes the symmetric tridiagonal matrix A and chooses a **shift** μ (the choice of μ will be discussed below), then forming the QR factorisation

$$A \quad \mu \; I = Q \quad R \quad .$$

We then multiply Q and R in the reverse order, and construct the new matrix A defined by

$$A \quad = R \quad Q \quad + \mu \; I \, .$$

Recalling that the matrix Q is orthogonal, it is a simple matter to see that A $= Q$ A Q , so that A and A have the same eigenvalues. As A $= A$, all matrices in the sequence (A) have the same eigenvalues as A itself. It is also easy to show that each of the matrices A is symmetric and tridiagonal. (See Exercise 7.)

The choice of the shift parameter μ is very important; if correctly chosen the sequence of matrices A converges very rapidly to a matrix in which one of the off-diagonal elements is zero. If this element is in the first or last row, we have thereby identified one of the eigenvalues; if it is one of the intermediate elements, we can split the matrix into two separate matrices of lower order. In either case we can repeat the iterative process with smaller matrices, until all the eigenvalues are found.

The usual simple choice of the shift parameter in the kth step is

$$\mu = a \quad ,$$

the last diagonal element of the matrix A . In general, after a few steps of the iteration the element at position $(n, n \quad 1)$ will become negligibly small. One of the eigenvalues of the resulting matrix is then the last diagonal element, and we continue the process with the matrix of order $n \quad 1$ obtained by removing the last row and column. There are special circumstances where this choice of shift is unsatisfactory, and other situations where another choice is more efficient, but we shall not discuss the details any further. The proof of the convergence of this method is long and technical; details will be found in the books cited in the Notes at the end of the chapter.

The method does not determine the eigenvalues in any particular order, so if we require only a small number of the largest eigenvalues, for example, the Sturm sequence method is preferable. The usual recommendation is that the QR algorithm should be used on a matrix of order n if more than about $\quad n$ of the eigenvalues are required.

Example 5.9 *We apply the QR algorithm to the tridiagonal matrix (5.31).*

After one step of the iteration the matrix $A = R \quad Q \quad + \mu \quad I$, with $\mu = a \quad = a \quad$, is

$$A = \begin{matrix} 7.034 & 2.271 & 0 & 0 & 0 \\ 2.271 & 2.707 & 0.744 & 0 & 0 \\ 0 & 0.744 & 5.804 & 3.202 & 0 \\ 0 & 0 & 3.202 & 0.464 & 1.419 \\ 0 & 0 & 0 & 1.419 & 2.082 \end{matrix} .$$

In successive iterations $k = 1, 2, 3, 4, 5$, the element a has the values 1.419, 1.262, 0.965, 0.223, 0.002, and after the next iteration a

vanishes to 10 decimal digits. The element a is 3.282, which is therefore an eigenvalue.

We then remove the last row and column, and continue the process on the resulting 4 4 matrix. After just one iteration the element at position $(4, 3)$ vanishes to 7 decimal digits, giving the eigenvalue 0.671. We remove the last row and column and continue with the resulting 3 3 matrix. After one iteration of the resulting 3 3 matrix the element at position $(3, 2)$ is 0.0005, and another iteration gives the accurate eigenvalue 1.690. We are now left with a 2 2 matrix, and the calculation of the last two eigenvalues is trivial. The number of iterations required to isolate each eigenvalue reduces as the algorithm reduces the size of the matrix; this sort of behaviour is typical.

The numerical values agree with those obtained by Jacobi's method, and the bisection method.

5.8 Inverse iteration for the eigenvectors

We saw in Section 5.3 that Jacobi's method can also, if required, produce the eigenvectors of the matrix, but the use of Householder's algorithm, in conjunction with the Sturm sequence method or the QR algorithm, only gives the eigenvalues. Suppose that A $^{\times}$ is a symmetric matrix, and assume that we have a good approximation ϑ to the required eigenvalue λ of A, and some approximation , = 1, to the associated eigenvector , = 1. It is implicitly assumed that $\vartheta = \lambda$ and that ϑ is not an eigenvalue of A, so that the matrix $A \quad \vartheta I$ is nonsingular. The method of **inverse iteration** defines the sequence of vectors , $k = 0, 1, \ldots$, as follows: given , find and then from

$$\begin{aligned} (A \quad \vartheta I) \quad &= \quad , \\ &= \quad c \quad , \end{aligned} \tag{5.36}$$

where $c = 1/$ $= 1/$. Hence, we conclude that $= 1$, $k = 0, 1, 2, \ldots$.

Theorem 5.10 *Suppose that* A $^{\times}$. *The sequence of vectors* ( ) *in* *defined in the process of inverse iteration (5.36) converges to the normalised eigenvector* *corresponding to the eigenvalue* λ *which is closest to* ϑ , *provided that* λ *is a simple eigenvalue and the initial vector* *is not orthogonal to the vector* .

Proof According to Theorem 5.1 (vii), the vector can be expressed

If the estimate ϑ is within rounding error of λ and the eigenvalues are well spaced, the convergence of the sequence ( ) will be extremely rapid: usually a couple of iterations will be sufficient.

The proof of Theorem 5.10 breaks down if $\alpha = 0$, *i.e.*, when the initial vector is exactly orthogonal to the required eigenvector. However, this does not mean that the iteration (5.36) will also break down; for the effect of rounding error will almost always introduce a small multiple of the vector into the expansion of in terms of the with $j = 1, 2, \ldots, n$, and the required eigenvector will then be obtained in a small number of iterations. This is a useful property of the method, since in practice it is not possible to check whether or not is orthogonal to , given that the eigenvector is unknown.

There will also be a problem if there is a multiple eigenvalue, or two eigenvalues are very close together: in the first case $\lambda \quad \vartheta / \lambda \quad \vartheta = 1$ for some $j = s$, and the proof of Theorem 5.10 breaks down; in the second case $\lambda \quad \vartheta / \lambda \quad \vartheta \quad 1$ for some $j = s$, leading to very slow convergence.

The computation of from (5.36) requires the solution of a system of linear equations whose matrix is $A \quad \vartheta I$. This matrix will usually be nearly singular – in fact, our objective in choosing ϑ was to make $A \quad \vartheta I$ exactly singular. In general the solution of such a system is extremely dangerous, because of the effect of rounding errors; in this case, however, the effect of rounding error will be to introduce a multiple of the dominant eigenvector, and this is exactly what is required. An analysis of the effect of rounding errors will confirm this fact, but would take too long here.

There are two ways in which we can implement the inverse iteration process. One obvious possibility would be to use the original matrix $A \quad ^{\times}$, as implied in (5.36). An alternative is to replace A in this equation by the tridiagonal matrix $T \quad ^{\times}$ supplied by Householder's method. The calculation is then very much quicker, but produces the eigenvector of T; to obtain the corresponding eigenvector of A we must then apply to this vector the sequence of Householder transformations which were used in the original reduction to tridiagonal form. It is easy to show that this is the most efficient method.

1 9 1 1 ,4 4* 4 .
, C0 J K #@. , &4?4# F404
F 494 (: ' % K 0 "
5 6 #@@?4

Inverse iteration with the original matrix A $\times$ requires the LU decomposition of A, followed by one or more forward and backsubstitution operations. As we saw in Section 2.6, the LU decomposition requires approximately $-n$ multiplications. The same process with the tridiagonal matrix T, using the Thomas algorithm, involves only a small multiple of n multiplications.

Having found an eigenvector of the tridiagonal matrix T $\times$, so that

$$T = \lambda ,$$

we use the fact that $Q\ AQ = T$ to write

$$AQ = \lambda Q ,$$

so that the vector Q is an eigenvector of A. Using Theorem 5.7, this means that the required eigenvector of A is

$$H - \ldots H ,$$

where the matrices H $\times$, $j = 2, \ldots, n$ 1, are Householder matrices. To multiply a vector by a Householder matrix $H = H( )$ we write

$$H = (I \alpha) = \alpha( ) .$$

Assuming that $\alpha = 2/( )$ is known, this requires the calculation of the scalar product , and then subtracting a multiple of the vector from the vector . This evidently involves $2n$ multiplications. Hence the calculation of Q requires only $2n(n 2)$ multiplications, and the work involved in the whole process is proportional to n, instead of n. In fact the total is less than $2n(n 2)$, since a more careful count can use the fact that many of the elements in the vector are known to be zero.

Example 5.10 *Returning to the tridiagonal matrix (5.31), the QR algorithm has given an accurate eigenvalue which is* 8.094 *to three decimal digits. Beginning the inverse iteration (5.36) with a randomly chosen vector* $\square$*, we find that*

$$= (0.0249, 0.0574, 0.3164, 0.4256, 0.8455) .$$

Successive iterations make no change in this vector, as might be expected, since the eigenvalue used was accurate to within rounding error.

This is therefore the eigenvector of the tridiagonal matrix (5.31), to

eigenvectors , $j = 1, 2, \ldots, n$, *as*

$$= \quad \alpha \quad , \tag{5.41}$$

then

$$R(\ ) = \frac{\lambda \ \alpha}{\alpha} . \tag{5.42}$$

On noting that is equal to 1 when $i = j$ and to 0 otherwise, (5.42) follows trivially by inserting (5.41) into (5.40).

Theorem 5.12 *Let* $A \quad ^{\times}$. *For any vector* ,

$$\lambda \quad R(\ ) \quad \lambda \quad , \tag{5.43}$$

where λ *and* λ *are respectively the least and greatest of the eigenvalues of* A. *These bounds are attained when is the corresponding eigenvector.*

Proof The inequalities follow immediately from (5.42) by noting that $\lambda \quad \lambda \quad \lambda$, $j = 1, 2, \ldots, n$. □

Theorem 5.13 *Suppose that is a normalised vector, that is,* $= 1$. *Assume, further, that is the* k*th normalised eigenvector of* $A \quad ^{\times}$, *and that*

$$= \quad (\varepsilon)$$

for a small ε . *Then,*

$$R(\ ) = \lambda \ + \quad (\varepsilon \ ) .$$

Proof It follows from (5.41) that $= \alpha$, and therefore,

$$\begin{aligned} &= \quad (\qquad) \ (\qquad) \\ &= \qquad 2 \qquad + \\ &= \quad 2(1 \quad \alpha \ ). \end{aligned}$$

Hence, $\alpha \ = 1 + \quad (\varepsilon \ )$. Further,

$$1 = \qquad = \qquad \alpha$$

$$= \alpha + \alpha$$
$$= 1 + (\varepsilon) + \alpha .$$

Consequently, $\alpha = (\varepsilon)$ for all $j = k$. The result then follows from (5.42) which (with $\alpha = = 1$) yields that

$$R( ) = \lambda \alpha + \lambda \alpha$$
$$= \lambda + (\varepsilon).$$

□

This important result means that if we have a fairly close approximation to an eigenvector of A, then the Rayleigh quotient $R( )$ gives very easily a much more accurate approximation to the corresponding eigenvalue.

5.10 Perturbation analysis

It is often necessary to have an estimate of how much the eigenvalues and eigenvectors of a matrix are affected by changes in the elements. Such perturbations may arise, for example, when the matrix elements are obtained by physical measurements which are inexact, or they might result from finite difference approximations of a differential equation, as will be seen in Chapter 13. The last two theorems in this chapter address some of these questions. We begin with the following preliminary result.

Theorem 5.14 *Let* M $^{\times}$ *, with eigenvalues* λ *and corresponding orthonormal eigenvectors* $, i = 1, 2, \ldots, n$*, and suppose that* $= \mathbf{0}$ *and* *are vectors in* *and* μ *is a real number such that*

$$(M \quad \mu I) = . \tag{5.44}$$

Then, at least one eigenvalue λ *of* M *satisfies*

$$\lambda \quad \mu \qquad / \quad .$$

Proof If μ is equal to one of the eigenvalues the proof is trivial, so we shall assume that $\mu = \lambda$, $k = 1, 2, \ldots, n$. We write the vectors and

as linear combinations of the eigenvectors of M, so that

$$= \quad \alpha \quad , \qquad = \quad \beta \quad .$$

Substituting in (5.44), we may equate coefficients of the linearly independent vectors $\quad , k = 1, 2, \ldots, n$, to deduce that

$$(\lambda \quad \mu)\alpha \; = \beta \; , \qquad k = 1, 2, \ldots, n \, .$$

Now suppose that λ is the eigenvalue which is closest to μ; this means that

$$\lambda \quad \mu \quad \lambda \quad \mu \; , \qquad k = 1, 2, \ldots, n \, .$$

Since the eigenvectors $\quad , i = 1, 2, \ldots, n$, are orthonormal in $\quad$, we have

$$\alpha \; = \quad , \qquad \beta \; = \quad .$$

Hence

$$\frac{\beta}{(\lambda \quad \mu)} = \quad ,$$

which gives

$$= \quad \beta \quad (\lambda \quad \mu) \quad \frac{\beta}{(\lambda \quad \mu)} = (\lambda \quad \mu) \quad ,$$

as required. □

We shall now use this result to show that in the case of a symmetric matrix A, small symmetric perturbations of A lead to small changes in the eigenvalues of A.

Theorem 5.15 (Bauer–Fike Theorem (symmetric case)) *Suppose that* $A, E \quad ^{\times}$ *and* $B = A \quad E$. *Assume, further, that the eigenvalues of* A *are denoted by* $\lambda \; , j = 1, 2, \ldots, n$, *and* μ *is an eigenvalue of* B. *Then, at least one eigenvalue* λ *of* A *satisfies*

$$\lambda \quad \mu \quad E \quad .$$

Proof This is a straightforward consequence of the previous theorem. Suppose that $\quad$ is the normalised eigenvector of B corresponding to the eigenvalue μ, so that $B \quad = \mu \quad$. Then,

$$(A \quad \mu I) \; = (B + E \quad \mu I) \; = E \quad .$$

It then follows from Theorem 5.14 that there is an eigenvalue λ of A such that

$$\lambda \quad \mu \qquad E \qquad\qquad E \qquad\qquad = \; E \quad ,$$

as required. □

Example 5.11 *Consider the* 3 3 *Hilbert matrix*

$$A = \begin{matrix} 1 & 1/2 & 1/3 \\ 1/2 & 1/3 & 1/4 \\ 1/3 & 1/4 & 1/5 \end{matrix}$$

and its perturbation

$$B = \begin{matrix} 1.0000 & 0.5000 & 0.3333 \\ 0.5000 & 0.3333 & 0.2500 \\ 0.3333 & 0.2500 & 0.2000 \end{matrix}$$

which results by rounding each entry of A *to four decimal digits.*

In this case, $E = A \quad B$ and $E \quad = 3.3 \quad 10^{-} \; .$ Let μ be an eigenvalue of B; then, according Theorem 5.15, at least one of the eigenvalues λ , λ , λ of the matrix A satisfies the inequality

$$\lambda \quad \mu \quad 3.3 \quad 10^{-} \; . \tag{5.45}$$

Indeed, the true eigenvalues of A and B are, respectively,

$$\lambda \; = 0.002687338072 \, , \quad \lambda \; = 0.1223270673 \, , \quad \lambda \; = 1.408318925 \, ,$$

and

$$\mu \; = 0.002664493933 \, , \quad \mu \; = 0.1223414532 \, , \quad \mu \; = 1.408294053 \, .$$

Therefore,

$$\lambda \quad \mu \; = 2.29 \quad 10^{-} \, , \quad \lambda \quad \mu \; = \; 1.44 \quad 10^{-} \, , \quad \lambda \quad \mu \; = 2.49 \quad 10^{-} \, ,$$

which is in agreement with (5.45).

5.11 Notes

Theorem 5.15 is a special case of the following general result, known as the Bauer–Fike Theorem.

[1] 94:4 4 4 9 " 8, (' #*&C
#@? 4

Theorem 5.16 *Assume that* $A \in \mathbb{C}^{n\times n}$ *is diagonalisable; i.e., there exists a nonsingular matrix* $X \in \mathbb{C}^{n\times n}$ *such that* $X^{-1}AX = \Lambda$, *where* Λ *is a diagonal matrix whose diagonal entries* λ_j, $j = 1, \ldots, n$, *are the eigenvalues of* A. *Suppose further that* $E \in \mathbb{C}^{n\times n}$, $B = A + E$, *and* μ *is an eigenvalue of* B. *Then, at least one eigenvalue* λ *of* A *satisfies*

$$|\lambda - \mu| \leq \kappa(X)\,\|E\|,$$

where $\kappa(X) = \|X\|\,\|X^{-1}\|$ *is the condition number of the matrix* X *in the matrix 2-norm on* $\mathbb{C}^{n\times n}$.

In the special case when $A, E \in \mathbb{C}^{n\times n}$, the matrix X can be chosen to be orthogonal; *i.e.*, $X^{-1} = X^{*}$. Therefore, $\|X\| = \|X^{-1}\| = 1$, and hence $\kappa(X) = 1$, in accordance with the inequality stated in Theorem 5.15. Theorems 5.15 and 5.16 estimate how far the eigenvalues of A are perturbed by changes in the elements of A. The question as to how large the changes in the eigenvectors may be is more difficult; it is discussed in detail in

- , *The Algebraic Eigenvalue Problem,* Clarendon Press, Oxford University Press, New York, 1988.

Chapter 8 of Wilkinson's book outlines the convergence proof of the QR iteration, while the convergence of Jacobi's method is covered in Chapter 5 of that book. For further details, see also Chapter 9 of

! $, *The Symmetric Eigenvalue Problem*, Prentice–Hall, Englewood Cliffs, NJ, 1980.

Exercises

5.1 Give a proof of Lemma 5.3.

5.2 Use Householder matrices to transform the matrix

$$A = \begin{matrix} 2 & 1 & 2 & 2 \\ 1 & 7 & 6 & 5 \\ 2 & 6 & 2 & 5 \\ 2 & 5 & 5 & 1 \end{matrix}$$

to tridiagonal form.

5.3 Use Sturm sequences to show that no eigenvalue of the matrix

$$A = \begin{matrix} 3 & 1 & 0 & 0 \\ 1 & 2 & 2 & 0 \\ 0 & 2 & 4 & \alpha \\ 0 & 0 & \alpha & 1 \end{matrix}$$

lies in the interval $(0, 1)$ if $5\alpha > 8$, and that exactly one eigenvalue of A lies in this interval if $5\alpha < 8$.

5.4 Given any two nonzero vectors and in , construct a Householder matrix H such that H is a scalar multiple of ; note that if $H = c$, then $c = /$. Is the matrix unique?

5.5 Suppose that the matrix $D \times$ is diagonal with distinct diagonal elements d , ..., d . Let $A \times$, with a 1 for all i, j $1, 2, \ldots, n$, and assume that ε is so small that ε can be neglected, and that the matrix $D + \varepsilon A$ has eigenvalue $\lambda + \varepsilon\mu$ and corresponding eigenvector $\mathbf{e} + \varepsilon\mathbf{u}$. Show that $\lambda = d$ for some j $1, 2, \ldots, n$ and that $\mu = a$. Write down the elements of $\mathbf{e}$, and show that

$$u = \frac{a}{d \quad \lambda}, \qquad i = j .$$

Explain why the requirement that eigenvectors should be normalised implies that $u = 0$.

5.6 With the same notation as in Exercise 5, suppose now that $d = d = = d$, that d , d , ..., d are distinct, and that ε can be neglected. Writing the matrices and the eigenvector in partitioned form, so that

$$\begin{matrix} d\ I + \varepsilon A & \varepsilon A & \mathbf{e} + \varepsilon & + \varepsilon \\ \varepsilon A & D - + \varepsilon A & \mathbf{f} + \varepsilon & + \varepsilon \end{matrix}$$

$$= (\lambda + \varepsilon\mu + \varepsilon\ \nu) \begin{matrix} \mathbf{e} + \varepsilon & + \varepsilon \\ \mathbf{f} + \varepsilon & + \varepsilon \end{matrix} ,$$

show that $\lambda = d$, $\mathbf{f} = \mathbf{0}$, and that μ is an eigenvalue of A with corresponding eigenvector $\mathbf{e}$. Show how is obtained from the solution of $(D - \lambda I) = A\ \mathbf{e}$, and that

$$(A \quad \mu) = \nu\mathbf{e} \quad A\ \mathbf{v} .$$

Explain how the vector can be obtained in terms of the eigenvectors and eigenvalues of the matrix A, assuming that these eigenvalues are distinct.

5.7 Suppose that A is tridiagonal, that $A - \mu I = QR$ and $B = RQ + \mu I$, where μ, Q is a product of plane rotations and R is upper triangular and tridiagonal. Show that B can be written as an orthogonal transformation of A, and that B is symmetric. Show also that the only nonzero elements in the matrix B which are below the diagonal lie immediately below the diagonal; deduce that B is tridiagonal.

5.8 Perform one step of the QR algorithm, using the shift $\mu = a$, for the matrix

$$A = \begin{matrix} 0 & 1 \\ 1 & 0 \end{matrix} .$$

Show that the QR algorithm does not converge for this matrix. (This is a special case in which a different shift must be used.)

5.9 Perform one step of the QR algorithm, using the shift $\mu = a$, for the matrix

$$A = \begin{matrix} 13 & 4 \\ 4 & 10 \end{matrix} .$$

5.10 Carry out two steps of inverse iteration for the matrix

$$A = \begin{matrix} 2 & 2 \\ 2 & 5 \end{matrix} ,$$

using the eigenvalue estimate $\vartheta = 5$ and the initial vector

$$= \begin{matrix} 1 \\ 1 \end{matrix} .$$

Verify that the elements of the vector agree with those of the true eigenvector with an accuracy of about 5%. Evaluate the Rayleigh quotient using the vector , and verify that the result agrees with the true eigenvalue to about 1 in 3000.

5.11 An eigenvalue and eigenvector of the matrix A may be evaluated by solving the system of nonlinear equations

$$\begin{aligned} (A \quad \lambda I) \quad &= \quad 0 , \\ &= \quad 1 \end{aligned}$$

for the unknowns λ and . Using Newton's method, starting

from estimates λ and , show that the next iteration is determined by

$$\begin{aligned} A \quad \delta\lambda \quad &= \quad (A \quad \lambda \quad I) \quad , \\ &= \; -(\qquad\qquad 1) \end{aligned}$$

and $= \quad + \quad , \lambda \quad = \lambda \quad + \delta\lambda$. Comment on the difference between this method and the method of inverse iteration in Section 5.8.

5.12 Suppose that $A \quad ^{\times}$ and that Jacobi's method has produced an orthogonal matrix R and a symmetric matrix B such that $B = R \;\; AR$. Suppose also that $b \quad < \varepsilon$ for all $i = j$. Show that, for each $j = 1, 2, \ldots, n$, there is at least one eigenvalue λ of A such that

$$\lambda \quad b \quad < \varepsilon \quad \overline{n}.$$

5.13 Suppose that $A \quad ^{\times}$ and that the Householder reduction and QR algorithm have produced an orthogonal matrix Q and a tridiagonal matrix T such that $T = Q \;\; AQ$. Suppose also that $t \quad - \quad < \varepsilon$. Show that there is at least one eigenvalue λ of A such that

$$\lambda \quad t \quad < \varepsilon.$$

6

Polynomial interpolation

6.1 Introduction

It is time to take a break from solving equations. In this chapter we consider the problem of polynomial interpolation; it involves finding a polynomial that agrees exactly with some information that we have about a real-valued function f of a single real variable x. This information may be in the form of values f(x), . . . , f(x) of the function f at some finite set of points x , . . . , x on the real line, and the corresponding polynomial is then called the **Lagrange interpolation polynomial** or, provided that f is differentiable, it may include values of the derivative of f at these points, in which case the associated polynomial is referred to as a **Hermite interpolation polynomial**.

Why should we be interested in constructing Lagrange or Hermite interpolation polynomials? If the function values f(x) are known for all x in a closed interval of the real line, then the aim of polynomial

interpolation is to approximate the function f by a polynomial over this interval. Given that any polynomial can be completely specified by its (finitely many) coefficients, storing the interpolation polynomial for f in a computer will be, generally, more economical than storing f itself.

Frequently, it is the case, though, that the function values $f(x)$ are only known at a finite set of points $x, \ldots, x$, perhaps as the results of some measurements. The aim of polynomial interpolation is then to attempt to reconstruct the unknown function f by seeking a polynomial p whose graph in the (x, y)-plane passes through the points with coordinates $(x, f(x))$, $i = 0, \ldots, n$. Of course, in general, the resulting polynomial p will differ from f (unless f itself is a polynomial of the same degree as p), so an error will be incurred. In this chapter we shall also establish results which provide bounds on the size of this error.

6.2 Lagrange interpolation

Given that n is a nonnegative integer, let $\&$ denote the set of all (real-valued) polynomials of degree n defined over the set of real numbers. The simplest interpolation problem can be stated as follows: given x and y in , find a polynomial $p \quad \&$ such that $p(x) = y$. The solution to this is, trivially, $p(x) \quad y$. The purpose of this section is to explore the following more general problem.

> Let $n \quad 1$, and suppose that $x, i = 0, 1, \ldots, n$, are *distinct* real numbers (*i.e.*, $x = x$ for $i = j$) and $y, i = 0, 1, \ldots, n$, are real numbers; we wish to find $p \quad \&$ such that $p(x) = y, i = 0, 1, \ldots, n$.

To prove that this problem has a unique solution, we begin with a useful lemma.

Lemma 6.1 *Suppose that* $n \quad 1$. *There exist polynomials* $L \quad \&$, $k = 0, 1, \ldots, n$, *such that*

$$L(x) = \begin{cases} 1, & i = k, \\ 0, & i = k, \end{cases} \tag{6.1}$$

for all $i, k = 0, 1, \ldots, n$. *Moreover,*

$$p(x) = \quad L(x)y \tag{6.2}$$

satisfies the above interpolation conditions; in other words, $p \quad \&$ *and* $p(x) = y, \ i = 0, 1, \ldots, n$.

Proof For each fixed k, $0 \quad k \quad n$, L is required to have n zeros $- x$, $i = 0, 1, \ldots, n$, $i = k$; thus, $L\ (x)$ is of the form

$$L\ (x) = C \overset{*}{\quad} (x \quad x\ ), \tag{6.3}$$

□

where C

Proof In view of Remark 6.1, for $n = 0$ the proof is trivial. Let us therefore suppose that $n \quad 1$. It follows immediately from Lemma 6.1 that the polynomial $p \quad \&$ defined by

$$p\ (x) = \quad L\ (x)y$$

satisfies the conditions (6.5), thus showing the *existence* of the required polynomial. It remains to show that p is the *unique* polynomial in $\&$ satisfying the interpolation property

$$p\ (x\ ) = y\ , \qquad i = 0, 1, \ldots, n\,.$$

Suppose, otherwise, that there exists $q \quad \&$, different from p , such that $q\ (x\ ) = y\ , \quad i = 0, 1, \ldots, n$. Then, $p \quad q \quad \&$ and $p \quad q$ has $n + 1$ distinct roots, $x\ , i = 0, 1, \ldots, n$; since a polynomial of degree n cannot have more than n distinct roots, unless it is identically 0, it follows that

$$p\ (x) \quad q\ (x) \quad 0\,,$$

which contradicts our assumption that p and q are distinct. Hence, there exists only one polynomial $p \quad \&$ which satisfies (6.5). □

Definition 6.1 *Suppose that* $n \quad 0$. *Let* $x\ , i = 0, \ldots, n$, *be distinct real numbers, and* $y\ , i = 0, \ldots, n$, *real numbers. The polynomial* p *defined by*

$$p\ (x) = \quad L\ (x)y\ , \tag{6.6}$$

with $L\ (x), k = 0, 1, \ldots, n$, *defined by (6.4) when* $n \quad 1$, *and* $L\ (x) \quad 1$ *when* $n = 0$, *is called the* **Lagrange interpolation polynomial** *of degree* n *for the set of points* $(x\ , y\ ): i = 0, \ldots, n$. *The numbers* $x\ ,$ $i = 0, \ldots, n$, *are called the* **interpolation points**.

Frequently, the real numbers y are given as the values of a real-valued function f, defined on a closed real interval $[a, b]$, at the (distinct) interpolation points $x \quad [a, b]$, $i = 0, \ldots, n$.

Definition 6.2 *Let* $n \quad 0$. *Given the real-valued function* f, *defined and continuous on a closed real interval* $[a, b]$, *and the (distinct) interpolation points* $x \quad [a, b]$, $i = 0, \ldots, n$, *the polynomial* p *defined by*

$$p\ (x) = \qquad L\ (x) f(x\ ) \tag{6.7}$$

is the **Lagrange interpolation polynomial of degree n (with interpolation points x , i = 0, . . . , n) for the function f .**

Example 6.1 *We shall construct the Lagrange interpolation polynomial of degree* 2 *for the function* f : x e *on the interval* [1, 1]*, with interpolation points* x = 1, x = 0, x = 1.

As n = 2, we have that

$$L\ (x) = \frac{(x \quad x\ )(x \quad x\ )}{(x \quad x\ )(x \quad x\ )} = -x(x \quad 1)\,.$$

Similarly, L (x) = 1 x and L (x) = −x(x + 1). Therefore,

$$p\ (x) = -x(x \quad 1)\,e^{-} + (1 \quad x\ )\,e\ + -x(x+1)\,e\ .$$

Thus, after some simplification, p (x) = 1 + x sinh 1 + x (cosh 1 1).

Although the values of the function f and those of its Lagrange interpolation polynomial coincide at the interpolation points, f (x) may be quite different from p (x) when x is *not* an interpolation point. Thus, it is natural to ask just how large the difference f (x) p (x) is when x = x , i = 0, . . . , n. Assuming that the function f is sufficiently smooth, an estimate of the size of the **interpolation error** f (x) p (x) is given in the next theorem.

Theorem 6.2 *Suppose that* n 0*, and that* f *is a real-valued function, defined and continuous on the closed real interval* [a, b]*, such that the derivative of* f *of order* n + 1 *exists and is continuous on* [a, b]*. Then, given that* x [a, b]*, there exists* ξ = ξ(x) *in* (a, b) *such that*

$$f(x) \quad p\ (x) = \frac{f \qquad (\xi)}{(n+1)!}\pi \qquad (x)\,, \tag{6.8}$$

where

$$\pi \qquad (x) = (x \quad x\ )\ldots(x \quad x\ )\,. \tag{6.9}$$

Moreover

$$f(x) \quad p\ (x) \qquad \frac{M}{(n+1)!}\,\pi \qquad (x)\ , \tag{6.10}$$

where

$$M = \max_{(\square} f$$

It is perhaps worth noting that since the location of ξ in the interval $[a, b]$ is unknown (to the extent that the exact dependence of ξ on x is not revealed by the proof of Theorem 6.2), (6.8) is of little practical value; on the other hand, given the function f, an upper bound on the maximum value of f over $[a, b]$ is, at least in principle, possible to obtain, and thereby we can provide an upper bound on the size of the interpolation error by means of inequality (6.10).

6.3 Convergence

An important theoretical question is whether or not a sequence (p) of interpolation polynomials for a continuous function f converges to f as n . This question needs to be made more specific, as p depends on the distribution of the interpolation points x , $j = 0, 1, \ldots, n$, not just on the value of n. Suppose, for example, that we agree to choose equally spaced points, with

$$x = a + j(b - a)/n, \qquad j = 0, 1, \ldots, n, \quad n \quad 1.$$

The question of convergence then clearly depends on the behaviour of M as n increases. In particular, if

$$\lim_{\to \infty} \frac{M}{(n+1)!} \max_{\square} \pi \quad (x) = 0,$$

then, by (6.10),

$$\lim_{\to \infty} \max_{\square} f(x) \quad p(x) = 0, \tag{6.12}$$

and we say that the sequence of interpolation polynomials (p), with equally spaced points on $[a, b]$, converges to f as n , uniformly on the interval $[a, b]$.

You may now think that if all derivatives of f exist and are continuous on $[a, b]$, then (6.12) will hold. Unfortunately, this is not so, since the sequence

$$M \quad \max_{\square} \pi \quad (x)$$

may tend to , as n , faster than the sequence $(1/(n+1)!)$ tends to 0.

In order to convince you of the existence of such 'pathological' functions, we consider the sequence of Lagrange interpolation polynomials

Table 6.1. *Runge phenomenon:* n *denotes the degree of the interpolation polynomial* p *to* f*, with equally spaced points on* [5, 5]. 'Max error' *signifies* max □ – f(x) p (x) .

E	F !
	%(
'	"
%	%
	'
	$
	& %%
'	) (
%	' (
	)'
	(($
'	( )

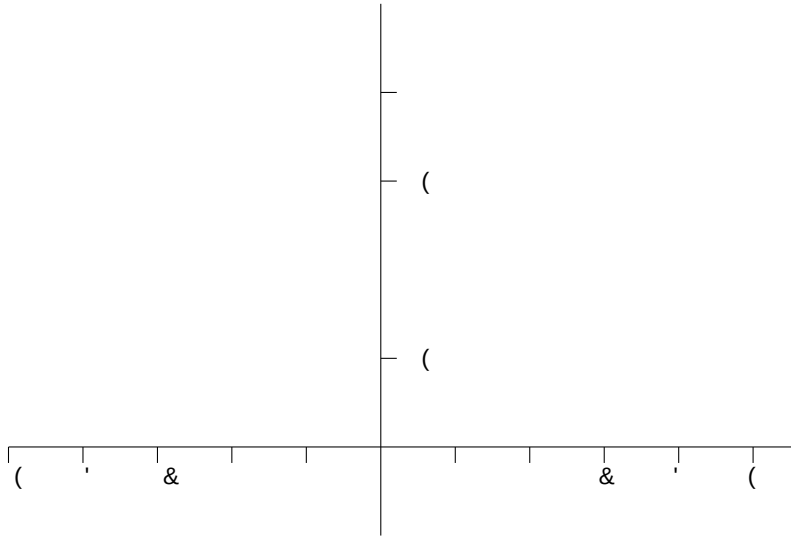
p , n = 0, 1, 2, . . ., with equally spaced interpolation points on the interval [5, 5], to

$$f(x) = \frac{1}{1+x}, \qquad x \quad [\ -5, 5].$$

This example is due to Runge, and the characteristic behaviour exhibited by the sequence of interpolation polynomials p in Table 6.1 is referred to as the **Runge phenomenon**: Table 6.1 shows the maximum difference between f(x) and p (x) for 5 x 5, for values of n from 2 up to 24. The numbers indicate clearly that the maximum error increases exponentially as n increases. Figure 6.1 shows the interpolation polynomial p , using the equally spaced interpolation points x = 5 + j, j = 0, 1, . . . , 10. The sizes of the local maxima near 5 grow exponentially as the degree n increases.

Note that, in many ways, the function f is well behaved; all its deriva-

1 / 5 - B* % #.>? F 6 C * K 6 #@ &
FI F 6D , 6 , 5 6 1 ' $
, 4 0 , #.. 1 J 64
F 6 , 1 , 2 5 $
, , 1 1 , G
8 E 1 1 , 1 , S, #..&
" 5 1 , 1 4) #@
, 1 1 % ' , FI 4 0 E
, 5 2 & 6 , 6 1
4 % 1 J 1 ,"4



The construction is similar to that of the Lagrange interpolation polynomial, but now requires two sets of polynomials H and K with $k = 0, \ldots, n$; these will be defined in the proof of the next theorem.

Theorem 6.3 (Hermite Interpolation Theorem) *Let* $n \quad 0$*, and suppose that* x , $i = 0, \ldots, n$*, are* distinct *real numbers. Then, given two sets of real numbers* y , $i = 0, \ldots, n$*, and* z , $i = 0, \ldots, n$*, there is a unique polynomial* $p \quad$ *in* $\& \quad$ *such that*

$$p \quad (x) = y , \qquad p' \quad (x) = z , \qquad i = 0, \ldots, n . \tag{6.13}$$

Proof Let us begin by supposing that $n \quad 1$. As in the case of Lagrange interpolation, we start by constructing a set of auxiliary polynomials; we consider the polynomials $H \quad$ and $K \quad , k = 0, 1, \ldots, n$, defined by

$$\begin{aligned} H \ (x) &= [L \ (x)] \ (1 \quad 2L' (x \ )(x \quad x \ )) , \\ K \ (x) &= [L \ (x)] \ (x \quad x \ ) , \end{aligned} \tag{6.14}$$

where

$$L \ (x) = \prod^{*} \frac{x \quad x}{x \quad x} .$$

Clearly $H \quad$ and $K \quad , k = 0, 1, \ldots, n$, are polynomials of degree $2n + 1$. It is easy to see that $H \ (x \ ) = K \ (x \ ) = 0$, $H' (x \ ) = K' (x \ ) = 0$ whenever $i, k \quad 0, 1, \ldots, n$ and $i = k$; moreover, a straightforward calculation verifies their values when $i = k$, showing that

$$H \ (x \ ) = \begin{matrix} 1 , & i = k , \\ 0 , & i = k , \end{matrix} \qquad H' (x \ ) = 0 , \qquad i, k = 0, 1, \ldots, n ,$$

$$K \ (x \ ) = 0 , \qquad K' (x \ ) = \begin{matrix} 1 , & i = k , \\ 0 , & i = k , \end{matrix} \qquad i, k = 0, 1, \ldots, n .$$

We deduce that

$$p \quad (x) = \quad [H \ (x) y \ + K \ (x) z \ ]$$

satisfies the conditions (6.13), and $p \quad$ is clearly an element of $\& \quad$.

To show that this is the only polynomial in $\& \quad$ satisfying these conditions, we suppose otherwise; then, there exists a polynomial q in $\& \quad$, distinct from $p \quad$, such that

$$q \quad (x \ ) = y \quad \text{and} \quad q' \quad (x \ ) = z , \qquad i = 0, 1, \ldots, n .$$

Consequently, p q has $n+1$ distinct zeros; therefore, Rolle's Theorem implies that, in addition to the $n+1$ zeros x, $i = 0, 1, \ldots, n$, p′ q′ vanishes at another n points which interlace the x. Hence p′ q′ & has $2n+1$ zeros, which means that p′ q′ is identically zero, so that p q is a constant function. However, (p q)(x) = 0 for $i = 0, 1, \ldots, n$, and hence p q 0, contradicting the hypothesis that p and q are distinct. Thus, p is unique.

When $n = 0$, we define H (x) 1 and K (x) x x, which correspond to taking L (x) 1 in (6.15). Clearly, p defined by

$$p\ (x) = H\ (x)y\ + K\ (x)z\ = y\ + (x\quad x\ )z$$

is the unique polynomial in & such that p (x) = y and p′ (x) = z. □

Definition 6.3 *Let* n 0*, and suppose that* x, $i = 0, \ldots, n$*, are distinct real numbers and* y, z, $i = 0, \ldots, n$*, are real numbers. The polynomial* p *defined by*

$$p\quad (x) = \quad [H\ (x)y\ + K\ (x)z\ ] \tag{6.15}$$

where H (x) *and* K (x) *are defined by (6.15), is called the* **Hermite interpolation polynomial** *of degree* $2n+1$ *for the set of values given in* (x, y, z): $i = 0, \ldots, n$.

Example 6.2 *We shall construct a cubic polynomial* p *such that*

$$p\ (0) = 0, \quad p\ (1) = 1, \quad p'\ (0) = 1 \quad \text{and} \quad p'\ (1) = 0.$$

Here $n = 1$, and since p (0) = p′ (1) = 0 the polynomial simplifies to

$$p\ (x) = H\ (x) + K\ (x).$$

We easily find that, with $n = 1$, x = 0 and x = 1,

$$L\ (x) = 1\quad x, \quad L\ (x) = x,$$

and then,

$$\begin{aligned} H\ (x) &= [L\ (x)]\ (1\quad 2L'\ (x\ )(x\quad x\ )) = x\ (3\quad 2x), \\ K\ (x) &= [L\ (x)]\ (x\quad x\ ) = (1\quad x)\ x. \end{aligned}$$

These yield the required Hermite interpolation polynomial,

$$p\ (x) = \quad x \ + x \ + x \, .$$

Definition 6.4 *Suppose that* f *is a real-valued function, defined on the closed interval* $[a, b]$ *of , and that* f *is continuous and differentiable on this interval. Suppose, further, that* $n \quad 0$ *and that* x , $i = 0, \ldots, n$, *are distinct points in* $[a, b]$. *Then, the polynomial* p *defined by*

$$p \quad (x) = \quad [H \ (x) f\,(x \ ) + K \ (x) f^{\square}(x \ )] \tag{6.16}$$

is the **Hermite interpolation polynomial of degree** $2n + 1$ **with interpolation points** x , $i = 0, \ldots, n$, **for** f. *It satisfies the conditions*

$$p \quad (x \ ) = f\,(x \ ), \quad p^{\square} \quad (x \ ) = f^{\square}(x \ ), \quad i = 0, \ldots, n \, .$$

Pictorially, the graph of p touches the graph of the function f at the points x , $i = 0, \ldots, n$.

To conclude this section we state a result, analogous to Theorem 6.2, concerning the error in Hermite interpolation.

Theorem 6.4 *Suppose that* $n \quad 0$ *and let* f *be a real-valued function, defined, continuous and* $2n + 2$ *times differentiable on the interval* $[a, b]$, *such that* f *is continuous on* $[a, b]$. *Further, let* p *denote the Hermite interpolation polynomial of* f *defined by (6.16). Then, for each* $x \quad [a, b]$ *there exists* $\xi = \xi(x)$ *in* (a, b) *such that*

$$f\,(x) \quad p \quad (x) = \frac{f \quad (\xi)}{(2n + 2)!}[\pi \quad (x)] \ , \tag{6.17}$$

where π *is as defined in (6.9). Moreover,*

$$f\,(x) \quad p \quad (x) \quad \frac{M}{(2n + 2)!}[\pi \quad (x)] \ , \tag{6.18}$$

where $M \quad = \max_{(\square} \quad f \quad (\zeta) \, .$

Proof The inequality (6.18) is a straightforward consequence of (6.17). In order to prove (6.17), we observe that it is trivially true if $x = x$

for some i, $i = 0, \ldots, n$; thus, it suffices to consider $x \quad [a, b]$ such that $x = x$, $i = 0, \ldots, n$. For such x, let us define the function

x_i, $i = 0, 1, \ldots, n$, *are distinct points in* $[a, b]$, *and that* p & *is the Lagrange interpolation polynomial for* f *defined by these points. Then, there exist distinct points* η_i, $i = 1, \ldots, n$, *in* (a, b), *and corresponding to each* x *in* $[a, b]$ *there exists a point* $\xi = \xi(x)$ *in* (a, b), *such that*

$$f^{(}(x) \quad p^{(}(x) = \frac{f \quad (\xi)}{n!}\pi^{(}(x), \tag{6.20}$$

where

$$\pi^{(}(x) = (x \quad \eta_i) \ldots (x \quad \eta_n).$$

Proof Since $f(x_i) \quad p(x_i) = 0$, $i = 0, 1, \ldots, n$, there exists a point η_i in (x_{i-1}, x_i) at which $f^{(}(\eta_i) \quad p^{(}(\eta_i) = 0$, for each $i = 1, \ldots, n$. This defines the points η_i, $i = 1, \ldots, n$. Now the proof closely follows that of Theorem 6.2.

When $x = \eta_i$ for some $i \quad 1, \ldots, n$, both sides of (6.20) are zero. Suppose then that x is distinct from all the η_i, $i = 1, \ldots, n$, and define the function $t \quad \chi(t)$ by

$$\chi(t) = f^{(}(t) \quad p^{(}(t) \quad \frac{f^{(}(x) \quad p^{(}(x)}{\pi^{(}(x)}\pi^{(}(t).$$

This function vanishes at every point η_i, $i = 1, \ldots, n$, and also at the point $t = x$. By successively applying Rolle's Theorem we deduce that χ vanishes at some point ξ. The result then follows as in the proof of Theorem 6.2. □

Corollary 6.1 *Under the conditions of Theorem 6.5,*

$$f^{(}(x) \quad p^{(}(x) \quad \frac{M}{n!}\pi^{(}(x) \quad \frac{(b \quad a)^{} \quad M}{n!}$$

for all x *in* $[a, b]$, *where* $M \quad = \max_{(} \quad f \quad (x)$.

In particular, we deduce that if f and all its derivatives are defined and continuous on the closed interval $[a, b]$, and

$$\lim_{\to \infty} \frac{(b \quad a)^{} \quad M}{n!} = 0,$$

then $\lim_{\to \infty} \max_{(} \quad f^{(}(x) \quad p^{(}(x) = 0$, showing the convergence of the sequence of interpolation polynomials $(p^{(})$ to $f^{(}$, uniformly on $[a, b]$.

The discussion in the last few paragraphs may give the impression that numerical differentiation is a straightforward procedure. In practice, however, things are much more complicated since the function values $f(x_i)$, $i = 0, 1, \ldots, n$, will be polluted by rounding errors.

Example 6.3

and $x = h$, tends to 0; in the latter case, the degree of the polynomial p_h tends to infinity and consequently the spacing between the increasing number of consecutive interpolation points shrinks. Nevertheless, Example 6.3 illustrates the issue that caution should be exercised in the course of numerical differentiation when rounding errors are present.

6.6 Notes

The interpolation polynomial (6.6) was discovered by Edward Waring (1736–1798) in 1776, rediscovered by Euler in 1783 and published by Joseph-Louis Lagrange (1736–1813) in his *Leçons élémentaires sur les mathématiques*, Paris, 1795.

Lagrange's interpolation theorem is a purely algebraic result, and it also holds in number fields different from the field of real numbers considered in this chapter. In particular, it holds if the numbers x_i and y_i, $i = 0, 1, \ldots, n$, are complex, and the polynomial p has complex coefficients. Theorem 6.2 is due to Augustin-Louis Cauchy (1789–1857). The interpolation polynomial (6.15) was discovered by Charles Hermite (1822–1901).

Before modern computers came into general use about 1960, the evaluation of a standard mathematical function for a given value of x required the use of published tables of the function, in book form. If x was not one of the tabulated values, the required result was obtained by interpolation, using tabulated values close to x. The tabulated values were given at equally spaced points, so that usually $x = jh$, where h is a fixed increment. In this case the Lagrange formula can be simplified; as this sort of interpolation had to be done frequently, various devices were used to make the calculations easy and quick. Older books, such as F.B. Hildebrand's *Introduction to Numerical Analysis*, published in 1956, contain extensive discussions of such special methods of interpolation, some of which date back to the time of Newton, but are now mainly of historical interest. A notable early contribution to the development of mathematical tables is the work of Henry Briggs (1560–1630), Savilian Professor of Geometry and fellow of Merton College in Oxford, entitled *Arithmetica logarithmica*, published in 1624. It contained extensive calculations of the logarithms of thirty thousand numbers to 14 decimal digits; these were the numbers from 1 to 20000 and from 90000 to 100000. It also contained tables of the sin function to 15 decimal digits, and of the tan and sec functions to 10 decimal digits.

Exercises

6.1 Construct the Lagrange interpolation polynomial p of degree 1, for a continuous function f defined on the interval $[-1, 1]$, using the interpolation points $x = -1$, $x = 1$. Show further that if the second derivative of f exists and is continuous on $[0, 1]$, then

$$f(x) - p(x) \leq \frac{M}{2}(1 - x) \leq \frac{M}{2}, \quad x \in [-1, 1],$$

where $M = \max_{-1 \leq x \leq 1} |f''(x)|$. Give an example of a function f, and a point x, for which equality is achieved.

6.2 (i) Write down the Lagrange interpolation polynomial of degree 1 for the function $f: x \to x^2$, using the points $x = 0$, $x = a$. Verify Theorem 6.2 by direct calculation, showing that in this case ξ is unique and has the value $\xi = \frac{1}{2}(x + a)$.
(ii) Repeat the calculation for the function $f: x \to (2x - a)^2$; show that in this case there are two possible values for ξ, and give their values.

6.3 Given the distinct points x_i, $i = 0, 1, \ldots, n+1$, and the points y_i, $i = 0, 1, \ldots, n+1$, let q be the Lagrange polynomial of degree n for the set of points (x_i, y_i): $i = 0, 1, \ldots, n$ and let r be the Lagrange polynomial of degree n for the points (x_i, y_i): $i = 1, 2, \ldots, n+1$. Define

$$p(x) = \frac{(x - x_0)r(x) - (x - x_{n+1})q(x)}{x_{n+1} - x_0}.$$

Show that p is the Lagrange polynomial of degree $n+1$ for the points (x_i, y_i): $i = 0, 1, \ldots, n+1$.

6.4 Let $n \geq 1$. The points x_j are equally spaced in $[-1, 1]$, so that

$$x_j = \frac{2j - n}{n}, \quad j = 0, \ldots, n.$$

With the usual notation

$$\pi_{n+1}(x) = (x - x_0) \ldots (x - x_n),$$

show that

$$\pi_{n+1}(1 + 1/n) = \frac{(2n)!}{2^n n!}.$$

Using Stirling's formula

$$n! \sim \sqrt{2\pi n}\, n^n e^{-n}, \quad n \to \infty,$$

verify that

$$\pi \quad (1 \quad 1/n) \qquad \frac{2 \qquad e^{-}}{n}$$

for large values of n.

6.5 Let $n \quad 1$. Suppose that x , $i = 0, 1, \ldots, n$, are distinct real numbers, and y , u , $i = 0, 1, \ldots, n$, are real numbers. Suppose, further, that there exists $p \qquad \& \qquad$ such that $p \qquad (x) = y$ for all $i = 0, 1, \ldots, n$, and $p^{\square} \quad (x) = u$, $i = 0, 1, \ldots, n$. Attempt to prove that $p \qquad$ is the unique polynomial with these properties, by adapting the uniqueness proofs in Sections 6.2 and 6.4, using Rolle's Theorem; explain where the proof fails. Show that there is no polynomial $p \quad \&$ such that $p \; (\quad 1) = 1$, $p \; (0) = 0$, $p \; (1) = 1$, $p^{\square}(\quad 1) = 0$, $p^{\square}(0) = 0$, $p^{\square}(1) = 0$, but that if the first condition is replaced by $p \; (\quad 1) = \quad 1$, then there is an infinite number of such polynomials. Give an explicit expression for the general form of these polynomials.

6.6 Suppose that $n \quad 1$. The function f and its derivatives of order up to and including $2n + 1$ are continuous on $[a, b]$. The points x , $i = 0, 1, \ldots, n$, are distinct and lie in $[a, b]$. Construct polynomials $l \; (x)$, $h \; (x)$, $k \; (x)$, $i = 1, \ldots, n$, of degree $2n$ such that the polynomial

$$p \quad (x) = l \; (x) f \, (x \;) + \qquad [h \; (x) f \, (x \;) + k \; (x) f^{\square}(x \;)]$$

satisfies the conditions

$$p \quad (x \;) = f \, (x \;), \quad i = 0, 1, \ldots, n,$$

and

$$p^{\square} \quad (x \;) = f^{\square}(x \;), \quad i = 1, \ldots, n.$$

Show also that for each value of x in $[a, b]$ there is a number η, depending on x, such that

$$f \, (x) \quad p \quad (x) = \frac{(x \quad x \;)^{+} \qquad (x \quad x \;)}{(2n + 1)!} f \qquad (\eta).$$

6.7 Suppose that $n \quad 2$. The function f and its derivatives of order up to and including $2n$ are continuous on $[a, b]$. The points x , $i = 0, 1, \ldots, n$, are distinct and lie in $[a, b]$. Explain how to

construct polynomials $l_i(x), l_i(x), h_i(x), k_i(x), i = 1, \ldots, n-1$, of degree $2n-1$ such that the polynomial

$$p_{2n-1}(x) = l_1(x)f(x_1) + l_1(x)f(x_1) + \cdots + [h_i(x)f(x_i) + k_i$$

Find the limit of this expression as $\varepsilon \to 0$, and deduce that $p - q - 0$ as $\varepsilon \to 0$, where $q -$ is the Hermite interpolation polynomial for f, using the points x, $i = 0, \ldots, n \quad 1$.

6.10 Construct the Hermite interpolation polynomial of degree 3 for the function $f: x \quad x$, using the points $x = 0$, $x = a$, and show that it has the form $p(x) = 3a\,x \quad 2a\,x$. Verify Theorem 6.4 by direct calculation, showing that in this case ξ is unique and has the value $\xi = -(x + 2a)$.

6.11 The complex function $z \quad f(z)$ of the complex variable z is holomorphic in the region D of the complex plane; the boundary of D is the simple closed contour C. The interpolation points x, $j = 0, 1, \ldots, n$, with $n \quad 1$, and the point x all lie in D. Determine the residues of the function g defined by

$$g(z) = \frac{f(z)}{z \quad x} \quad \frac{x \quad x}{z \quad x}$$

at its poles in D, and deduce that

$$f(x) \quad p(x) = \frac{1}{2\pi\imath} \int_{C} \frac{f(z)}{z \quad x} \quad \frac{x \quad x}{z \quad x} \, dz,$$

where p is the Lagrange interpolation polynomial for the function f using the interpolation points x, $j = 0, 1, \ldots, n$.

Now, suppose that the real number x and the interpolation points x, $j = 0, 1, \ldots, n$, all lie in the real interval $[a, b]$, and that D consists of all the points z such that $z \quad t < K$ for all $t \quad [a, b]$, where K is a constant with $K > b \quad a$. Show that the length of the contour C is $2(b \quad a) + 2\pi K$, and that

$$f(x) \quad p(x) < \frac{(b \quad a + \pi K)M}{\pi} \quad \frac{b \quad a}{K} \quad ,$$

where M is such that $f(z) \quad M$ on C. Deduce that the sequence $(p \quad)$ converges to f, uniformly on $[a, b]$.

Show that these conditions are not satisfied by the function $f: x \quad 1/(1 + x)$ for x in the interval $[\ 5, 5]$. For what values of a are the conditions satisfied by f for x in the interval $[\ a, a]$?

6.12 With the same notation as in Example 6.3, let

$$E(h) = \frac{(f(h) + \varepsilon \quad) \quad (f(\ h) + \varepsilon_{-})}{2h} \quad f''(0).$$

Suppose that $f'''(x)$ exists and is continuous at all $x \quad [\ h, h]$.

By expanding $f(h)$ and $f(\ \ h)$ into Taylor series about the point 0, show that there exists $\xi \quad (\ \ h, h)$ such that

$$E(h) = \frac{1}{6}h\ \ f^{\square\square\square}(\xi) + \frac{\varepsilon \qquad \varepsilon_{-}}{2h}.$$

Hence deduce that

$$E(h) \qquad \frac{1}{6}h\ \ M\ \ + \frac{\varepsilon}{h}$$

where $M\ \ = \max_{\square\ -**} \quad f^{\square\square\square}(x)$ and $\varepsilon = \max(\ \varepsilon\ \ ,\ \varepsilon_{-}\ )$. Show further that the right-hand side of the last inequality achieves its minimum value when

$$h = \quad \frac{3\varepsilon}{M} \qquad .$$

7

Numerical integration – I

7.1 Introduction

The problem of evaluating definite integrals arises both in mathematics and beyond, in many areas of science and engineering. At some point in our mathematical education we all learned to calculate simple integrals such as

$$\int e \, dx \quad \text{or} \quad \int \cos x \, dx$$

using a table of integrals, so you will know that the values of these are $e - 1$ and 0 respectively; but how about the innocent-looking

$$\int e^{x^2} dx \quad \text{and} \quad \int \cos(x^2) dx,$$

or the more exotic

$$\int \exp(\sin(\cos(\sinh(\cosh(\tan^{-1}(\log(x))))))) \, dx?$$

Please try to evaluate these using a table of integrals and see how far you can get! It is not so simple, is it? Of course, you could argue that the last example was completely artificial. Still, it illustrates the point that it is relatively easy to think of a continuous real-valued function f defined on a closed interval $[a, b]$ of the real line such that the definite integral

$$\int f(x) \, dx \tag{7.1}$$

is very hard to reduce to an entry in the table of integrals by means of the usual tricks of variable substitution and integration by parts. If you have access to the computer package Maple, you may try to type

at the Maple command line. In about the same time as it will take you to correctly type the command at the keyboard, as if by magic, the result 1514.780678 will pop up on the screen. How was this number arrived at?

The purpose of this chapter, and its continuation, Chapter 10, is to answer this question. Specifically, we shall address the problem of evaluating (7.1) approximately, by applying the results of Chapter 6 on polynomial interpolation to derive formulae for numerical integration (also called numerical quadrature rules). We shall also explain how one can estimate the associated approximation error. What does polynomial interpolation have to do with evaluating definite integrals? The answer will be revealed in the next section which is about a class of quadrature formulae bearing the names of two English mathematicians: Newton and Cotes.

7.2 Newton–Cotes formulae

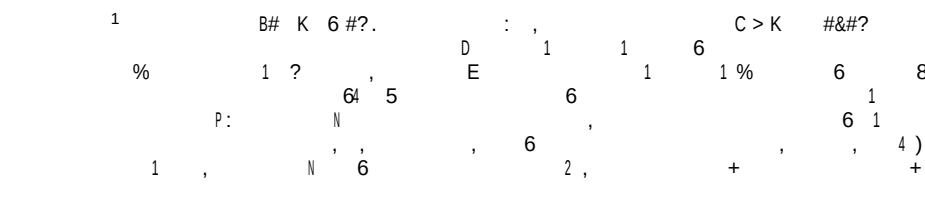
Let f be a real-valued function, defined and continuous on the closed real interval $[a, b]$, and suppose that we have to evaluate the integral

$$\int f(x)dx .$$

Since polynomials are easy to integrate, the idea, roughly speaking, is to approximate the function f by its Lagrange interpolation polynomial p of degree n, and integrate p instead. Thus,

$$\int f(x)dx \approx \int p(x)dx . \tag{7.2}$$

For a positive integer n, let x_i, $i = 0, 1, \ldots, n$, denote the interpolation



points; for the sake of simplicity, we shall assume that these are equally spaced, that is,

$$x_i = a + ih, \qquad i = 0, 1, \ldots, n,$$

where

$$h = (b - a)/n.$$

The Lagrange interpolation polynomial of degree n for the function f, with these interpolation points, is of the form

$$p(x) = \sum_{k=0}^{n} L_k(x) f(x_k) \quad \text{where} \quad L_k(x) = \prod_{\substack{j=0 \\ j \neq k}}^{n} \frac{x - x_j}{x_k - x_j}.$$

Inserting the expression for p into the right-hand side of (7.2) yields

$$\int_a^b f(x)\,dx \approx \sum_{k=0}^{n} w_k f(x_k), \tag{7.3}$$

where

$$w_k = \int_a^b L_k(x)\,dx, \qquad k = 0, 1, \ldots, n. \tag{7.4}$$

The values w_k, $k = 0, 1, \ldots, n$, are referred to as the **quadrature weights**, while the interpolation points x_k, $k = 0, 1, \ldots, n$, are called the **quadrature points**. The numerical quadrature rule (7.3), with quadrature weights (7.4) and equally spaced quadrature points, is called the **Newton–Cotes formula** of order n. In order to illustrate the general idea, we consider two simple examples.

Trapezium rule. In this case we take $n = 1$, so that $x_0 = a$, $x_1 = b$; the Lagrange interpolation polynomial of degree 1 for the function f is simply

$$\begin{aligned} p(x) &= L_0(x) f(a) + L_1(x) f(b) \\ &= \frac{x - b}{a - b} f(a) + \frac{x - a}{b - a} f(b) \\ &= \frac{1}{b - a}[(b - x) f(a) + (x - a) f(b)]. \end{aligned}$$

Integrating $p(x)$ from a to b yields

$$\int_a^b f(x)\,dx \approx \frac{b - a}{2}[f(a) + f(b)].$$

This numerical integration formula is called the trapezium rule. The

terminology stems from the fact that the expression on the right is the area of the trapezium with vertices $(a, 0)$, $(b, 0)$, $(a, f(a))$, $(b, f(b))$.

Simpson's rule. A slightly more sophisticated quadrature rule is obtained by taking $n = 2$. In this case $x_0 = a$, $x_1 = (a+b)/2$ and $x_2 = b$, and the function f is approximated by a quadratic Lagrange interpolation polynomial.

The quadrature weights are calculated from

$$\begin{aligned} w_0 &= \int_a^b L_0(x)\,dx \\ &= \int_a^b \frac{(x - x_1)(x - x_2)}{(x_0 - x_1)(x_0 - x_2)}\,dx \\ &= \int_{-1}^{1} \frac{t(t-1)}{2}\,\frac{b-a}{2}\,dt \\ &= \frac{b-a}{6}, \end{aligned}$$

where it is convenient to make the change of variable

$$x = \frac{b-a}{2}\,t + \frac{b+a}{2}.$$

Similarly, $w_2 = -(b-a)$, and it is easy to see that $w_0 = w_2$ by symmetry. This gives

$$\int_a^b f(x)\,dx \approx \frac{b-a}{6}\left[f(a) + 4f\left(\frac{a+b}{2}\right) + f(b)\right],$$

a numerical integration formula known as Simpson's rule.

It is very important to notice that the weights w_k defined in (7.4) depend only on n and k, not on the function f. Their values can therefore

be calculated in advance, as in the trapezium rule and Simpson's rule. The evaluation of the approximation to the integral (7.1) is then a trivial matter; it is only necessary to compute $f(x_k)$ at each of the quadrature points x_k, $k = 0, 1, \ldots, n$, multiply by the known weights w_k for $k = 0, 1, \ldots, n$, and form the sum on the right-hand side of (7.3).

7.3 Error estimates

Our next task is to estimate the size of the error in the numerical integration formula (7.3), that is, the error that has been committed by integrating the interpolating Lagrange polynomial of f instead of f itself. The error in (7.3) is defined by

$$E_n(f) = \int_a^b f(x)\,dx - \sum_{k=0}^{n} w_k f(x_k).$$

The next theorem provides a useful bound on $E_n(f)$ under the additional hypothesis that the function f is sufficiently smooth.

Theorem 7.1 *Let $n \geq 1$. Suppose that f is a real-valued function, defined and continuous on the interval $[a, b]$, and let $f^{(n+1)}$ be defined and continuous on $[a, b]$. Then,*

$$E_n(f) \leq \frac{M_{n+1}}{(n+1)!} \int_a^b \pi_n(x)\,dx, \tag{7.5}$$

where $M_{n+1} = \max_{\xi \in [a,b]} |f^{(n+1)}(\xi)|$ and $\pi_n(x) = (x - x_0)\ldots(x - x_n)$.

Proof Recalling the definition of the weights w_k from (7.4), we can write $E_n(f)$ as follows:

$$\begin{aligned} E_n(f) &= \int_a^b f(x)\,dx - \int_a^b L_n(x) f(x)\,dx \\ &= \int_a^b [f(x) - p_n(x)]\,dx. \end{aligned}$$

Thus,

$$|E_n(f)| \leq \int_a^b |f(x) - p_n(x)|\,dx.$$

The desired error estimate (7.5) follows by inserting (6.8) into the right-hand side of this inequality. □

Let us use this theorem to estimate the size of the error which arises from applying the trapezium rule to the integral $\int_a^b f(x)\,dx$. In this case, with $n = 1$ and $\pi(x) = (x-a)(x-b)$, the bound (7.5) reduces to

$$\begin{aligned} E(f) &\le \frac{M}{2}\int_a^b (x-a)(x-b)\,dx \\ &= \frac{M}{2}\int_a^b (b-x)(x-a)\,dx \\ &= \frac{(b-a)^3}{12}M\,. \end{aligned} \tag{7.6}$$

An analogous but slightly more tedious calculation shows that, for Simpson's rule,

$$\begin{aligned} E(f) &\le \frac{M}{6}\int_a^b (x-a)(x-(a+b)/2)(x-b)\,dx \\ &= \frac{(b-a)^5}{196}M\,. \end{aligned} \tag{7.7}$$

Unfortunately, (7.7) gives a considerable overestimate of the error in Simpson's rule; in particular it does not bring out the fact that $E(f) = 0$ whenever f is a polynomial of degree 3. The next theorem will allow us to give a sharper bound on the error in Simpson's rule which illustrates this fact. More generally, it is quite easy to prove that when n is odd the Newton–Cotes formula (7.3) (with w_k defined by (7.4)) is exact for all polynomials of degree n, while when n is even it is also exact for all polynomials of degree $n+1$ (see Exercise 2 at the end of the chapter).

Theorem 7.2 *Suppose that f is a real-valued function, defined and continuous on the interval $[a,b]$, and that $f' = f^{(4)}$, the fourth derivate of f, is continuous on $[a,b]$. Then,*

$$\int_a^b f(x)\,dx - \frac{b-a}{6}[f(a) + 4f((a+b)/2) + f(b)] = -\frac{(b-a)^5}{2880}f^{(4)}(\xi)\,, \tag{7.8}$$

for some ξ in (a,b).

Proof Making the change of variable

$$x = \frac{a+b}{2} + \frac{b-a}{2}t\,, \qquad t \in [-1,1]\,,$$

and defining the function $t \mapsto F(t)$ by $F(t) = f(x)$, we see that

$$\int f(x)\,dx - \frac{b-a}{6}[f(a) + 4f((a+b)/2) + f(b)]$$
$$= \frac{b-a}{2}\int_{-} F(\tau)\,d\tau - \frac{1}{3}[F(-1) + 4F(0) + F(1)] \quad . \quad (7.9)$$

We now introduce the function $t \mapsto G(t)$ by

$$G(t) = \int_{-} F(\tau)\,d\tau - \frac{t}{3}[F(-t) + 4F(0) + F(t)], \quad t \in [-1, 1];$$

the right-hand side of (7.9) is then simply $-(b-a)G(1)$.

The remainder of the proof is devoted to showing that $-(b-a)G(1)$ is, in turn, equal to the right-hand side of (7.8) for some ξ in (a, b). To do so, we define

$$H(t) = G(t) - t\,G(1), \quad t \in [-1, 1],$$

and apply Rolle's Theorem repeatedly to the function H. Noting that $H(0) = H(1) = 0$, we deduce that there exists $\zeta_1 \in (0, 1)$ such that $H'(\zeta_1) = 0$. But it is easy to show that $H'(0) = 0$, so there exists $\zeta_2 \in (0, \zeta_1)$ such that $H''(\zeta_2) = 0$. Again we see that $H''(0) = 0$, so there exists $\zeta_3 \in (0, \zeta_2)$ such that $H'''(\zeta_3) = 0$. Now,

$$G'''(t) = -\frac{t}{3}[F'''(t) - F'''(-t)],$$

and therefore

$$H'''(\zeta_3) = -\frac{\zeta_3}{3}[F'''(\zeta_3) - F'''(-\zeta_3)] - 60\zeta_3\,G(1).$$

Applying the Mean Value Theorem to the function F''' this shows that there exists $\zeta_4 \in (-\zeta_3, \zeta_3)$ such that

$$H'''(\zeta_3) = -\frac{\zeta_3}{3}[2\zeta_3\,F''(\zeta_4)] - 60\zeta_3\,G(1)$$
$$= -\frac{2\zeta_3}{3}[F''(\zeta_4) + 90G(1)].$$

Since $H'''(\zeta_3) = 0$ and $\zeta_3 \neq 0$, this means that

$$G(1) = -\frac{1}{90}F''(\zeta_4) = -\frac{(b-a)}{1440}f''(\xi),$$

and the required result follows. □

Table 7.1. I *is the result of the Newton–Cotes formula of degree* n *for the approximation of the integral (7.12)*

	& '%
	%)$')
&	'(
'	&)'
(	&)%$
%	&) '(
)	$ $$
	('$
$	&$ %
	' %)&&
	& "))
	& $'
&	$ $
'	) $$('
(	' (((%

7.4 The Runge phenomenon revisited

By looking at the right-hand side of the error bound (7.5) we may be led to believe that by increasing n, that is by approximating the integrand by Lagrange interpolation polynomials of increasing degree and integrating these exactly, we shall reduce the size of the quadrature error $E(f)$. However, this is not always the case, even for very smooth functions f. An example of this behaviour uses the same function as in Section 6.3; Table 7.1 gives the results of applying Newton–Cotes formulae of increasing degree to the evaluation of the integral

$$\int_{-}^{!} \frac{1}{1+x}\,dx\,. \tag{7.12}$$

These results do not evidently converge as n increases, and in fact they eventually increase without bound. This behaviour is related to the fact that the weights w in the Newton–Cotes formula are not all positive when $n > 8$. We shall return to this point in Theorem 10.2.

A better approach to improving accuracy is to divide the interval $[a, b]$ into an increasing number of subintervals of decreasing size, and then to use a numerical integration formula of fixed order n on each

of the subintervals. Quadrature rules based on this approach are called composite formulae; in the next section we shall describe two examples.

7.5 Composite formulae

We shall consider only some very simple composite quadrature rules: the composite trapezium rule and the composite Simpson rule.

Suppose that f is a function, defined and continuous on a nonempty closed interval $[a, b]$ of the real line. In order to construct an approximation to

$$\int f(x)\,dx,$$

we now select an integer $m \geq 2$ and divide the interval $[a, b]$ into m equal subintervals, each of width $h = (b - a)/m$, so that

$$\int f(x)\,dx = \sum \int f(x)\,dx, \tag{7.13}$$

where

$$x_i = a + ih = a + \frac{i}{m}(b - a), \quad i = 0, 1, \ldots, m.$$

Each of the integrals is then evaluated by the trapezium rule,

$$\int f(x)\,dx \approx \frac{1}{2}h[f(x_{i-1}) + f(x_i)]; \tag{7.14}$$

summing these over $i = 1, 2, \ldots, m$ leads to the following definition.

Definition 7.1 (Composite trapezium rule)

$$\int f(x)\,dx \approx h\left[\frac{1}{2}f(x_0) + f(x_1) + \cdots + f(x_{m-1}) + \frac{1}{2}f(x_m)\right]. \tag{7.15}$$

1 , 1 , 1 6 , ," " 1 $
, 1 N ,
1 , 5 1 $ $, 7 ,
6 5 6 " 1% , 1 6 , B .&
6 , B) 6D C # 6 , B) 6DD4 % , N $
"" , 1 ,
, N 1 ! " #@@. , 1 %
F 6 ' 6 %2
4

The error in the composite trapezium rule can be estimated by using the error bound (7.6) for the trapezium rule on each individual subinterval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, m$. For this purpose, let us define

$$\begin{aligned} \varepsilon(f) &= \int_a^b f(x)\,dx - \frac{h}{2}\left[f(x_0) + f(x_1) + \cdots + f(x_{m-1}) + f(x_m)\right] \\ &= \sum_{i=1}^{m} \int_{x_{i-1}}^{x_i} f(x)\,dx - \frac{h}{2}\left[f(x_{i-1}) + f(x_i)\right]. \end{aligned}$$

Applying (7.6) to each of the terms under the summation sign we obtain

$$\varepsilon(f) \le \frac{1}{12}h^3 \max_{\xi \in [a,b]} |f''(\xi)| \le \frac{(b-a)}{12m^2} h^2 M, \tag{7.16}$$

where $M = \max_{\xi \in [a,b]} |f''(\xi)|$.

For Simpson's rule, let us suppose that the interval $[a, b]$ has been divided into $2m$ intervals by the points $x_i = a + ih$, $i = 0, 1, \ldots, 2m$, with $m \ge 2$ and

$$h = \frac{b-a}{2m},$$

and let us apply Simpson's rule on each of the intervals $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, m$, giving

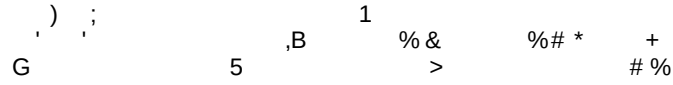
$$\int_{x_{i-1}}^{x_i} f(x)\,dx = \int_{x_{i-1}}^{x_i} f(x)\,dx - \frac{2h}{6}\left[f(x_{i-1}) + 4f(x_{i-\frac{1}{2}}) + f(x_i)\right].$$

This leads to the following definition.

Definition 7.2 (Composite Simpson rule)

$$\int_a^b f(x)\,dx \approx \frac{h}{3}\left[f(x_0) + 4f(x_1) + 2f(x_2) + 4f(x_3) + \cdots + 2f(x_{m-2}) + 4f(x_{m-1}) + f(x_m)\right]. \tag{7.17}$$

A schematic view of the pattern in which the coefficients 1, 4 and 2 appear in the composite Simpson rule is shown in Figure 7.1.



In order to estimate the error in the composite Simpson rule, we proceed in the same way as for the composite trapezium rule. Let us define

$$\begin{aligned} \%(f) &= \int_{x_{-1}}^{x_1} f(x)\,dx - \frac{h}{3}[f(x_{-1}) + 4f(x_0) + f(x_1)] \\ &= \sum_{r=1}^{m} \int_{x_{2r-2}}^{x_{2r}} f(x)\,dx - \frac{h}{3}[f(x_{2r-2}) + 4f(x_{2r-1}) + f(x_{2r})]\,. \end{aligned}$$

Applying (7.10) to each individual term in the sum and recalling that $b - a = 2mh$ we obtain the following error bound:

$$|\%(f)| \leq \frac{(b-a)}{2880m^4} M_4\,, \tag{7.18}$$

where $M_4 = \max_{a \leq \zeta \leq b} |f^{(4)}(\zeta)|$.

The composite rules (7.15) and (7.17) provide greater accuracy than the basic formulae considered in Section 7.2; this is clearly seen by comparing the error bounds (7.16) and (7.18) for the two composite rules with (7.6) and (7.8), the error estimates for the basic trapezium rule and Simpson rule respectively. The inequalities (7.16) and (7.18) indicate that, as long as the function f is sufficiently smooth, the errors in the composite rules can be made arbitrarily small by choosing a sufficiently large number of subintervals.

7.6 The Euler–Maclaurin expansion

We have seen in (7.16) that the error in the composite trapezium rule is bounded by a term involving $1/m^2$, where m is the number of subdivi-

sions of the interval $[a, b]$; the **Euler –Maclaurin** expansion expresses this error as a series in powers of $1/m$, and makes it possible to improve accuracy by extrapolation methods.

We first define a sequence of polynomials.

Definition 7.3 *Consider the sequence of polynomials* q , $r = 1, 2, \ldots$, *defined by their properties, as follows:*

(i) q *is a polynomial of degree* r*;*
(ii) *for each positive integer* r, q $= q$ *;*
(iii) q *is an odd function if* r *is odd, and an even function if* r *is even;*
(iv) *if* $r > 1$ *is odd, then* $q(-1) = 0$ *and* $q(1) = 0$*;*
(v) $q(t) = -t$.

Using these conditions it is easy to construct the polynomials q in succession. From (v) and (ii) we get

$$q(t) = -t + A, \qquad q(t) = -t + A\,t + A,$$

where A and A are constants. From (iii) we see that $A = 0$; then, from (iv) it follows that $A = -$. Hence,

$$q(t) = -t + -, \qquad q(t) = -t + -t.$$

We can then go on to construct q and q , and so on.

1 : B#> % #& & + C #. #&.* $
D E, , 1
1 , 6 , 1 , 6
1 1 1 4 1 #* , 2
, 5 , 5 6
, 6 1 4
, , 1 5 J 6 6 J G
, , , 6 6 $ 6 , 6
,' , 5 6 1 6 , ,4)
9 #&?> 1 1 6 ,
, , 4 : +N J , , N
1 R 8 , 6 4 , , 1
1 , 1 1 , 1 J G L
1 1 B D 1
1 , 1 1 1 G 1 # U U2
1 E , E J , 4
2 ' , B9 6 #?@. % 6 , C # K #& ?
, D , 5 6 1 F
1 ## , 1 # 4) #&#@ 1 #
, 9 1 6 , 64 0 7 " 1 &?* 5
0 % E 6 , 8 1 N
4 ' , N " , 8
, 1 4

where $t = t(x) = \quad 1 + \frac{}{*}(x \quad x_{-}\ )$ *for* $x \quad [x_{-}, x\ ]$, $i = 1, \ldots, m$, *and* $c\ = q\ \ (1)/2 \quad$ *for* $r = 1, \ldots, k$.

Proof We express the integral as a sum over the m subintervals $[x_{-}, x\ ]$, $i = 1, \ldots, m$, as in (7.13). In the interval $[x_{-}, x\ ]$ we change the variable by writing $x = x_{-}\ + h(t+1)/2$, so that

$$\int_{}^{} f(x)\mathrm{d}x = \frac{h}{2}\int_{-}^{} g(t)\mathrm{d}t\,,$$

where $f(x) = g(t)$. According to Theorem 7.3, then,

$$\begin{aligned}
&\int_{}^{} f(x)\mathrm{d}x \quad \frac{h}{2}[f(x_{-}) + f(x\ )] \\
&= \frac{h}{2}\int_{-}^{} g(t)\mathrm{d}t \quad [g(\ \ 1) + g(1)] \\
&= \frac{h}{2}\sum_{}^{} q\ \ (1)[g^{-}\ (1) \quad g^{-}\ (\ \ 1)] \\
&\qquad \int_{-}^{} \sum_{}^{} q\ \ (t)g\ \ (t)\mathrm{d}t\ \ .
\end{aligned}$$

On noting that $g\ \ (t) = (h/2)\ f\ \ (x)$, $\square = 1, 2, \ldots, 2k$, $\mathrm{d}t = (2/h)\,\mathrm{d}x$, summation over all the subintervals $[x_{-}, x\ ]$, for $i = 1, \ldots, m$, gives the required result. The important point is the symmetry of the polynomials $q\ $, which ensures that $q\ \ (1) = q\ \ (\ \ 1)$, so that all the derivatives of f at the internal points $x\ $ cancel in the course of summation, leaving only the derivatives at a and b. □

Remark 7.1 *By successively computing the polynomials* $q\ (t)$*, we can determine the values of* $c\ = q\ \ (1)/2\ \ $, $r = 1, 2, 3, \ldots$. *For example,*

$$c\ = \quad —, c\ = —, c\ = \quad ——, c\ = ———, c\ = \quad ———, \ldots.$$

It can be shown that $c\ = \quad ——$ *for all* $r = 1, 2, 3, \ldots$, *where* $B\ \ $ *are the Bernoulli numbers with even index, which can be determined from*

[illegible]

the Taylor series expansion

$$-\coth\ -\ =\ \sum^{\infty}\ \frac{B\quad x}{(2r)!}\,.$$

Easier still, typing ! *at the Maple command line gives* c = ————————; c , c , . . . *can be found in the same way.*

An interesting consequence of Theorem 7.4 concerns the numerical integration of smooth periodic functions. Suppose that f is a continuous function defined on (,) such that all derivatives of f, up to and including order 2k, are defined and continuous on (,), and f is periodic on (,) with period b a; *i.e.*, f(x + b a) f(x) = 0 for all x . Hence, by successive differentiation of this equality and taking x = a we deduce that, in particular,

$$f^{-}(b)\quad f^{-}(a) = 0 \qquad \text{for } r = 1, 2, \ldots, k\,.$$

Therefore, according to (7.20), we have that

$$I\quad T(m) = \quad (h\ )\,.$$

The fact that for k 1 this integration error is much smaller than the (h) error that will be observed in the case of a nonperiodic function indicates that the composite trapezium rule is particularly well suited for the numerical integration of smooth periodic functions.

A second application of the Euler–Maclaurin expansion concerns extrapolation methods. This subject will be discussed in the next section.

7.7 Extrapolation methods

In general the calculation of the higher derivatives involved in the Euler–Maclaurin expansion (7.20) is not possible. However, the existence of the expansion allows us to eliminate successive terms by repeated calculation of the trapezium rule approximation.

For example, the case k = 2 of (7.20) may be written in the form

$$\int_a^b f(x)\,dx\quad T(m) = C\ h\ +\ \ (m^{-}\ ),$$

where C = c [f′(b) f′(a)] and h = (b a)/m. This also means that

$$\int_a^b f(x)\,dx\quad T(2m) = C\ (h/2)\ +\ \ (m^{-}\ ).$$

We can eliminate the term in h from these two equalities, giving

$$\int f(x)dx = \frac{4T(2m) - T(m)}{3} + \quad (h\ ).$$

The same elimination process could be used for any two values of m, from the calculation of $T(m\ )$ and $T(m\ )$; the advantage of using m and $2m$ is that in the computation of $T(2m)$ half the required values of $f(x\ )$ are already known from $T(m)$, and we do not have to calculate them again. This process of eliminating the term in h from the expansion of the error is known as **Richardson extrapolation** or h **extrapolation**. It is easy to extend the process to higher-order terms. For example,

$$\int f(x)dx \quad T(m) = C\ h\ + C\ h\ + C\ h\ + \quad (h\ ).$$

Hence

$$\int f(x)dx \quad \frac{4T(2m) \quad T(m)}{3} = \quad -C\ h \quad -C\ h\ + \quad (h\ ),$$

which leads to

$$\int f(x)dx \quad \frac{16T\ (2m) \quad T\ (m)}{15} = \quad (h\ ),$$

where

$$T\ (m) = \frac{4T(2m) - T(m)}{3}.$$

Therefore,

$$T\ (m) = \frac{16T\ (2m) \quad T\ (m)}{15}$$

approximates the integral $\int f(x)dx$ to accuracy $(h\ )$. Adopting the notational convention

$$T\ (m) = T(m)$$

and proceeding recursively,

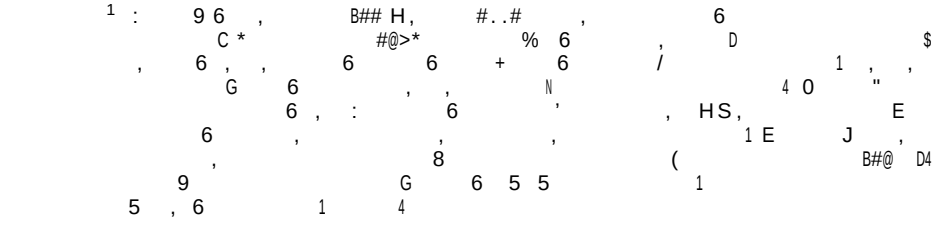


Table 7.2. *Romberg table.*

	& * +	&1* +	&2* +	&3* +	&4* +
'	&*'+	&1*'+	&2*'+	&3*'+	&4*'+
	&* +	&1* +	&2* +	&3* +	
%	&* %+	&1* %+	&2* %+		
&	&*& +	&1*& +			
%'	&*%'+				

$$T\ (m) = \frac{4\ T\ -\ (2m) \quad T\ -\ (m)}{4 \quad 1}, \quad k = 1, 2, 3, \ldots, \tag{7.21}$$

will approximate $'\ f(x)dx$ to accuracy $(h\ \ )$, provided of course that f exists and is continuous on the closed interval $[a, b]$. This extrapolation process is known as the **Romberg** integration method.

The intermediate results in Romberg's method are often arranged in the form of a table, known as the Romberg table. For example, if we start with $m = 4$ subdivisions of the closed interval $[a, b]$, each of length $h = (b \quad a)/4$, and proceed by doubling the number of subdivisions in each step (and thereby halving the spacing h between the quadrature points from the previous step), then the associated Romberg table is as shown in Table 7.2, where we took, successively, $m = 4, 8, 16, 32, 64$ subdivisions of the interval $[a, b]$ of length $h = (b \quad a)/m$ each. After $T\ (4) = T(4), \ldots, T\ (64) = T(64)$ have been computed, we calculate $T\ (4), \ldots, T\ (32)$ using (7.21) with $k = 1$, then we compute $T\ (4), \ldots, T\ (16)$ using (7.21) with $k = 2$, then $T\ (4), T\ (8)$ using (7.21) with $k = 3$, and finally $T\ (4)$ using (7.21) with $k = 4$. Provided that the integrand is sufficiently smooth, the numbers in the $T(m)$ column approximate the integral to within an error $(h\ )$; the numbers in the $T\ (m)$ column to within $(h\ )$, those in the $T\ (m)$ column to $(h\ )$, those in the $T\ (m)$ column to $(h\ )$, and those in the $T\ (m)$ column to within $(h\ \ )$.

1 1) 1 % ' ,
5 6 1 0 F 64 8 ,
(1 , ,) <F = (5
0 $ * C*? #@>>4

An example is shown in Table 7.3. This gives the results of calculating the integral

$$\int \frac{e^{-}}{1+4x}\,dx$$

by Romberg's method; first the trapezium rule is used successively with $m = 4, 8, 16, 32$ and 64 equal subdivisions of the interval $[0, 1]$ of length $h = (b \quad a)/m$ each. There are then four stages of extrapolation: Stage 1 involves computing $T\ (m)$ for $m = 4, 8, 16, 32$; Stage 2 computes $T\ (m)$ for $m = 4, 8, 16$; Stage 3 calculates $T\ (m)$ for $m = 4, 8$; and Stage 4 then computes $T\ (m)$ for $m = 4$. Not only does the extrapolation give an accurate result, but the consistency of the numerical values in the last two columns gives a good deal of confidence in quoting the result 0.220458 correct to six decimal digits. Note that none of the individual composite trapezium rule calculations in the $T(m)$ column gives a result correct to more than three decimal digits – not even $T(64)$ which uses 64 equal subdivisions of $[0, 1]$.

Table 7.3. *Romberg table for the calculation of* $\int (e^{-} \quad /(1+4x))dx$.

& * +	&1* +	&2* +	&3* +	&4* +
' '	&	')	'(	'(
)$)$	((	'($	'(	
% &)'	'%	'(		
& $'	'(			
%' ()$				

The success of Romberg integration is only justified if the integrand f satisfies the hypotheses of the Euler–Maclaurin Theorem. As an illustration of this, Table 7.4 shows the result of the same calculation, but for the integral

$$\int x \quad dx\,.$$

The function $x \quad x$ is not differentiable at $x = 0$, so the required conditions are not satisfied for any extrapolation. The numerical results bear this out; they are quite close to the correct value, $3/4$, but the behaviour of the extrapolation does not give any confidence in the accuracy of the result. In fact the extrapolation has not given much improvement

on T(64). The calculation of integrals involving this sort of singularity requires special methods which we shall not discuss here.

We have reached the end of this chapter, but do not despair: the story about numerical integration rules will continue. In Chapter 10 we shall discuss a class of quadrature formulae, generally referred to as Gaussian quadrature rules, which are distinct from the Newton–Cotes formulae considered here. Before doing so, however, in Chapters 8 and 9 we make a brief excursion into the realm of approximation theory.

Table 7.4. *Romberg table for the calculation of* ' x dx.

& * +	&1* +	&2* +	&3* +	&4* +
') ((	)' "	)'%$(	)' $	)'$(&'
	)&&	)'%% %	)')$	)'$(&
%	)'& &	)' %(&	)'$(	
&	)') $)	)'$'%(		
%'	)' $ &			

7.8 Notes

The material presented in this chapter is classical. For further details on the theory and practice of numerical integration, we refer to the following texts:

$ $ % (." *Methods of Numerical Integration,* Second Edition, Computer Science and Applied Mathematics, Academic Press, Orlando, FL, 1984;

, " *Approximate Calculation of Integrals,* translated from Russian by Arthur H. Stroud, ACM Monograph Series, Macmillan, New York, 1962;

/ " *Numerical Quadrature and Cubature,* Computational Mathematics and Applications, Academic Press, London, 1980.

The first of these is a standard text and contains a huge bibliography of more than 1500 entries. Concerning the implementation of numerical integration rules into mathematical software, the reader is referred to

) % , 0 " *Computational Integration*, SIAM, Philadelphia, 1998.

It includes a comprehensive overview of computational integration techniques based on both numerical and symbolical methods, and an exposition of some more recent number-theoretical, pseudorandom and lattice algorithms; these topics are beyond the scope of the present text.

Exercises

7.1 With the usual notation for the Newton–Cotes quadrature formula and using the equally spaced quadrature points $x = a+kh$ for $k = 0, 1, \ldots, n$ and n 1, show that $w = w$ – for $k = 0, 1, \ldots, n$.

7.2 By considering the polynomial $[x \quad (a+b)/2]$, n 1, and the result of Exercise 1, or otherwise, show that the Newton–Cotes formula using $n+1$ points x , $k = 0, 1, \ldots, n$, is exact for all polynomials of degree $n+1$ whenever n is even.

7.3 A quadrature formula on the interval $[\ 1, 1]$ uses the quadrature points $x = \alpha$ and $x = \alpha$, where $0 < \alpha$ 1:

$$\int_{-}^{!} f(x)\,dx \quad w\ f(\ \alpha) + w\ f(\alpha).$$

The formula is required to be exact whenever f is a polynomial of degree 1. Show that $w = w = 1$, independent of the value of α. Show also that there is one particular value of α for which the formula is exact also for all polynomials of degree 2. Find this α, and show that, for this value, the formula is also exact for all polynomials of degree 3.

7.4 The Newton–Cotes formula with $n = 3$ on the interval $[\ 1, 1]$ is

$$\int_{-}^{!} f(x)\,dx \quad w\ f(\ 1) + w\ f(\ 1/3) + w\ f(1/3) + w\ f(1).$$

Using the fact that this formula is to be exact for all polynomials of degree 3, or otherwise, show that

$$\begin{aligned} 2w + 2w &= 2, \\ 2w + -w &= -, \end{aligned}$$

and hence find the values of the weights w , w , w and w .

7.5 For each of the functions $1, x, x$, $\ldots, x$, find the difference between $'_{-} f(x)dx$ and (i) Simpson's rule, (ii) the formula derived in Exercise 4.

Deduce that for every polynomial of degree 5 formula (ii) is

more accurate than formula (i). Find a polynomial of degree 6 for which formula (i) is more accurate than formula (ii).

7.6 Write down the errors in the approximation of

$$\int x \, dx \quad \text{and} \quad \int x \, dx$$

by the trapezium rule and Simpson's rule. Hence find the value of the constant C for which the trapezium rule gives the correct result for the calculation of

$$\int (x \quad Cx \ ) \, dx,$$

and show that the trapezium rule gives a more accurate result than Simpson's rule when $- < C < -$.

7.7 Determine the values of c , $j = \quad 1, 0, 1, 2$, such that the quadrature rule

$$Q(f) = c_{-} \ f(\quad 1) + c \ f(0) + c \ f(1) + c \ f(2)$$

gives the correct value for the integral

$$\int f(x) \, dx$$

when f is any polynomial of degree 3. Show that, with these values of the weights c , and under appropriate conditions on the function f,

$$\int f(x) \, dx \quad Q(f) \qquad — M \ .$$

Give suitable conditions for the validity of this bound, and a definition of the quantity M .

7.8 Writing $T(m)$ for the composite trapezium rule defined in (7.15) and $S(2m)$ for the composite Simpson's rule defined in (7.17), show that

$$S(2m) = -T(2m) \quad -T(m).$$

7.9 Suppose that the function f has a continuous fourth derivative on the interval $[a, b]$, and that $T(m)$ denotes the composite trapezium rule approximation to $\int f(x)dx$, using m subintervals. Show that

$$\frac{T(m) \quad T(2m)}{T(2m) \quad T(4m)} \qquad 4 \quad \text{as } m \qquad .$$

Using the information in Table 7.3 evaluate this expression for $m = 4, 8, 16$.

7.10 With the same notation as in Exercise 9, suppose that the fourth derivative of f is not continuous on $[a, b]$, but that

$$\int_{-1}^{1} f(x)\,dx - T(m) = A/m^{\alpha} + E(m),$$

where $\alpha > 0$ and A are constants and $\lim_{m \to \infty} m^{\alpha} E(m) = 0$. Determine

$$\lim_{m \to \infty} \frac{T(m) - T(2m)}{T(2m) - T(4m)}.$$

Suggest a value of α which is consistent with the values of $T(m)$ given in Table 7.4.

7.11 The function f has a continuous fourth derivative on the interval $[-1, 1]$. Construct the Hermite interpolation polynomial of degree 3 for f using the interpolation points $x = -1$ and $x = 1$. Deduce that

$$\int_{-1}^{1} f(x)\,dx - [f(-1) + f(1)] = -[f'(-1) - f'(1)] + E,$$

where

$$E = -\max_{-1 \le x \le 1} f^{(4)}(x).$$

7.12 Construct the polynomials q_1, q_2, q_3 and q_4 given by Definition 7.3. Hence show that, in the notation of Theorem 7.4,

$$c_1 = -1/12, \quad c_2 = 1/720, \quad c_3 = -1/30240.$$

7.13 Using the relations

$$2\sin\tfrac{1}{2}x \sin jx = \cos\tfrac{1}{2}x - \cos(j + \tfrac{1}{2})x,$$

$$2\sin\tfrac{1}{2}x \cos jx = \sin(j + \tfrac{1}{2})x - \sin\tfrac{1}{2}x,$$

where m is a positive integer, show that the composite trapezium rule (7.15) with m subintervals will give the exact result for each of the integrals

$$\int_{-1}^{1} \cos rx\,dx, \qquad \int_{-1}^{1} \sin rx\,dx,$$

for any integer value of r which is not a multiple of m.

What values are given by the composite trapezium rule for these integrals when $\mathsf{r} = \mathsf{mk}$ and k is a positive integer?

8

Polynomial approximation in the -norm

8.1 Introduction

In Chapter 6 we considered the problem of interpolating a function by polynomials of a certain degree. Here we shall discuss other types of approximation by polynomials, the overall objective being to find the polynomial of given degree n which provides the 'best approximation' from & to a given function in a sense that will be made precise below.

8.2 Normed linear spaces

In order to be able to talk about 'best approximation' in a rigorous manner we need to recall from Chapter 2 the concept of *norm*; this will allow us to compare various approximations quantitatively and select the one which has the smallest approximation error. The definition given in Section 2.7 applies to a linear space consisting of functions in the same way as to the finite-dimensional linear spaces considered in Chapter 2.

Definition 8.1 *Suppose that is a linear space over the field of real numbers. A nonnegative function defined on whose value at f is denoted by $\|f\|$ is called a* **norm** *on if it satisfies the following axioms:*

$\|f\| = 0$ *if, and only if,* $f = 0$ *in ;*
$\|\lambda f\| = |\lambda| \|f\|$ *for all* λ *, and all* f *in ;*
$\|f+g\| \leq \|f\| + \|g\|$ *for all* f *and* g *in (the triangle inequality).*

A linear space , equipped with a norm, is called a **normed linear space**.

Throughout this chapter [a, b

Lemma 8.1 *(i) Suppose that the real-valued weight function* w *is defined, continuous, positive and integrable on the interval* (a, b)*. Then, for any function* f C[a, b]*,*

$$f \quad W \ f \ _{\infty}, \quad \textit{where } W = \quad w(x)dx \quad .$$

(ii) Given any two positive numbers ε *(however small) and* M *(however large), there exists a function* f C[a, b] *such that*

$$f \ < \varepsilon, \quad f \ _{\infty} > M .$$

Proof The proof is left as an exercise (see Exercise 1). □

The definitions (2.33) and (2.34) of the vector norms $_{\infty}$ and on imply that

$$n^{-} \quad _{\infty} \quad n \quad _{\infty} \quad , \tag{8.3}$$

which means that, to all intents and purposes, these two norms are interchangeable. Lemma 8.1 indicates that a similar chain of inequalities cannot possibly hold for the norms (8.1) and (8.2) on C[a, b], and the choice between them may therefore significantly influence the outcome of the analysis.

Stimulated by the first axiom of *norm*, we shall think of f C[a, b] as being well approximated by a polynomial p on [a, b] if f p is small, where is either $_{\infty}$ or defined, respectively, by (8.1) or (8.2). In the light of Lemma 8.1, it should come as no surprise that the mathematical tools for the analysis of smallness of f p $_{\infty}$ are quite different from those that ensure smallness of f p . We have therefore chosen to discuss these two matters separately: the present chapter focuses on the -norm (8.1), while Chapter 9 explores the use of the 2-norm (8.2).

Despite the fundamental differences between the norms (8.1) and (8.2) which we have alluded to above, there is a common underlying feature which is independent of the choice of norm: if *no limitation is imposed*

on the degree of the approximating polynomial p, then the approximation error f p can be made *arbitrarily small* in both norms. This is a central result in the theory of polynomial approximation and is formulated in the next theorem.

Theorem 8.1 (Weierstrass Approximation Theorem) *Suppose that* f *is a real-valued function, defined and continuous on a bounded closed interval* [a, b] *of the real line; then, given any* ε > 0*, there exists a polynomial* p *such that*

$$\| f - p \|_\infty < \varepsilon.$$

Further, if w *is a real-valued function, defined, continuous, positive and integrable on* (a, b)*, then an analogous result holds in the* 2*-norm over the interval* [a, b] *with weight function* w*.*

This is an important theorem in classical analysis, and several proofs are known. It is evidently sufficient to consider only the interval [0, 1]; a simple change of variable will then extend the proof to any bounded closed interval [a, b]. For a real-valued function f , defined and continuous on the interval [0, 1], Bernstein's proof uses the polynomial

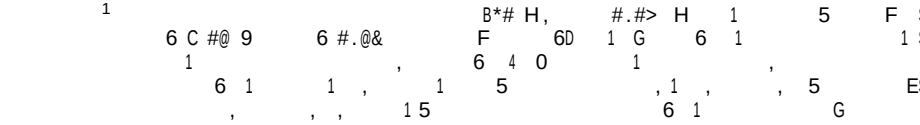
$$p_n(x) = \sum_{k=0}^{n} f(k/n)\, p_{n,k}(x), \qquad x \in [0,1],$$

where the **Bernstein polynomials** $p_{n,k}(x)$ are defined by

$$p_{n,k}(x) = \binom{n}{k} x^k (1-x)^{n-k}, \qquad x \in [0,1].$$

It can then be shown that, for any ε > 0, there exists n = n(ε) such that $\| f - p_n \|_\infty < \varepsilon$. The second part of the theorem is a direct consequence of this result, using part (i) of Lemma 8.1.

The details of the proof are given in Exercise 12. For an alternative proof, the reader is referred to Theorem 6.3 in M.J.D. Powell, *Approximation Theory and Methods*, Cambridge University Press, 1996.



8.3 Best approximation in the -norm

According to the Weierstrass Approximation Theorem any function f in $C[a,b]$ can be approximated arbitrarily well from the set of *all* polynomials. Clearly, if instead of the set of all polynomials we restrict ourselves to the set of polynomials $\mathcal{P}$ of degree n or less, with n *fixed*, then it is no longer true that, for any f $C[a,b]$ and any $\varepsilon > 0$, there exists p $\mathcal{P}$ such that

$$\| f - p \|_\infty < \varepsilon .$$

Consider, for example, the function x $\sin x$ defined on the interval $[0,\pi]$ and fix $n = 0$; then $\| f - q \|_\infty$ $1/2$ for any q $\mathcal{P}$, and therefore there is no q in $\mathcal{P}$ such that $\| f - q \|_\infty < 1/2$. A similar situation will arise if $\mathcal{P}$ is replaced by $\mathcal{P}$, with the polynomial degree n fixed.

It is therefore relevant to enquire just how well a given function f in $C[a,b]$ may be approximated by polynomials of a fixed degree n 0. This question leads us to the following approximation problem.

(A) Given that f $C[a,b]$ and n 0, fixed, find p $\mathcal{P}$ such that

$$\| f - p \|_\infty = \inf_{q \in \mathcal{P}} \| f - q \|_\infty ;$$

such a polynomial p is called a **polynomial of best approximation of degree n to the function f in the -norm**.

The next theorem establishes the existence of a polynomial of best approximation, showing, in particular, that the infimum of $\| f - q \|_\infty$ over q $\mathcal{P}$ is attained. We shall consider the question of uniqueness of the polynomial of best approximation later on, in Theorem 8.5.

Theorem 8.2 *Given that* f $C[a,b]$*, there exists a polynomial* p $\mathcal{P}$ *such that* $\| f - p \|_\infty = \min_{q \in \mathcal{P}} \| f - q \|_\infty$.

Proof Let us define the function $(c_0, \ldots, c_n)$ $E(c_0, \ldots, c_n)$ of $n+1$ real variables by

$$E(c_0, \ldots, c_n) = \| f - q \|_\infty, \quad \text{where } q(x) = c_0 + \cdots + c_n x^n .$$

1 1 , 1 6 E8 , 1 < =L
1 8 ,
1

H 6 1 6
< =2 6 , 1 , , 1 1 6
, 5 G , 1 6 B 1 , D < =4

We shall first show that E is continuous; this will imply that E attains its bounds on any bounded closed set in . We shall then construct a nonempty bounded closed set $'$ such that the lower bound of E on $'$ is the same as its lower bound over the whole of .

To show that E is continuous at each point $(c_0, \ldots, c_n)$, consider any $(\delta_0, \ldots, \delta_n)$ and define the polynomial η & by $\eta(x) = \delta_0 + \cdots + \delta_n x^n$. We see from the triangle inequality that

$$\begin{aligned} E(c_0+\delta_0, \ldots, c_n+\delta_n) &= \|f - (q+\eta)\|_\infty \\ &\le \|f - q\|_\infty + \|\eta\|_\infty \\ &= E(c_0, \ldots, c_n) + \|\eta\|_\infty . \end{aligned}$$

Now, for any given positive number ε, choose $\delta = \varepsilon/(1 + \cdots + K^n)$, where $K = \max(|a|, |b|)$. Consider any $(\delta_0, \ldots, \delta_n)$ such that $|\delta_i| \le \delta$ for all $i = 0, \ldots, n$. Then,

$$\begin{aligned} E(c_0+\delta_0, \ldots, c_n+\delta_n) - E(c_0, \ldots, c_n) &\le \|\eta\|_\infty \\ &\le \max_{\square} (|\delta_0| + |\delta_1| x| + \cdots + |\delta_n| x^n|) \\ &\le \delta(1 + \cdots + K^n) \\ &< \varepsilon . \end{aligned} \tag{8.4}$$

Similarly,

$$\begin{aligned} E(c_0, \ldots, c_n) &= \|f - q\|_\infty = \|f - (q+\eta) + \eta\|_\infty \\ &\le \|f - (q+\eta)\|_\infty + \|\eta\|_\infty \\ &\le E(c_0+\delta_0, \ldots, c_n+\delta_n) + \varepsilon, \end{aligned}$$

and therefore

$$E(c_0, \ldots, c_n) - E(c_0+\delta_0, \ldots, c_n+\delta_n) \le \varepsilon . \tag{8.5}$$

From (8.4) and (8.5) we deduce that

$$|E(c_0+\delta_0, \ldots, c_n+\delta_n) - E(c_0, \ldots, c_n)| < \varepsilon$$

for all $(\delta_0, \ldots, \delta_n)$ such that $|\delta_i| < \delta$, $i = 0, \ldots, n$, where now $\delta = \varepsilon/(1 + \cdots + K^n)$ and $K = \max(|a|, |b|)$. Hence E is continuous at $(c_0, \ldots, c_n)$. Since $(c_0, \ldots, c_n)$ is an arbitrary point in , it follows that E is continuous on the whole of .

Let us denote by $'$ the set of all points $(c_0, \ldots, c_n)$ in such that $E(c_0, \ldots, c_n) \le \|f\|_\infty + 1$. The set $'$ is evidently bounded and closed in ; further, $'$ is nonempty since $E(0, \ldots, 0) = \|f\|_\infty \le \|f\|_\infty + 1$, so that $(0, \ldots, 0) \in '$. Hence the continuous function E attains its

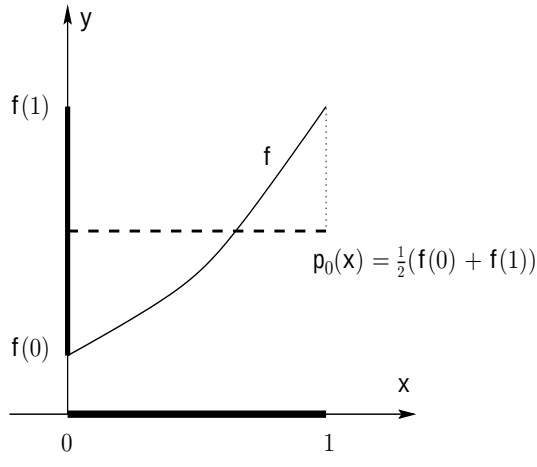
lower bound over the set ' ; let us denote this lower bound by d and let $(c^{\square}, \ldots, c^{\square})$ denote the point in ' where it is attained.

Since $(0, \ldots, 0)$ ', it follows that

$$d = \min_{+ \; + \; \square S} E(c, \ldots, c) \quad E(0, \ldots, 0) = f \; _{\infty} .$$

According to the definition of ',

$$E(c, \ldots, c$$



F ! ! $0 0 B

The polynomial p will be of the form p (x) c , and we need to determine c so that

$$\| f - p \|_\infty = \max_{\square} | f(x) - c |$$

is minimal. Since f is monotonic increasing, f (x) c attains its minimum at x = 0 and its maximum at x = 1; therefore f (x) c reaches its maximum value at one of the endpoints of [0, 1], *i.e.*,

$$E(c) = \max_{\square} | f(x) - c | = \max \{ | f(0) - c |, | f(1) - c | \} .$$

Clearly,

$$E(c) = \begin{cases} f(1) - c & \text{if } c < -(f(0) + f(1)) , \\ c - f(0) & \text{if } c \geq -(f(0) + f(1)) . \end{cases}$$

Drawing the graph of the function c E(c) shows that the minimum is attained when c = -(f (0) + f (1)). Consequently, the desired minimax polynomial of degree 0 for the function f is

$$p(x) \equiv -(f(0) + f(1)) , \qquad x \in [0, 1] .$$

The function f and its minimax approximation p & are depicted in Figure 8.1.

More generally, if f C[a, b] (not necessarily monotonic), and ξ and η denote two points in [a, b] where f attains its minimum and maximum

values, respectively, then the minimax polynomial of degree 0 to f on [a, b] is

$$p \ (x) \quad -(f(\xi) + f(\eta)) , \qquad x \quad [a, b] .$$

This example shows that the minimax polynomial p of degree *zero* for f C[a, b] has the property that the approximation error f p attains its extrema at *two* points, $x = \xi$ and $x = \eta$, with the error

$$f(x) \quad p \ (x) = \frac{1}{2}(f(x) \quad f(\xi)) + \frac{1}{2}(f(x) \quad f(\eta))$$

being *negative at one point*, $x = \xi$, and *positive at the other*, $x = \eta$. We shall prove that a property of this kind holds in general; the precise formulation of the general result is given in Theorem 8.4 which is, due to the oscillating nature of the approximation error, usually referred to as the Oscillation Theorem: it gives a complete characterisation of the minimax polynomial and provides a method for its construction. We begin with a preliminary result due to de la Vallée Poussin.

Theorem 8.3 (De la Vallée Poussin's Theorem) *Let* f C[a, b] *and* r & . *Suppose that there exist* n + 2 *points* x < < x *in the interval* [a, b]*, such that* f(x) r(x) *and* f(x) r(x) *have opposite signs, for* $i = 0, \ldots, n$. *Then,*

$$\min_{\square P} \quad f \quad q \ _\infty \qquad \min \qquad f(x) \quad r(x) . \tag{8.6}$$

Proof The condition on the signs of f(x) r(x) is usually expressed by saying that f r has alternating signs at the points x , $i = 0, 1, \ldots, n+1$. Let us denote the right-hand side of (8.6) by μ. Clearly, μ 0; when $\mu = 0$ the statement of the theorem is trivially true, so we shall assume that $\mu > 0$. Suppose that (8.6) is false; then, for a minimax polynomial approximation p & to the function f we have

$$f \quad p \ _\infty = \min_{\square P} \quad f \quad q \ _\infty < \mu .$$

[illegible]

Therefore,

$$p(x) \quad f(x) < r(x) \quad f(x), \quad i = 0, 1, \ldots, n+1.$$

Now,

$$r(x) \quad p(x) = [r(x) \quad f(x)] \quad [p(x) \quad f(x)], \quad i = 0, 1, \ldots, n+1.$$

Since the first term on the right always exceeds the second term in absolute value, it follows that $r(x) \quad p(x)$ and $r(x) \quad f(x)$ have the same sign for $i = 0, 1, \ldots, n+1$. Hence $r \quad p$, which is a polynomial of degree n, changes sign $n+1$ times. Thus, the assumption that (8.6) is false has led to a contradiction, and the proof is complete. □

Theorem 8.3 gives a clue to formulating a constructive characterisation of the *minimax polynomial*: indeed, we shall show that if the quantities $f(x) \quad r(x)$, $i = 0, 1, \ldots, n+1$, in Theorem 8.3 are all equal to $f \quad r \quad \infty$, then $r \quad \&$ is, in fact, a minimax polynomial of degree n for the function f on the interval $[a, b]$.

Theorem 8.4 (The Oscillation Theorem) *Suppose that* $f \quad C[a, b]$. *A polynomial* $r \quad \&$ *is a minimax polynomial for* f *on* $[a, b]$ *if, and only if, there exists a sequence of* $n+2$ *points* x, $i = 0, 1, \ldots, n+1$, *such that* $a \quad x < \quad < x \quad b$,

$$f(x) \quad r(x) = f \quad r \quad \infty, \quad i = 0, 1, \ldots, n+1,$$

and

$$f(x) \quad r(x) = [f(x \quad) \quad r(x \quad)], \quad i = 0, \ldots, n.$$

The statement of the theorem is often expressed by saying that $f \quad r$ attains its maximum absolute value with alternating signs at the points x. The points x, $i = 0, 1, \ldots, n+1$, in the Oscillation Theorem are referred to as **critical points**.

Proof of theorem If $f \quad \&$, then the result is trivially true, with $r = f$ and any sequence of $n+2$ distinct points x, $i = 0, 1, \ldots, n+1$, contained in $[a, b]$. Thus, we shall suppose throughout the proof that $f \; / \; \&$, *i.e.*, f is such that there is no polynomial $p \quad \&$ whose restriction to $[a, b]$ is identically equal to f.

The sufficiency of the condition stated in the theorem is easily shown. Suppose that the sequence of points x, $i = 0, 1, \ldots, n+1$, exists with

the given properties. Define

$$L = \|f - r\|_\infty \quad \text{and} \quad E(f) = \min_{q \in P} \|f - q\|_\infty .$$

From De la Vallée Poussin's Theorem, Theorem 8.3, it follows that $E(f) \geq L$. By the definition of $E(f)$ we also see that $E(f) \leq \|f - r\|_\infty = L$. Hence $E(f) = L$, and the given polynomial r is a minimax polynomial.

For the necessity of the condition, suppose that the given polynomial $r \in \mathcal{E}$ is a minimax polynomial for f on $[a, b]$. As $x \mapsto f(x) - r(x)$ is a continuous function on the bounded closed interval $[a, b]$, there exists a point in $[a, b]$ at which $f(x) - r(x)$ attains its maximum value, $L > 0$; let

$$x^* = \min \{x \in [a, b] : |f(x) - r(x)| = L\} .$$

Now, $x^* = b$ would imply that $f(x) - r(x) = L$ for all $x \in [a, b]$. As f is continuous on $[a, b]$, it would then follow that either $f(x) = r(x) + L$ for all $x \in [a, b]$ or $f(x) = r(x) - L$ for all $x \in [a, b]$; either way, we would find that $r \in \mathcal{E}$, which is assumed not to be the case. Therefore, $x^* \in [a, b)$; we may assume without loss of generality that $f(x^*) - r(x^*) = L > 0$.

Now, we shall prove the existence of the next critical point, $x^{**} \in (x^*, b]$ such that $f(x^{**}) - r(x^{**}) = -L$. Suppose otherwise, for contradiction; then, $-L < f(x) - r(x) < L$ for all x in $[a, b]$. Thus, by the continuity of f, there exists $\delta \in (0, L)$ such that $-L + \delta \leq f(x) - r(x) \leq L$ for all $x \in [a, b]$. Let us define $r^* \in \mathcal{E}$ by

$$r^*(x) = r(x) + \varepsilon ,$$

where $0 < \varepsilon < \min\{\delta, L - \delta\}$. Then, for all $x \in [a, b]$,

$$f(x) - r^*(x) = f(x) - r(x) - \varepsilon < L - \delta - \varepsilon < L$$

and

$$f(x) - r^*(x) = f(x) - r(x) - \varepsilon > -L + \delta - \varepsilon > -L ,$$

which means that

$$\|f - r^*\|_\infty < L = \|f - r\|_\infty .$$

Hence, $r^* \in \mathcal{E}$ is a better approximation to f on $[a, b]$ than $r \in \mathcal{E}$ is. This, however, contradicts our hypothesis that r is a polynomial of best approximation to f on $[a, b]$ from $\mathcal{E}$, and implies the existence of

$$x^{**} = \inf \{x \in (x^*, b] : |f(x) - r(x)| = L\} .$$

Consequently, $f(x) \quad r(x) = \quad L$ and $x \quad (x, b]$, as required; thus if $n = 0$, the proof is complete.

Let us, therefore, suppose that n

on each of the intervals $[\eta_i, \eta_{i+1}]$, $i = 0, 1, \ldots, m$ (whose union is $[a, b]$). We shall prove that, for $\varepsilon > 0$ sufficiently small,

$$\|f(x) - r^*(x)\| < L = \|f - r\|_\infty$$

for all x in $[\eta_i, \eta_{i+1}]$ and all $i = 0, 1, \ldots, m$; *i.e.*, $\|f - r^*\|_\infty < \|f - r\|_\infty$, contradicting the fact that $r \in \mathcal{R}_{m,n}$ is a minimax polynomial for f on $[a, b]$, and refuting the hypothesis that $1 \le m \le n$.

Take, for example, the interval $[\eta_0, \eta_1]$. For each x in $[\eta_0, \eta_1)$ we have $v(x) > 0$ and therefore, by the definition of $r^*(x)$ and property (a) above,

$$f(x) - r^*(x) \le L - \varepsilon v(x) < L, \qquad x \in [\eta_0, \eta_1).$$

Further, as $v(\eta_0) = 0$, it follows from (d) that

$$f(\eta_0) - r^*(\eta_0) = f(\eta_0) - r(\eta_0) < L.$$

Therefore, $f(x) - r^*(x) < L$ for each x in $[\eta_0, \eta_1]$. For a lower bound on $f(x) - r^*(x)$, note that by (a) and (c), $f(x) - r(x) > -L$ for all x in $[\eta_0, \eta_1]$. As $f - r$ is a continuous function on $[\eta_0, \eta_1]$, there exists $\delta \in (0, L)$ such that $f(x) - r(x) \le -L + \delta$ for all x in $[\eta_0, \eta_1]$. Thus, for $0 < \varepsilon < \min\{L, \delta/\|v\|_\infty\}$, where

$$\varepsilon = \frac{\delta}{\max_{[\eta_0, \eta_1]} v(x)},$$

we have that

$$f(x) - r^*(x) \le -L + \delta - \varepsilon v(x) < -L, \qquad x \in [\eta_0, \eta_1).$$

Further, by (d) above,

$$f(\eta_0) - r^*(\eta_0) = f(\eta_0) - r(\eta_0) > -L.$$

Hence, $f(x) - r^*(x) > -L$ for all $x \in [\eta_0, \eta_1]$, for $0 < \varepsilon < \min\{L, \delta/\|v\|_\infty\}$. Combining the upper and lower bounds on $f(x) - r^*(x)$, we deduce that

$$|f(x) - r^*(x)| < L = \|f - r\|_\infty, \qquad x \in [\eta_0, \eta_1].$$

Arguing in the same manner on each of the other intervals $[\eta_i, \eta_{i+1}]$, $i = 1, \ldots, m$, with $0 < \varepsilon < \min\{L, \delta_i/\|v\|_\infty\}$, $i = 1, \ldots, m$, and δ_i and ε_i defined analogously to δ and ε above, we conclude that

$$|f(x) - r^*(x)| < L = \|f - r\|_\infty, \qquad x \in [\eta_i, \eta_{i+1}], \quad i = 0, 1, \ldots, m,$$

and hence, for $0 < \varepsilon < \min\{L, \delta_0/\|v\|_\infty, \ldots, \delta_m/\|v\|_\infty\}$,

$$\|f - r^*\|_\infty < L = \|f - r\|_\infty.$$

of replacing r by $r^{\square}(x) = r(x)$ $\varepsilon v(x)$, with $\varepsilon > 0$, is indicated by the arrows. Since f $r^{\square} = f$ $r + \varepsilon v(x)$ and v is negative for x (η, η) and positive outside (η, η), f $r^{\square}$ will be smaller than f r at each of the points P, $i = 0, 1, 2$. There are two other local extrema for the error function f r: a minimum at Q and a maximum at R. Since both these points are to the right of η, where $v(x) > 0$, we shall have f $r^{\square} > f$ r at both of Q and R, and f $r^{\square} > f$ r at R. The magnitude of the extra term $\varepsilon v(x)$ must therefore be limited by the need to avoid the new difference f $r^{\square}$ becoming too large at R. We can achieve this by selecting $\varepsilon > 0$ sufficiently small. In this illustration the polynomial r & is not a minimax approximation to f on the given interval, since we can construct a better approximation $r^{\square}$ which is also in & .

We can now apply the Oscillation Theorem to prove that the minimax polynomial is unique.

Theorem 8.5 (Uniqueness Theorem) *Suppose that* $[a, b]$ *is a bounded closed interval of the real line. Each* f $C[a, b]$ *has a unique minimax polynomial* p & *on* $[a, b]$.

Proof Suppose that q & is also a minimax polynomial for f, and that p and q are distinct. Then,

$$f \quad p \quad _{\infty} = f \quad q \quad _{\infty} = E \ (f),$$

where, as in the proof of the Oscillation Theorem, we have used the notation

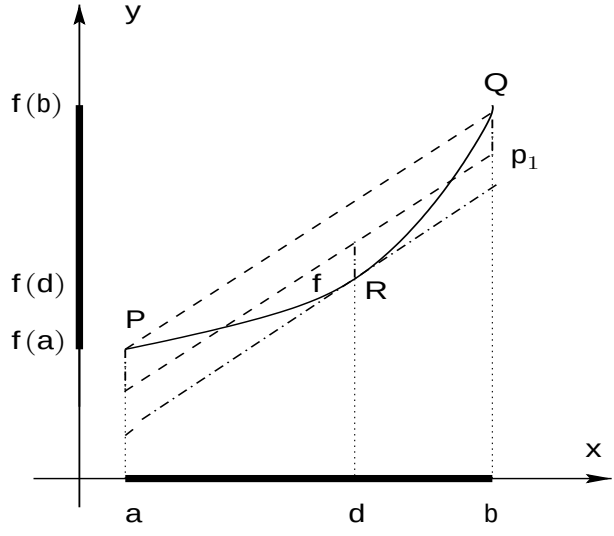
$$E \ (f) = \min_{\square P} f \quad q \quad _{\infty}.$$

This implies, by the triangle inequality, that

$$\begin{aligned} f \quad -(p \ + q \ ) \ _{\infty} \ &= \ -(f \quad p \ ) + -(f \quad q \ ) \ _{\infty} \\ &\quad - f \quad p \quad _{\infty} + - f \quad q \quad _{\infty} \\ &= \ E \ (f). \end{aligned}$$

Therefore $-(p \ + q \ )$ & is also a minimax polynomial approximation to f on $[a, b]$. By the Oscillation Theorem there exists a sequence of $n + 2$ critical points x, $i = 0, 1, \ldots, n + 1$, at which

$$f(x \ ) \quad -(p \ (x \ ) + q \ (x \ )) = E \ (f), \quad i = 0, 1, \ldots, n + 1.$$



& 8 ! B

equations

$$\begin{array}{ll} f(a) & (c\,a + c\,) = A, \\ f(d) & (c\,d + c\,) = \quad A, \\ f(b) & (c\,b + c\,) = A, \end{array} \tag{8.7}$$

where either $A = L$ or $A = \quad L$, with $L = \max \square \quad f(x) \quad p\,(x)$. Along with the condition

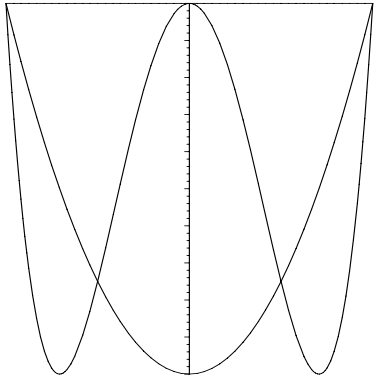
$$f^{\square}(d) = c \tag{8.8}$$

this gives four equations to determine the unknowns d, c , c and A.

Subtracting the first equation in (8.7) from the third equation, we get $f(b) \quad f(a) = c\,(b \quad a)$, whereby $c = (f(b) \quad f(a))\,/\,(b \quad a)$. Now, by the Mean Value Theorem, Theorem A.3, with this choice of c equation (8.8) has at least one solution, d, in the open interval in (a, b). In fact, the value of d is uniquely determined by (8.8), as $f^{\square}$ is continuous and strictly monotonic increasing. Next, c can be determined by adding the second equation in (8.7) to the first. Having calculated both c and c we insert them into the first equation in (8.7) to obtain A; finally $L = A$.

The construction of the minimax polynomial p is illustrated in Figure 8.3; R is the point at which the tangent to the curve $y = f(x)$ is parallel to the chord PQ; the graph of $p\,(x)$ is parallel to these two lines, and lies half-way between them.

Table 8.1.



they are all real and distinct, and lie in (1, 1);

(v) $T(x) \quad 1$ *for all* $x \quad [\ 1, 1]$ *and all* $n \quad 0$;

(vi) *for* $n \quad 1$, $T(x) = \quad 1$, *alternately at the* $n + 1$ *points* $x = \cos(k\pi/n)$, $k = 0, 1, \ldots, n$.

We can now apply the Oscillation Theorem to construct the minimax polynomial of degree n for $f : x \quad x$ on the interval $[\ 1, 1]$.

Theorem 8.6 *Suppose that* $n \quad 0$. *The polynomial* $p \quad \&$ *defined by*

$$p(x) = x \qquad 2^{-} T \quad (x), \qquad x \quad [\ 1, 1],$$

is the minimax approximation of degree n *to the function* $x \quad x$ *on the interval* $[\ 1, 1]$.

Proof By part (ii) of Lemma 8.2, $p \quad \&$. Since

$$x \qquad p(x) = 2^{-} T \quad (x),$$

by parts (v) and (vi) of Lemma 8.2, the difference $x \qquad p(x)$ does not exceed 2^{-} in the interval $[\ 1, 1]$, and attains this value with alternating signs at the $n + 2$ points $x = \cos(k\pi/(n + 1))$, $k = 0, 1, \ldots, n + 1$. Therefore, by the Oscillation Theorem, p is the (unique) minimax polynomial approximation from $\&$ to the function $x \quad x$ over $[\ 1, 1]$. □

A polynomial of degree n whose leading coefficient, the coefficient of x , is equal to 1, is called a **monic polynomial** of degree n. For example, the polynomial $r \quad \&$ defined by $r(x) = x \qquad q(x)$ with $q \quad \&$, is a monic polynomial of degree $n + 1$.

Corollary 8.1 *Suppose that* $n \quad 0$. *Among all monic polynomials of degree* $n + 1$ *the polynomials* $2^{-} T \quad$ *and* $2^{-} T \quad$ *have the smallest* -*norm on the interval* $[\ 1, 1]$.

Proof Let $\&$ denote the set of all monic polynomials of degree $n + 1$. Any $r \quad \&$ can be regarded as the difference between the function $x \quad x$ and a polynomial of lower degree, *i.e.*, $r(x) = x \qquad q(x)$ with $q \quad \&$. Hence, by Theorem 8.6,

$$\begin{aligned} \min_{\square P} \quad r \ _{\infty} &= \min_{\square P} \quad x \qquad q \ _{\infty} \\ &= \quad x \qquad (x \qquad 2^{-} T \quad) \ _{\infty} \\ &= \quad 2^{-} T \quad \ _{\infty}; \end{aligned}$$

the minimum is, therefore, achieved when r &

Proof Let $\tau_j = \cos\left((j + \tfrac{1}{2})\pi/(n+1)\right)$, $j = 0, 1, \ldots, n$, denote the zeros of the polynomial $T_{n+1}(t)$ (in the interval $(-1, 1)$). Hence,

$$(t - \tau_0)^* \cdots (t - \tau_n)^* = 2^{-n} T_{n+1}(t), \qquad t \in [-1, 1].$$

Let us define the points ξ_j, $j = 0, 1, \ldots, n$, as in the statement of the theorem. Clearly $\xi_j \in (a, b)$ is the image of $\tau_j \in (-1, 1)$ under the linear transformation $t \to x = \tfrac{1}{2}(b - a)t + \tfrac{1}{2}(b + a)$; we note in passing that the inverse of this mapping is $x \to t(x) = (2x - a - b)/(b - a)$; thus,

$$(x - \xi_0)^* \cdots (x - \xi_n)^* = \left(\frac{b-a}{2}\right)^* \cdots \left(t(x) - \tau_n\right)^* = \left(\frac{b-a}{2}\right)^{n+1} 2^{-n} T_{n+1}(t(x)).$$

The required bound now follows from (8.9), since $|T_{n+1}(t(x))| \le 1$ for all $x \in [a, b]$, and therefore $\|\pi_{n+1}\|_\infty \le (b - a)^{n+1} 2^{-n}$. □

The De la Vallée Poussin Theorem, Theorem 8.3, suggests the notion of a **near-minimax** polynomial, which is a polynomial $p \in \mathcal{P}_n$ such that the difference $f(x) - p(x)$ changes sign at $n + 1$ points ξ_j, $j = 0, 1, \ldots, n$, with $a < \xi_0 < \cdots < \xi_n < b$; for the difference $f(x) - p(x)$ then attains a local maximum or minimum with alternating signs in each of the intervals $[a, \xi_0), (\xi_0, \xi_1), \ldots, (\xi_n, b]$. The positions of these alternating local maxima and minima are then the points x_i, $i = 0, 1, \ldots, n + 1$, required by Theorem 8.3, and we therefore know that the ∞-norm of the error of the minimax polynomial lies between the least and greatest of the absolute values of these local maxima and minima. In particular, we should expect that if the sizes of these local maxima and minima are not greatly different, then the error of the near-minimax approximation should not be very much larger than the error of the minimax approximation.

Given any set of points ξ_i, $i = 0, 1, \ldots, n$, with $a < \xi_0 < \cdots < \xi_n < b$, the polynomial $\pi_{n+1}(x) = (x - \xi_0) \ldots (x - \xi_n)$ changes sign at the $n + 1$ points ξ_j, $j = 0, 1, \ldots, n$. Let us assume that $f \in C[a, b]$, $f^{(n+1)}$ exists and is continuous on $[a, b]$, and $f^{(n+1)}$ has the same sign on the whole of (a, b). It then follows that the product $f^{(n+1)}(\eta)\pi_{n+1}(x)$ has exactly $n + 1$ sign-changes in the open interval (a, b) for any $\eta \in (a, b)$. Thus, according to (8.9), the Lagrange interpolation polynomial p of degree n for the function f, with interpolation points ξ_j, $j = 0, 1, \ldots, n$, contained in the open interval (a, b), is a near-minimax polynomial from $\mathcal{P}_n$ for f on $[a, b]$.

We have therefore just shown that if $f^{(n+1)}$ exists and is continuous on the closed interval $[a, b]$, and has the same sign on the open interval

so we obtain a polynomial approximation p (x) by taking the terms of this series up to the one involving x . Then, clearly,

$$e \quad p(x) = \sum^{\infty} \frac{k\,x}{r!}.$$

Over the interval [0, 1], for example, this difference is nonnegative and monotonic increasing; it does not change sign at all. Hence the polynomial p & thus constructed is quite certainly not a near-minimax approximation for x e on [0, 1]. Nevertheless, max □ e p (x) can be made arbitrarily small by choosing n sufficiently large.

8.6 Notes

For further details on the topics presented in this chapter, we refer to

+ $ (" *Approximation Theory and Methods*, Cambridge University Press, Cambridge, 1996.

The Weierstrass Theorem is discussed in Chapter 6 of that book, and is stated in its Theorem 6.3. Although the proof presented by Powell uses the Bernstein polynomials, it is different from the more elementary but slightly lengthier argument proposed in Exercise 12 here: it relies on a proof of Bohman and Korovkin based on properties of monotone operators; see, also, p. 66 in Chapter 3 of

/ " *Introduction to Approximation Theory,* McGraw–Hill, New York, 1966.

The notes contained on pp. 224–233 of Cheney's book are particularly illuminating.

The proof of the Weierstrass Theorem as proposed in Exercise 12, including the definition of what we today call Bernstein polynomials, stem from a paper of Sergei Natanovich Bernstein (1880–1968), entitled 'Démonstration du théorème de Weierstrass fondée sur le calcul des probabilités', *Comm. Soc. Math. Kharkow* **13**, 1–2, 1912/13.

Weierstrass' main contributions to approximation theory, as well as those of other mathematicians (including Picard, Volterra, Runge, Lebesgue, Mittag-Leffler, Fejér, Landau, de la Vallée Poussin, Bernstein), are reviewed in the extensive historical survey by Allan Pinkus, Weierstrass and approximation theory, *J. Approx. Theory* **107**, 1–66, 2000. Further details about the history of the subject can be found at

the history of approximation theory website maintained by Allan Pinkus and Carl de Boor: ˜ ./0

The second part of Theorem 8.1 concerning the approximability of a continuous function by polynomials in the 2-norm is not usually presented as part of the classical Weierstrass Theorem which is posed in the -norm. Here, we have chosen to state these results together in order to highlight the analogy, as well as to motivate the use of the 2-norm in polynomial approximation in the next chapter, Chapter 9.

In both Cheney's and Powell's books minimax approximation is treated in the more general framework of Haar systems. An $(n+1)$-dimensional linear subspace (of C[a, b] is said to satisfy the *Haar condition* if, for every nonzero p in (, the number of roots of the equation $p(x) = 0$ in the interval [a, b] is less than $n + 1$. The concept of Haar system is due to Alfred Haar (1885–1933), Die Minkowskische Geometrie und die Annäherung an stetige Funktionen, *Math. Ann.* **78**, 294–311, 1918; this paper contains Haar's Theorem which characterises finite-dimensional Haar systems in spaces of continuous functions. The *Characterisation Theorem*, formulated as Theorem 7.2 in Powell's book, shows that the Oscillation Theorem, Theorem 8.4 of the present chapter, remains valid in a more general setting when the set of polynomials $1, x, \ldots, x$ is replaced by an $(n + 1)$-dimensional Haar system of functions contained in C[a, b].

Exercises

8.1 Give a proof of Lemma 8.1.

8.2 Suppose that the real-valued function f is continuous and even on the interval [a, a], that is, $f(x) = f(\ x)$ for all x [a, a]. By using the Uniqueness Theorem, or otherwise, show that the minimax polynomial approximation of degree n is an even function. Deduce that the minimax polynomial approximation of degree 2n is also the minimax polynomial approximation of degree $2n+1$. What does this imply about the sequence of critical points for the minimax polynomial p ?

8.3 State and prove similar results to those in Exercise 2, for the case where f is an odd function, that is, $f(x) = \ f(\ x)$ for all x [a, a].

8.4 (i) Construct the minimax polynomial p & on the interval [1, 1] for the function g defined by $g(x) = \sin x$.

(ii) Construct the minimax polynomial p & on the interval $[\ 1, 1]$ for the function h defined by $h(x) = \cos x$.
(Use the results of Exercises 2 and 3.)

8.5 The function H is defined by $H(x) = 1$ if $x > 0$, $H(x) = \ 1$ if $x < 0$, and $H(0) = 0$. Show that for any $n \quad 0$ and any $p \quad \&$, $H \quad p \quad \infty \quad 1$ on the interval $[\ 1, 1]$. Construct the polynomial, of degree 0, of best approximation to H on the interval $[\ 1, 1]$, and show that it is unique. (Note that since H is discontinuous most of the theorems in this chapter are not applicable.)

Show that the polynomial of best approximation, of degree 1, to H on $[\ 1, 1]$ is not unique, and give an expression for its most general form.

8.6 Suppose that $t < t < \quad < t$ are k distinct points in the interval $[a, b]$; for any function f defined on $[a, b]$, write $Z\ (f) = \max \quad f(x)$. Explain why $Z\ (\ )$ is not a norm on the space of functions which are continuous on $[a, b]$; show that it is a norm on the space of polynomials of degree n, provided that $k > n$.

In the case $k = 3$, with $t \ = 0,\ t \ = -,\ t \ = 1$, where we wish to approximate the function $f : x \quad e$ on the interval $[0, 1]$, explain graphically, or otherwise, why the polynomial p of degree 1 which minimises $Z\ (f \quad p\ )$ satisfies the conditions

$$f(0) \quad p\ (0) = \quad [f(-) \quad p\ ](-) = f(1) \quad p\ (1).$$

Hence construct this polynomial p . Now suppose that $k = 4$, with $t \ = 0,\ t \ = -,\ t \ = -,\ t \ = -$; use a similar method to construct the polynomial of degree 1 which minimises $Z\ (f \quad p\ )$.

8.7 Among all polynomials $p \quad \&$ of the form

$$p\ (x) = Ax \ + \quad \bar{a}\ x\ ,$$

where A is a fixed nonzero real number, find the polynomial of best approximation for the function $f(x) \quad 0$ on the closed interval $[\ 1, 1]$.

8.8 Find the minimax polynomial $p \quad \&$ on the interval $[\ 1, 1]$ for the function f defined by

$$f(x) = \quad a\ x\ ,$$

where $a = 0$.

8.9 Construct the minimax polynomial p & on the interval $[1, 2]$ for the function f defined by $f(x) = x$.

8.10 Give a proof of Lemma 8.2.

8.11 Give an example of a continuous real-valued function f defined on the closed interval $[a, b]$ such that the set of critical points for the minimax approximation of f by polynomials from & does not contain either of the points a and b.

8.12 For each nonnegative integer n, and $x \in [0, 1]$, define the Bernstein polynomials p & by

$$p (x) = \frac{n!}{k!(n k)!} x (1 x) , \quad k = 0, \ldots, n .$$

Show that

$$(1 x + tx) = p (x)t ;$$

by differentiating this relation successively with respect to t and putting $t = 1$, show that, for any $x \in [0, 1]$,

$$p (x) = 1 ,$$

$$kp (x) = nx ,$$

$$k(k 1)p (x) = n(n 1)x ,$$

and deduce that

$$(x k/n) p (x) = \frac{x(1 x)}{n} , \qquad x [0, 1] .$$

Define M to be the upper bound of $f(x)$ on $[0, 1]$. Given $\varepsilon > 0$, we can choose $\delta > 0$ such that $f(x) f(y) < \varepsilon/2$ for any x and y in $[0, 1]$ such that $x y < \delta$. Now define the polynomial p & by

$$p (x) = f(k/n)p (x) ,$$

and choose a fixed value of x in $[0,1]$; show that

$$f(x) \quad p\ (x) \qquad f(x) \quad f(k/n)\, p \quad (x).$$

Using the notation

$$= \quad +$$

where denotes the sum over those values of k for which $x \quad k/n < \delta$, and denotes the sum over those values of k for which $x \quad k/n \quad \delta$, show that

$$f(x) \quad f(k/n)\, p \quad (x) < \varepsilon/2.$$

Show also that

$$f(x) \quad f(k/n)\, p \quad (x) \quad (2M/\delta\ ) \quad (x \quad k/n)\ p \quad (x).$$

Now, choose $N = M/(\delta\ \varepsilon)$, and show that

$$f(x) \quad p\ (x) < \varepsilon \qquad x \quad [0,1],$$

if $n \quad N$. Deduce that

$$f \quad p \quad _{\infty} < \varepsilon, \quad \text{if } n \quad N,$$

where $_{\infty}$ denotes the -norm on the interval $[0,1]$.

9

Approximation in the 2-norm

9.1 Introduction

In Chapter 8 we discussed the idea of best approximation of a continuous real-valued function by polynomials of some fixed degree in the -norm. Here we consider the analogous problem of best approximation in the 2-norm. Why, you might ask, is it necessary to consider best approximation in the 2-norm when we have already developed a perfectly adequate theory of best approximation in the -norm? As our first example in Section 9.3 will demonstrate, the choice of norm can significantly influence the outcome of the problem of best approximation: the polynomial of best approximation of a certain fixed degree to a given continuous function in one norm need not bear any resemblance to the polynomial of best approximation of the same degree in another norm. Ultimately, in a practical situation, the choice of norm will be governed by the sense in which the given continuous function has to be well approximated.

As will become apparent, best approximation in the 2-norm is closely related to the notion of orthogonality and this in turn relies on the concept of *inner product*. Thus, we begin the chapter by recalling from linear algebra the definition of *inner product space*.

Throughout the chapter $[a, b]$ will denote a nonempty, bounded, closed interval of the real line, and (a, b) will signify a nonempty bounded open interval of the real line.

Now,

$$\begin{aligned} , \quad ! \quad &= \\ &= \qquad (\cos \alpha \cos \beta + \sin \alpha \sin \beta) \\ &= \qquad \cos(\alpha \quad \beta) \\ &= \qquad \cos(\vartheta \, " \,) , \end{aligned}$$

where $\vartheta \, " \, = \alpha \quad \beta$ *is the angle between the vectors* *and* *. The vector* *is orthogonal to* *if, and only if,* $\vartheta \, "$ *is* $\pi/2$ *or* $3\pi/2$*; either way,* $\cos(\vartheta \, " \,) = 0$*, and hence* $\quad , \quad ! = 0$*. We note in passing that if* $\quad = \quad ,$ *then* $\vartheta \, " \, = 0$ *and therefore*

$$, \quad ! = \qquad .$$

This last observation motivates our next definition.

Definition 9.3 *Suppose that* *is an inner product space over the field of real numbers, with inner product* $\quad , \quad !$*. For* f *in* *, we define the*

$$\mathsf{f} \quad = \quad \mathsf{f}, \mathsf{f} \, ! \quad . \tag{9.1}$$

Although our terminology and our notation appear to imply that (9.1) defines a norm on , this is by no means obvious. In order to show that $\mathsf{f} \qquad \mathsf{f}, \mathsf{f} \, !$ is indeed a norm, we begin with the following result which is a direct generalisation of the Cauchy–Schwarz inequality (2.35) from Chapter 2.

Lemma 9.1 (Cauchy–Schwarz inequality)

$$\mathsf{f}, \mathsf{g} \, ! \qquad \mathsf{f} \quad \mathsf{g} \qquad \mathsf{f}, \mathsf{g} \qquad . \tag{9.2}$$

Proof The proof is analogous to that of (2.35). Recalling the definition of from (9.1) and noting the first three axioms of inner product, we find that, for f, g ,

$$0 \qquad \lambda \mathsf{f} + \mathsf{g} \quad = \lambda \quad \mathsf{f} \quad + 2\lambda \; \mathsf{f}, \mathsf{g} \, ! + \quad \mathsf{g} \qquad \lambda \qquad . \tag{9.3}$$

Denoting, for f, g fixed, the quadratic polynomial in λ on the right-hand side by $\mathsf{A}(\lambda)$, the condition for $\mathsf{A}(\lambda)$ to be nonnegative for all λ in is that $[2 \; \mathsf{f}, \mathsf{g} \, !] \quad 4 \; \mathsf{f} \quad \mathsf{g} \qquad 0$; this gives the inequality (9.2). □

Now, putting $\lambda = 1$ in (9.3) and using (9.2) on the right yields

$$\mathsf{f} + \mathsf{g} \qquad \mathsf{f} \quad + \quad \mathsf{g} \qquad \mathsf{f}, \mathsf{g} \qquad .$$

Consequently, obeys the triangle inequality, the third axiom of norm. The first two axioms of norm, namely that

$f \quad 0$ for all f , and $f = 0$ if, and only if, $f = 0$ in , and
$\lambda f = \lambda f$ for all λ and all f ,

follow directly form (9.1) and from the last three axioms of inner product stated in Definition 9.1.

We have thus shown the following result.

Theorem 9.1 *An inner product space over the field of real numbers, equipped with the induced norm , is a normed linear space over .*

We conclude this section with a relevant example of an inner product space, whose induced norm is the 2-norm considered at the beginning of Chapter 8.

Example 9.3 *The set* $C[a,b]$ *of continuous real-valued functions defined on the closed interval* $[a,b]$ *is an inner product space with*

$$\langle f, g \rangle = \int_a^b w(x) f(x) g(x) dx, \tag{9.4}$$

where w *is a* **weight function***, defined, positive, continuous and integrable on the open interval* (a,b)*. The norm , induced by this inner product and given by*

$$\|f\| = \left(\int_a^b w(x) f(x) \, dx \right)^{1/2}, \tag{9.5}$$

is referred to as the 2*-norm on* $C[a,b]$ *(see Example 8.2). For the sake of simplicity, we have chosen not to distinguish in terms of our notation between the* 2*-norm on* $C[a,b]$ *defined above and the* 2*-norm for vectors introduced in Chapter 2; it will always be clear from the context which of the two is intended.*

Clearly, it is not necessary to demand the continuity of the function f on the closed interval $[a,b]$ to ensure that $\|f\|$ is finite. For example, $f: x \quad \text{sgn}\, x \quad -(a+b)$, $x \quad [a,b]$, has finite 2-norm, despite the fact that it has a jump discontinuity at $x = -(a+b)$.

Motivated by this observation, and the desire to develop a theory of approximation in the 2-norm whose range of applicability extends beyond the linear space of continuous functions on a bounded closed interval, we denote by $L_{\cdot}(a,b)$ the set of all real-valued functions f

defined on (a,b) such that $w(x)\,f(x)$ is integrable on (a,b); the set $L_2(a,b)$ is equipped with the inner product (9.4) and the induced 2-norm (9.5). Obviously, $C[a,b]$ is a proper subset of $L_2(a,b)$.

In this broader context, is frequently referred to as the L_2-norm; for the sake of simplicity we shall continue to call it the 2-norm. As before, w is assumed to be a real-valued function, defined, positive, continuous and integrable on the open interval (a,b). When $w(x) \equiv 1$ on (a,b), we shall write $L_2(a,b)$ instead of $L_2(a,b)$.

We are now ready to consider best approximation in the 2-norm.

9.3 Best approximation in the 2-norm

The problem of best approximation in the 2-norm can be formulated as follows:

(B) Given that $f \in L_2(a,b)$, find $p \in \mathbb{P}$ such that

$$\|f - p\| = \inf_{q \in \mathbb{P}} \|f - q\|;$$

such p is called a **polynomial of best approximation of degree n to the function f in the 2-norm on (a,b)**.

The existence and uniqueness of p will be shown in Theorem 9.2. However, we shall first consider some simple examples.

Example 9.4 *Suppose that $\varepsilon > 0$ and let $f(x) = 1 - e^{-x/\varepsilon}$ with x in $[0,1]$. For $\varepsilon = 10^{-3}$, the function f is depicted in Figure 9.1. We shall construct the polynomial of best approximation of degree 0 in the 2-norm, with weight function $w(x) \equiv 1$, for f on $(0,1)$, and compare it with the minimax polynomial of degree 0 for f on $[0,1]$.*

The best approximation to f by a polynomial of degree 0 in the 2-norm on the interval $(0,1)$, with weight function $w(x) \equiv 1$, is determined by minimising $\|f - c\|$ over all $c \in \mathbb{R}$; equivalently, we need to minimise

$$\int_0^1 (f(x) - c)^2\,dx = \int_0^1 f(x)^2\,dx - 2c\int_0^1 f(x)\,dx + c^2$$

which tends to as ε 0+. These examples indicate that the polynomial of best approximation from &

where

$$M = \int_0^1 x^{i+j}\,dx = \frac{1}{k+j+1},$$
$$b = \int_0^1 f(x)x^i\,dx.$$

Equivalently, recalling that the inner product associated with the 2-norm (in the case of $w(x) \equiv 1$) is defined by

$$\langle g, h\rangle = \int_0^1 g(x)h(x)dx,$$

M and b can be written as

$$M = \langle x^j, x^i\rangle, \qquad b = \langle f, x^i\rangle. \tag{9.7}$$

By solving the system of linear equations (9.6) for $c_0, \ldots, c_n$, we obtain the coefficients of the polynomial of best approximation of degree n to the function f in the 2-norm on the interval $(0, 1)$. We can proceed in the same manner on any interval (a, b) with any positive, continuous and integrable weight function w defined on (a, b).

This approach is straightforward for small values of n, but soon becomes impractical as n increases. The source of the computational difficulties is the fact that the matrix M is the Hilbert matrix, discussed in Section 2.8. The Hilbert matrix is well known to be ill-conditioned for large n, so any solution to (9.6), computed with a fixed number of decimal digits, loses all accuracy due to accumulation of rounding errors. Fortunately, an alternative method is available, and is discussed in the next section.

9.4 Orthogonal polynomials

In the previous section we described a method for constructing the polynomial of best approximation $p \in \mathcal{P}_n$ to a function f in the 2-norm; it was based on seeking p as a linear combination of the polynomials x^j, $j = 0, \ldots, n$, which form a basis for the linear space $\mathcal{P}_n$. The approach was not entirely satisfactory because it gave rise to a system of linear equations with a full matrix that was difficult to invert. The central idea of the alternative approach that will be described in this section is to expand p in terms of a different basis, chosen so that the resulting system of linear equations has a diagonal matrix; solving this

linear system is then a trivial exercise. Of course, the nontrivial ingredient of this alternative approach is to find a suitable basis for $\mathcal{P}$ which achieves the objective that the matrix of the linear system is diagonal. The expression for M in (9.7) gives us a clue how to proceed.

Suppose that ϕ_j, $j = 0, \ldots, n$, form a basis for $\mathcal{P}$, $n \geq 0$; let us seek the polynomial of best approximation as

$$p(x) = \gamma_0 \phi_0(x) + \cdots + \gamma_n \phi_n(x),$$

where $\gamma_0, \ldots, \gamma_n$ are real numbers to be determined. By the same process as in the previous section, we arrive at a system of linear equations of the form (9.6):

$$M\gamma = \beta, \qquad j = 0, \ldots, n,$$

where now

$$M = \langle \phi_j, \phi_k \rangle \qquad \text{and} \qquad \beta = \langle f, \phi_j \rangle,$$

with the inner product $\langle \cdot, \cdot \rangle$ defined by

$$\langle g, h \rangle = \int_a^b w(x) g(x) h(x)\, dx,$$

and the weight function w assumed to be positive, continuous and integrable on the interval (a, b).

Thus, $M = (M_{jk})$ will be a diagonal matrix provided that the basis functions ϕ_j, $j = 0, \ldots, n$, for the linear space $\mathcal{P}$ are chosen so that $\langle \phi_j, \phi_k \rangle = 0$, for $j \neq k$; in other words, ϕ_j is required to be orthogonal to ϕ_k for $j \neq k$, in the sense of Definition 9.2. This observation motivates the following definition.

Definition 9.4 *Given a weight function w, defined, positive, continuous and integrable on the interval (a, b), we say that the sequence of polynomials ϕ_j, $j = 0, 1, \ldots$, is a* **system of orthogonal polynomials** *on the interval (a, b) with respect to w, if each ϕ_j is of exact degree j, and if*

$$\int_a^b w(x)\phi_j(x)\phi_k(x)\,dx \begin{cases} \neq 0 & \text{for all } k = j, \\ = 0 & \text{when } k \neq j. \end{cases}$$

Next, we show that a system of orthogonal polynomials exists on any interval (a, b) and for any weight function w which satisfies the conditions in Definition 9.4. We proceed inductively.

Let $\phi \ (x)$ 1, and suppose that ϕ has already been constructed for $j = 0, \dots, n$, with n 0. Then,

$$! \quad w(x)\phi \ (x)\phi \ (x)\mathrm{d}x = 0 , \qquad k \quad 0, \dots, n \qquad j \ .$$

Let us now define the polynomial

$$q(x) = x \qquad a \ \phi \ (x) \qquad a \ \phi \ (x) ,$$

where

$$a = \frac{' \ w(x) \, x \quad \phi \ (x)\mathrm{d}x}{' \ w(x)[\phi \ (x)] \ \mathrm{d}x} , \qquad j = 0, \dots, n .$$

It then follows that

$$\begin{aligned} ! \ w(x)q(x)\phi \ (x)\mathrm{d}x &= ! \ w(x)x \quad \phi \ (x)\mathrm{d}x \\ &\quad a \quad ! \ w(x)[\phi \ (x)] \ \mathrm{d}x \\ &= 0 \qquad \text{for } 0 \quad j \quad n , \end{aligned}$$

where we have used the orthogonality of the sequence ϕ , $j = 0, \dots, n$. Thus, with this choice of the numbers a we have ensured that q is orthogonal to all the previous members of the sequence, and ϕ can now be defined as any nonzero-constant multiple of q. This procedure for constructing a system of orthogonal polynomials is usually referred to as **Gram–Schmidt orthogonalisation**.

Example 9.5 *We shall construct a system of orthogonal polynomials* $\phi \ , \phi \ , \phi$ *on the interval* $(0, 1)$ *with respect to the weight function* $w(x)$ 1.

We put $\phi \ (x)$ 1, and we seek ϕ in the form

$$\phi \ (x) = x \quad c \ \phi \ (x)$$

such that $\phi \ , \phi \ ! = 0$; that is,

$$x, \phi \ ! \quad c \quad \phi \ , \phi \ ! = 0 .$$

[1] KX F B & K #.> / " C @ % #@#? $ / "DL , B#* K 6 #.&? / B D C ? / , #@>@ F 6D4

Hence,

$$c = \frac{x, \phi}{\phi, \phi} = -$$

and therefore,

$$\phi(x) = x - \phi(x) = x -$$

By construction, $\phi, \phi = \phi, \phi = 0$.

We now seek ϕ in the form

$$\phi(x) = x (d\phi(x) + d\phi(x))$$

such that $\phi, \phi = 0$ and $\phi, \phi = 0$. Thus,

$$\begin{aligned} x, \phi \; d \phi, \phi \; d \phi, \phi &= 0, \\ x, \phi \; d \phi, \phi \; d \phi, \phi &= 0. \end{aligned}$$

As $\phi, \phi = 0$ and $\phi, \phi = 0$, we have that

$$\begin{aligned} d &= \frac{x, \phi}{\phi, \phi} = 1, \\ d &= \frac{x, \phi}{\phi, \phi} = -, \end{aligned}$$

and therefore

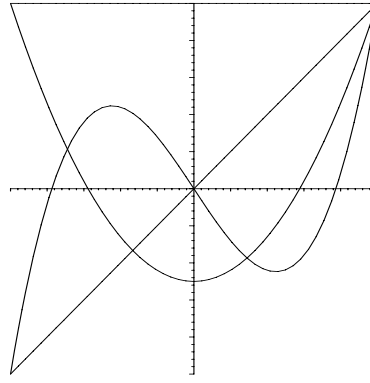
$$\phi(x) = x \quad x + -. \tag{9.8}$$

Clearly, $\phi, \phi = 0$ for $j = k$, $j, k \quad 0, 1, 2$, and ϕ is of exact degree j, $j = 0, 1, 2$. Thus we have found the required system ϕ, ϕ, ϕ of orthogonal polynomials on the interval $(0, 1)$ with respect to the given weight function w.

By continuing this procedure, we can construct a system of orthogonal polynomials $\phi, \phi, \ldots, \phi$, with respect to the weight function $w(x) \quad 1$ on the interval $(0, 1)$, for any $n \quad 1$. For example, when $n = 3$, we shall find ϕ, ϕ, ϕ, ϕ, with ϕ, ϕ, ϕ, as above, and

$$\phi(x) = x \quad -x + -x \quad -.$$

Having generated a system of orthogonal polynomials on the interval $(0, 1)$ with respect to the weight function $w(x) \quad 1$, by performing the linear mapping $x \quad (b \quad a)x + a$ we may obtain a system of orthogonal polynomials on any open interval (a, b) with respect to the weight function $w(x) \quad 1$. For example, when $(a, b) = (\ 1, 1)$, the mapping $x \quad 2x \quad 1$ leads to the system of Legendre polynomials on $(\ 1, 1)$.



$ 5 3 B * +

Example 9.6 (Legendre polynomials) *We wish to construct a system of orthogonal polynomials on* $(a, b) = (-1, 1)$ *with respect to the weight function* $w(x) \equiv 1$.

On replacing x by

$$\frac{x - a}{b - a} = -(x + 1), \qquad x \in (a, b) = (-1, 1),$$

in $\phi_2(x)$, $\phi_3(x)$, $\phi_4(x)$, $\phi_5(x)$ from Example 9.5, we obtain, on normalising each of these polynomials so that its value at $x = 1$ is equal to 1, the polynomials $\phi_0, \phi_1, \phi_2, \phi_3$, defined by

$$\begin{aligned}
\phi_0(x) &= 1, \\
\phi_1(x) &= x, \\
\phi_2(x) &= \tfrac{1}{2}x^2 - \tfrac{1}{2}, \\
\phi_3(x) &= \tfrac{1}{2}x^3 - \tfrac{3}{2}x.
\end{aligned}$$

These are the first four elements of the system of Legendre polynomials, orthogonal on the interval $(-1, 1)$ with respect to the weight function $w(x) \equiv 1$. They are depicted in Figure 9.2. An alternative normalisation would have been to divide each ϕ_n by $\|\phi_n\|$ so as to ensure that the 2-norm of the resulting scaled polynomial is equal to 1.

Example 9.7 *The Chebyshev polynomials* $T_n : x \mapsto \cos(n \cos^{-1} x)$, $n = 0, 1, \ldots$, *introduced in Section 8.4, form an orthogonal system on the interval* $(-1, 1)$ *with respect to the positive, continuous and integrable weight function* $w(x) = (1 - x^2)^{-\frac{1}{2}}$.

The proof of this is simple: let , ! denote the inner product in L. (1, 1)

Such a system of polynomials is said to be **orthonormal**. The polynomials ψ_j, $j = 0, \ldots, n$, are linearly independent and form a basis for the linear space S; therefore, each element $q \in S$ can be expressed as a suitable linear combination,

$$q(x) = \beta_0 \psi_0(x) + \cdots + \beta_n \psi_n(x).$$

We wish to choose β_j, $j = 0, \ldots, n$, so as to ensure that the corresponding polynomial q minimises $\|f - q\|$ over all $q \in S$. Let us, therefore, consider the function $E\colon (\beta_0, \ldots, \beta_n) \to E(\beta_0, \ldots, \beta_n)$ defined by $E(\beta_0, \ldots, \beta_n) = \|f - q\|^2$, where $q(x) = \beta_0 \psi_0(x) + \cdots + \beta_n \psi_n(x)$. Then,

$$\begin{aligned} E(\beta_0, \ldots, \beta_n) &= \langle f - q, f - q \rangle \\ &= \langle f, f \rangle - 2\langle f, q \rangle + \langle q, q \rangle \\ &= \|f\|^2 - 2\sum_{j=0}^{n} \beta_j \langle f, \psi_j \rangle + \sum_{j=0}^{n}\sum_{k=0}^{n} \beta_j \beta_k \langle \psi_j, \psi_k \rangle \\ &= \|f\|^2 - 2\sum_{j=0}^{n} \beta_j \langle f, \psi_j \rangle + \sum_{j=0}^{n} \beta_j^2 \\ &= \sum_{j=0}^{n} [\beta_j - \langle f, \psi_j \rangle]^2 + \|f\|^2 - \sum_{j=0}^{n} \langle f, \psi_j \rangle^2. \end{aligned}$$

The function $(\beta_0, \ldots, \beta_n) \to E(\beta_0, \ldots, \beta_n)$ achieves its minimum value at $(\beta_0^*, \ldots, \beta_n^*)$, where

$$\beta_j^* = \langle f, \psi_j \rangle, \quad j = 0, \ldots, n.$$

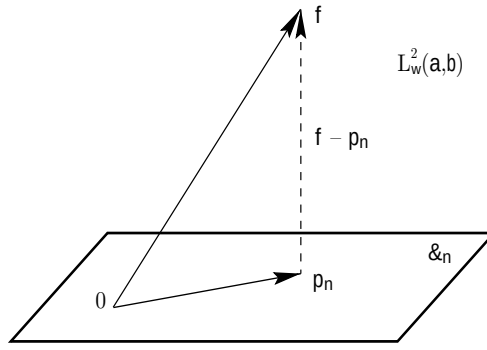
Hence $p \in S$ defined by

$$p(x) = \beta_0^* \psi_0(x) + \cdots + \beta_n^* \psi_n(x)$$

is the unique polynomial of best approximation of degree n to the function $f \in L_2(a, b)$ in the 2-norm on the interval (a, b). □

Remark 9.1 *As* $E(\beta_0^*, \ldots, \beta_n^*) = \|f - p\|^2 \geq 0$, *it follows from the proof of Theorem 9.2 that if* $f \in L_2(a, b)$, *and* $\psi_0, \psi_1, \ldots$ *is an orthonormal system of polynomials in* $L_2(a, b)$, *then*

$$\sum_{j=0}^{\infty} \langle f, \psi_j \rangle^2 \leq \|f\|^2.$$



$\$ \& \ 7 \qquad B \quad B \qquad \$ \ ) \ \# \qquad)$
$! \qquad \$ \qquad B \quad , \quad !$
$3^2 * \quad + \quad " \qquad \$ \qquad 3 \quad ^2 * \quad +$
$5 \qquad :$
$* \qquad B \quad 3 \quad ^2 * \quad ++ \quad B$
$B : \qquad * \qquad +$

for each $n \quad 0$. *This result is known as* **Bessel's inequality**.

The next theorem, in conjunction with the use of orthogonal polynomials, will be our key tool for constructing the polynomial of best approximation in the 2-norm.

Theorem 9.3 *A polynomial* $p \quad \mathcal{S}$ *is the polynomial of best approximation of degree* n *to a function* $f \quad L_{\cdot}(a,b)$ *in the* 2*-norm if, and only if, the difference* $f \quad p$ *is orthogonal to every element of* $\mathcal{S}$, *i.e.,*

$$f \quad p \ , q! = 0 \qquad q \quad \mathcal{S} \ . \tag{9.9}$$

A geometrical illustration of the property (9.9) is given in Figure 9.3.

Proof of theorem Suppose that (9.9) holds. Then,

$$f \quad p \ , p \quad q! = 0 \qquad q \quad \mathcal{S} \ ,$$

given that $p \quad q \quad \mathcal{S}$ for each q in $\mathcal{S}$. Therefore,

$$\begin{aligned} f \quad p \qquad &= \quad f \quad p \ , f \quad p \ ! \\ &= \quad f \quad p \ , f \quad q! + f \quad p \ , q \quad p \ ! \\ &= \quad f \quad p \ , f \quad q! \qquad q \quad \mathcal{S} \ . \end{aligned}$$

Hence, by the Cauchy–Schwarz inequality (9.2),

$$f \quad p \qquad f \quad p \qquad f \quad q \qquad q \quad \& \ .$$

This implies that

$$f \quad p \qquad f \quad q \qquad q \quad \& \ .$$

On choosing $q = p$ on the right-hand side, equality will hold and therefore

$$f \quad p \quad = \min_{\square P} \quad f \quad q \quad .$$

Conversely, suppose that p is the polynomial of best approximation to $f \quad L_{\cdot}(a,b)$. We have seen in the proof of Theorem 9.2 that p can be written in terms of the orthonormal polynomials $\psi \ , \ k = 0, \ldots, n$, as

$$p \ (x) = \beta^{\square}\psi \ (x) + \quad + \beta^{\square}\psi \ (x) \, ,$$

where

$$\beta^{\square} = \ f, \psi \ ! \, , \qquad k = 0, \ldots, n \, . \tag{9.10}$$

On recalling that $\psi \ , \psi \ ! = \delta \ , \ j,k \qquad 0, \ldots, n$, where δ is the Kronecker delta, we deduce from (9.10) that

$$\begin{aligned} f \quad p \ , \psi \ ! \ &= \quad f, \psi \ ! \qquad \beta^{\square} \ \psi \ , \psi \ ! \\ &= \quad f, \psi \ ! \qquad \beta^{\square}\delta \\ &= \quad f, \psi \ ! \quad \beta^{\square} = 0 \, , \qquad j = 0, \ldots, n \, . \end{aligned} \tag{9.11}$$

Since $\& \ = \text{span} \ \psi \ , \ldots, \psi$, it follows from (9.11) that $f \quad p \ , q ! = 0$ for all $q \quad \&$, as required. □

An equivalent, but slightly more explicit, form of writing (9.9) is

$$\int w(x)(f(x) \quad p \ (x))q(x) \, dx = 0 \qquad q \quad \& \ .$$

Theorem 9.2 provides a simple method for determining the polynomial of best approximation $p \qquad \&$ to a function $f \quad L_{\cdot}(a,b)$ in the 2-norm. First, proceeding as described in the discussion following Definition 9.4, we construct the system of orthogonal polynomials $\phi \ , \ j = 0, \ldots, n$, on the interval (a,b) with respect to the weight function w, if this system

is not already known. We normalise the polynomials ϕ_j, $j = 0, \ldots, n$, by setting

$$\psi_j = \frac{\phi_j}{\|\phi_j\|}, \qquad j = 0, \ldots, n,$$

to obtain the system of orthonormal polynomials ψ_j, $j = 0, \ldots, n$, on (a, b). We then evaluate the coefficients $\beta_j = \langle f, \psi_j \rangle$, $j = 0, \ldots, n$, and form $p_n(x) = \beta_0 \psi_0(x) + \cdots + \beta_n \psi_n(x)$.

We may avoid the necessity of determining the normalised polynomials ψ_j by writing

$$\begin{aligned} p_n(x) &= \beta_0 \psi_0(x) + \cdots + \beta_n \psi_n(x) \\ &= \beta_0 \frac{\phi_0}{\|\phi_0\|} \frac{1}{\|\phi_0\|} \phi_0(x) + \cdots + \beta_n \frac{\phi_n}{\|\phi_n\|} \frac{1}{\|\phi_n\|} \phi_n(x) \\ &= \gamma_0 \phi_0(x) + \cdots + \gamma_n \phi_n(x), \end{aligned} \tag{9.12}$$

where

$$\gamma_j = \frac{\langle f, \phi_j \rangle}{\langle \phi_j, \phi_j \rangle}, \qquad j = 0, \ldots, n. \tag{9.13}$$

Thus, as indicated at the beginning of the section, with this approach to the construction of the polynomial of best approximation in the 2-norm, we obtain the coefficients γ_j explicitly and there is no need to solve a system of linear equations with a full matrix.

Example 9.8 *We shall construct the polynomial of best approximation of degree* 2 *in the* 2*-norm to the function* $f(x) = e^x$ *over* $(0, 1)$ *with weight function* $w(x) \equiv 1$.

We already know a system of orthogonal polynomials ϕ_0, ϕ_1, ϕ_2 on this interval from Example 9.5; thus, we seek $p_2 = \gamma_0 \phi_0 + \gamma_1 \phi_1 + \gamma_2 \phi_2$ in the form

$$p_2(x) = \gamma_0 \phi_0(x) + \gamma_1 \phi_1(x) + \gamma_2 \phi_2(x), \tag{9.14}$$

where, according to (9.13),

$$\gamma_j = \frac{\int_0^1 e^x \phi_j(x) dx}{\int_0^1 \phi_j(x) dx}, \qquad j = 0, 1, 2.$$

Recalling from Example 9.5 that

$$\phi_0(x) \equiv 1, \quad \phi_1(x) = x - \tfrac{1}{2}, \quad \phi_2(x) = x^2 - x + \tfrac{1}{6},$$

we then have that

$$\begin{aligned} \gamma &= \frac{e \quad 1}{1} = e \quad 1, \\ \gamma &= \frac{3/2 \quad e/2}{1/12} = 18 \quad 6e, \\ \gamma &= \frac{7e/6 \quad 19/6}{1/180} = 210e \quad 570. \end{aligned} \tag{9.15}$$

Substituting the values of γ , γ and γ into (9.14), we conclude that the polynomial of best approximation of degree 2 for the function $f : x \quad e$ in the 2-norm is

$$p (x) = (210e \quad 570)x + (588 \quad 216e)x + (39e \quad 105).$$

The approximation error is

$$f \quad p = 0.005431,$$

to six decimal digits.

We conclude this section by giving a property of orthogonal polynomials that will be required in the next chapter.

Theorem 9.4 *Suppose that* ϕ , $j = 0, 1, \ldots,$ *is a system of orthogonal polynomials on the interval* (a, b) *with respect to the positive, continuous and integrable weight function* w *on* (a, b)*. It is understood that* ϕ *is a polynomial of exact degree* j*. Then, for* $j \quad 1$*, the zeros of the polynomial* ϕ *are real and distinct, and lie in the interval* (a, b)*.*

Proof Suppose that ξ , $i = 1, \ldots, k$, are the points in the open interval (a, b) at which $\phi (x)$ changes sign. Let us note that $k \quad 1$, because for $j \quad 1$, by orthogonality of $\phi (x)$ to $\phi (x) \quad 1$, we have that

$$\int w(x)\phi (x)dx = 0.$$

Thus, the integrand, being a continuous function that is not identically zero on (a, b), must change sign on (a, b); however, w is positive on (a, b), so ϕ must change sign at least once on (a, b). Therefore $k \quad 1$.

Let us define

$$\pi (x) = (x \quad \xi) \ldots (x \quad \xi). \tag{9.16}$$

Now the function $\phi (x)\pi (x)$ does not change sign in the interval (a, b), since at each point where $\phi (x)$ changes sign $\pi (x)$ changes sign also. Hence,

$$\int w(x)\phi (x)\pi (x)dx = 0.$$

However, ϕ is orthogonal to every polynomial of lower degree with respect to the weight function w, so the degree of the polynomial π must be at least j; thus, $k \geq j$. On the other hand, k cannot be greater than j, since a polynomial of exact degree j cannot change sign more than j times. Therefore $k = j$; *i.e.*, the points $\xi_i \in (a,b)$, $i = 1, \ldots, j$, are the zeros (and all the zeros) of $\phi(x)$. □

9.5 Comparisons

We can show that the polynomial of best approximation in the 2-norm for a function $f \in C[a,b]$ is also a near-best approximation in the ∞-norm for f on $[a,b]$ in the sense defined in Section 8.5.

Theorem 9.5 *Let $n \geq 0$ and assume that f is defined and continuous on the interval $[a,b]$, and $f \notin \mathcal{P}_n$. Let p_n be the polynomial of best approximation of degree n to f in the 2-norm on $[a,b]$, where the weight function w is positive, continuous and integrable on (a,b). Then, the difference $f - p_n$ changes sign at no less than $n+1$ distinct points in the interval (a,b).*

Proof The proof is very similar to that of Theorem 9.4; we shall give an outline and leave the details as an exercise.

As $\langle f - p_n, 1 \rangle = 0$, *i.e.*,

$$\int_a^b w(x)(f(x) - p_n(x))dx = 0\,,$$

and $w(x) > 0$ for all $x \in (a,b)$, it follows that $f - p_n$ changes sign in (a,b). Let ξ_j, $j = 1, \ldots, k$, denote distinct points in (a,b) where $f - p_n$ changes sign. We shall prove that $k \geq n+1$.

Define the polynomial $\pi(x)$ as in (9.16); then, $w(x)[f(x) - p_n(x)]\pi(x)$ does not change sign in (a,b), and so its integral over (a,b) is not zero. Therefore, $\langle f - p_n, \pi \rangle \neq 0$. On the other hand, according to Theorem 9.3, $f - p_n$ is orthogonal to every polynomial of degree n or less. Hence the degree of $\pi(x)$ must be greater than n, and so $k \geq n+1$. □

We return to the example illustrated by Figure 8.5, and consider the difference $f - p_n$ for the function $f: x \to e^x$ on the interval $(0,1)$. Figure 9.4 shows this difference for two polynomial approximations of degree 4: the minimax approximation of Section 8.5 and the best approximation in the 2-norm with weight function $w(x) \equiv 1$. It is clear that the

which gives greater weight near the ends of the interval, it seems likely that the extrema of the error might be more nearly equal. This can be achieved by using the weight function $w(x) = [x(1 \quad x)]^{-}$, so that the orthogonal polynomials are the Chebyshev polynomials adapted to the interval $(0, 1)$. Figure 9.5 shows the corresponding difference $f \quad p$, and we now see that the two best approximations, in the -norm and the weighted 2-norm, are very close.

Polynomials of best approximation in the 2-norm have a special property which is often useful. Suppose that we have constructed the best polynomial approximation, p , of degree n, in the 2-norm, but that p does not achieve the required accuracy. To construct the best polynomial approximation of degree $n + 1$ all we need is to calculate γ from

$$\gamma \quad = \frac{f \quad p \ , \phi \quad !}{\phi}$$

and then let $p \quad (x) = p \ (x) + \gamma \quad \phi \quad (x)$. By noting that

$$f \quad p \quad , \phi \ ! = 0 , \qquad j = 0, 1, \ldots , n + 1 ,$$

it follows that p is best least squares approximation to f from & . If we are constructing the minimax approximation of degree $n + 1$, or using Lagrange interpolation with equally spaced points, the work involved in constructing p is lost, and the construction of p must begin completely afresh.

9.6 Notes

We give some pointers to the vast literature on orthogonal polynomials. The following are classical sources on the subject.

1. " *Orthogonal Polynomials,* Pergamon Press, Oxford, New York, 1971.

$ 1 " *Orthogonal Polynomials,* Memoirs of the American Mathematical Society, no. 213, American Mathematical Society, Providence, RI, 1979.

1 &. 2" *Orthogonal Polynomials*, Colloquium publications (American Mathematical Society), 23, American Mathematical Society, Providence, RI, 1959.

Tables of orthogonal polynomials are found in

+) (. #) & (Editors), 'Orthogonal polynomials', Ch. 22 in *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables*, ninth printing, Dover, New York, pp. 771–802, 1972.

Computational aspects of the theory of orthogonal polynomials are discussed in the edited volume

" " ' (Editors), *Applications and Computation of Orthogonal Polynomials*, Conference at the Mathematical Research Institute, Oberwolfach, Germany, March 22–28, 1998, Birkhäuser, Basel, 1999.

A recent survey of the theory and application of orthogonal polynomials in numerical computations is contained in

" Orthogonal polynomials: applications and computation, *Acta Numerica* **5** (A. Iserles, ed.), Cambridge University Press, Cambridge, pp. 45–119, 1996.

Finally, we refer to the books of Powell and Cheney, cited in the Notes at the end of the previous chapter, concerning the application of orthogonal polynomials in the field of best least squares approximation.

Exercises

9.1 Construct orthogonal polynomials of degrees 0, 1 and 2 on the interval $(0, 1)$ with the weight function $w(x) = \ln x$.

9.2 Let the polynomials ϕ_j, $j = 0, 1, \ldots$, form an orthogonal system on the interval $(-1, 1)$ with respect to the weight function $w(x) \equiv 1$. Show that the polynomials $\phi_j((2x - a - b)/(b - a))$, $j = 0, 1, \ldots$, represent an orthogonal system for the interval (a, b) and the same weight function. Hence obtain the polynomials in Example 9.5 from the Legendre polynomials in Example 9.6.

9.3 Suppose that the polynomials ϕ_j, $j = 0, 1, \ldots$, form an orthogonal system on the interval $(0, 1)$ with respect to the weight function $w(x) = x^{\alpha}$, $\alpha > 0$. Find, in terms of ϕ_j, a system of orthogonal polynomials for the interval $(0, b)$ and the same weight function.

9.4 Show, by induction or otherwise, that, for $0 \leq k \leq n$,

$$\frac{d}{dx}\{(1-x^2)^n\} = (1-x^2)^{n-k} q_k(x),$$

where q_k is a polynomial of degree k. Deduce that all the derivatives of the function $(1-x^2)^n$ of order less than n vanish at $x = \pm 1$.

Define $\phi_n(x) = (d/dx)^n (1-x^2)^n$, and show by repeated integration by parts that

$$\int_{-1}^{1} \phi_n(x)\phi_j(x)dx = 0, \qquad 0 \leq j < k.$$

Hence verify the expressions in Example 9.6 for the Legendre polynomials of degrees 0, 1, 2 and 3.

9.5 Show, by induction or otherwise, that, for $0 \leq k \leq j$,

$$\frac{d}{dx}\{x^j e^{-x}\} = x^{j-k} q_k(x) e^{-x},$$

where $q_k(x)$ is a polynomial of degree k.

The function ϕ_j is defined for $j \geq 0$ by

$$\phi_j(x) = e^x \frac{d}{dx}(x^j e^{-x}).$$

Show that, for each $j \geq 0$, ϕ_j is a polynomial of degree j, and that these polynomials form an orthogonal system on the interval $(0, \infty)$ with respect to the weight function $w(x) = e^{-x}$. Write down the polynomials with $j = 0, 1, 2$ and 3.

9.6 Suppose that ϕ_j, $j = 0, 1, \ldots$, form a system of orthogonal polynomials with weight function $w(x)$ on the interval (a, b). Show that, for some value of the constant C, $\phi_{j+1}(x) - Cx\phi_j(x)$ is a polynomial of degree j, and hence that

$$\phi_{j+1}(x) - Cx\phi_j(x) = \sum_{k=0}^{j} \alpha_k \phi_k(x), \qquad \alpha_k = \ldots.$$

Use the orthogonality properties to show that $\alpha_k = 0$ for $k < j-1$, and deduce that the polynomials satisfy a recurrence relation of the form

$$\phi_{j+1}(x) - (Cx + D_j)\phi_j(x) + E_j \phi_{j-1}(x) = 0, \qquad j \geq 1.$$

9.7 In the notation of Exercise 6 suppose that the normalisation of the polynomials is so chosen that for each j the coefficient of x in $\phi_j(x)$ is positive. Show that $C_j > 0$ for all j. By considering

$$\int_a^b w(x)\phi_j(x)[\phi_j(x) - C_{j-1}x\phi_{j-1}(x)]\,dx$$

show that

$$\int_a^b w(x)x\phi_{j-1}(x)\phi_j(x)\,dx > 0,$$

and deduce that $E_j > 0$ for all j. Hence show that for all positive values of j the zeros of ϕ_j and ϕ_{j-1} interlace. (See the proof of Theorem 5.8.)

9.8 Using the weight function w on the interval (a, b) apply a similar argument to that for Theorem 8.6 to find the best polynomial approximation p_n of degree n in the 2-norm to the function x^{n+1}. Show that

$$\|x^{n+1} - p_n\|_2^2 = \int_a^b w(x)\phi_{n+1}\,dx/[c_{n+1}],$$

where c_{n+1} is the coefficient of x^{n+1} in $\phi_{n+1}(x)$.

Write down the best polynomial approximation of degree 2 to the function x^3 in the 2-norm with $w(x) \equiv 1$ on the interval $(-1, 1)$, and evaluate the 2-norm of the error.

9.9 Suppose that the weight w is an even function on the interval $(-a, a)$, and that a system of orthogonal polynomials ϕ_j, $j = 0, \ldots, n$, on the interval $(-a, a)$ is constructed by the Gram–Schmidt process. Show that, if j is even, then ϕ_j is an even function, and that, if j is odd, then ϕ_j is an odd function.

Now suppose that the best polynomial approximation of degree n in the 2-norm to the function f on the interval $(-a, a)$ is expressed in the form

$$p_n(x) = \gamma_0\phi_0(x) + \cdots + \gamma_n\phi_n(x).$$

Show that if f is an even function, then all the odd coefficients γ_{j-1} are zero, and that if f is an odd function, then all the even coefficients γ_j are zero.

9.10 The function $H(x)$ is defined by $H(x) = 1$ if $x > 0$, and $H(-x) = -H(x)$. Construct the best polynomial approximations of degrees 0, 1 and 2 in the 2-norm to this function over the interval $(-1, 1)$ with weight function $w(x) \equiv 1$. (It may not

appear very useful to consider a polynomial approximation to a discontinuous function, but representations of such functions by Fourier series will be familiar to most readers. Note that the function H belongs to $\mathrm{L}_{\cdot}(\ \ 1,1)$.)

10

Numerical integration – II

10.1 Introduction

In Section 7.2 we described the Newton–Cotes family of formulae for numerical integration. These were constructed by replacing the integrand by its Lagrange interpolation polynomial with equally spaced interpolation points and integrating this exactly. Here, we consider another family of numerical integration rules, called Gauss quadrature formulae, which are based on replacing the integrand f by its Hermite interpolation polynomial and choosing the interpolation points x in such a way that, after integrating the Hermite polynomial, the derivative values $f'(x)$ do not enter the quadrature formula. It turns out that this can be achieved by requiring that the x are roots of a polynomial of a certain degree from a system of orthogonal polynomials.

10.2 Construction of Gauss quadrature rules

Suppose that the function f is defined on the closed interval $[a, b]$ and that it is continuous and differentiable on this interval. Suppose, further, that w is a weight function, defined, positive, continuous and integrable on (a, b). We wish to construct quadrature formulae for the approximate evaluation of the integral

$$\int_a^b w(x) f(x) \, dx .$$

For a nonnegative integer n, let x_i, $i = 0, \ldots, n$, be $n + 1$ points in the interval $[a, b]$; the precise location of these points will be determined later on. The Hermite interpolation polynomial of degree $2n + 1$ for the

function f is given by the expression (see Section 6.4)

$$p(x) = H(x)f(x) + K(x)f'(x), \tag{10.1}$$

where

$$\begin{aligned} H(x) &= [L(x)]^2(1 - 2L'(x)(x - x)), \\ K(x) &= [L(x)]^2(x - x). \end{aligned} \tag{10.2}$$

Further, for $n \geq 1$, $L(x)$ is defined by

$$L(x) = \prod_{\substack{k=0 \\ k \neq}}^{n} \frac{x - x}{x - x}, \qquad k = 0, 1, \ldots, n;$$

if $n = 0$, we let $L(x) \equiv 1$ and thereby $H(x) \equiv 1$ and $K(x) = x - x$ for this value of n. Thus, we deduce from (10.1) that

$$\int w(x)f(x)dx \approx \int w(x)p(x)dx = W f(x) + V f'(x), \tag{10.3}$$

where

$$W = \int w(x)H(x)dx, \qquad V = \int w(x)K(x)dx.$$

There is an obvious advantage in choosing the points x in such a way that all the coefficients V are zero, for then the derivative values $f'(x)$ are not required. Recalling the form of the polynomial K and inserting it into the defining expression for V, we have

$$\begin{aligned} V &= \int w(x)[L(x)]^2(x - x)dx \\ &= C \int w(x)\pi(x)L(x)dx, \end{aligned} \tag{10.4}$$

where $\pi(x) = (x - x)\ldots(x - x)$ and

$$C = \begin{cases} \prod (x - x)^{-1} & \text{if } n \geq 1, \\ 1 & \text{if } n = 0. \end{cases}$$

Since π is of degree $n + 1$ while $L(x)$ is of degree n for each k, $0 \leq k \leq n$, each V will be zero if the polynomial π is orthogonal to every polynomial of lower degree with respect to the weight function

w. We can therefore construct the required quadrature formula (10.3) with V = 0, k = 0, . . . , n, by choosing the points x , k = 0, . . . , n, to be the zeros of the polynomial of degree n + 1 in a system of orthogonal polynomials over the interval (a, b) with respect to the weight function w; we know from Theorem 9.4 that these zeros are real and distinct, and all lie in the open interval (a, b).

Having chosen the location of the points x , we now consider W :

$$\begin{aligned} W &= \int w(x)H(x)dx \\ &= \int w(x)[L(x)]\,(1 - 2L'(x)(x - x))dx \\ &= \int w(x)[L(x)]\,dx - 2L'(x)V . \end{aligned} \tag{10.5}$$

Since V = 0, the second term in the last line vanishes and thus we obtain the following numerical integration formula, known as the **Gauss quadrature** rule:

$$\int w(x)f(x)dx \approx (f) = W f(x), \tag{10.6}$$

where the **quadrature weights** are

$$W = \int w(x)[L(x)]\,dx, \tag{10.7}$$

and the **quadrature points** x , k = 0, . . . , n, are chosen as the zeros of the polynomial of degree n + 1 from a system of orthogonal polynomials over the interval (a, b) with respect to the weight function w. Since this quadrature rule was obtained by exact integration of the Hermite interpolation polynomial of degree 2n + 1 for f, it gives the exact result whenever f is a polynomial of degree 2n + 1 or less.

Example 10.1 *Consider the case* n = 1*, with the weight function* w(x) 1 *over the interval* (0, 1).

The quadrature points x , x are then the zeros of the polynomial ϕ constructed in Example 9.5 and given by (9.8),

$$\phi(x) = x - + -, \tag{10.8}$$

1 9 , F ' %
#.# 4

and therefore

$$x = - \quad -, \quad x = - + \quad -.$$

Clearly, x and x belong to the open interval $(0,1)$, in accordance with Theorem 9.4. The weights are obtained from (10.7):

$$\begin{aligned} W &= \int \frac{x \quad x}{x \quad x} \, dx \\ &= 3 \int (x \quad 2x\, x + x\,) \, dx \\ &= 3(- \quad x \; + x\,) \\ &= -, \end{aligned} \tag{10.9}$$

and $W = -$ in the same way. We thus have the Gauss quadrature rule

$$\int f(x)dx \quad -f(- \quad -) + -f(- + \quad -), \tag{10.10}$$

which is exact whenever f is a polynomial of degree $2 \quad 1 + 1 = 3$ or less.

10.3 Direct construction

The calculation of the weights and the quadrature points in a Gauss quadrature rule requires little work when the system of orthogonal polynomials is already known. If this is not known, at the very least it is necessary to construct the polynomial from the system whose roots are the quadrature points; in that case a straightforward approach, which avoids this construction, may be easier.

Suppose, for example, that we wish to find the values of A , A , x and x such that the quadrature rule

$$\int f(x)dx \quad A\, f(x\,) + A\, f(x\,) \tag{10.11}$$

is exact for all $f \quad \&$.

We have to determine four unknowns, A , A , x and x , so we need four equations; thus we take, in turn, $f(x) \quad 1$, $f(x) = x$, $f(x) = x$ and $f(x) = x$ and demand that the quadrature rule (10.11) is exact (that is, the integral of f is equal to the corresponding approximation obtained by inserting f into the right-hand side of (10.11)). Hence,

$$1 = A + A \;, \tag{10.12}$$

$$- \quad = \quad A\ x \ + A\ x \ , \tag{10.13}$$

$$- \quad = \quad A\ x \ + A\ x \ , \tag{10.14}$$

$$- \quad = \quad A\ x \ + A\ x \ . \tag{10.15}$$

It remains to solve this system. To do so, we consider the quadratic polynomial π defined by

$$\pi\ (x) = (x \quad x\ )(x \quad x\ )$$

whose roots are the unknown quadrature points x and x . In expanded form, $\pi\ (x)$ can be written as

$$\pi\ (x) = x \ + px + q\,.$$

First we shall determine p and q; then we shall find the roots x and x of π . We shall then insert the values of x and x into (10.13) and solve the linear system (10.12), (10.13) for A and A .

To find p and q, we multiply (10.12) by q, (10.13) by p and (10.14) by 1, and we add up the resulting equations to deduce that

$$\begin{aligned} - + -p + q \quad &= \quad A\ (x \ + px \ + q) + A\ (x \ + px \ + q) \\ &= \quad A\ \pi\ (x\ ) + A\ \pi\ (x\ ) = A \quad 0 + A \quad 0 = 0\,. \end{aligned}$$

Therefore,

$$- + -p + q = 0\,. \tag{10.16}$$

Similarly, we multiply (10.13) by q, (10.14) by p and (10.15) by 1, and we add up the resulting equations to obtain

$$\begin{aligned} - + -p + -q \quad &= \quad A\ x\ (x \ + px \ + q) + A\ x\ (x \ + px \ + q) \\ &= \quad A\ x\ \pi\ (x\ ) + A\ x\ \pi\ (x\ ) = A \quad 0 + A \quad 0 = 0\,. \end{aligned}$$

Thus,

$$- + -p + -q = 0\,. \tag{10.17}$$

From (10.16) and (10.17) we immediately find that $p = \quad 1$ and $q = -$. Having determined p and q, we see that

$$\pi\ (x) = x \quad x + -\,,$$

in agreement with (10.8). We then find the roots of this quadratic polynomial to give x and x as before. With these values of x and x we deduce from (10.12) and (10.13) that

$$\begin{aligned} A \ + A \quad &= \quad 1\,, \\ A\ (- + \quad —) \quad A\ (- \quad —) \quad &= \quad 0\,, \end{aligned}$$

and therefore $A = A = -$. Thus, we conclude that the required quadrature rule is (10.10), as before.

It is easy to see that equations (10.16) and (10.17) express the condition that the polynomial $x + px + q$ is orthogonal to the polynomials 1 and x respectively. This alternative approach has simply constructed a quadratic polynomial from a system of orthogonal polynomials by requiring that it is orthogonal to every polynomial of lower degree, instead of building up the whole system of orthogonal polynomials.

A straightforward calculation shows that, in general, the quadrature rule (10.10) is not exact for polynomials of degree higher than 3 (take $f(x) = x$, for example, to verify this).

Example 10.2 *We shall apply the quadrature rule (10.10) to compute an approximation to the integral* $I = \int e \, dx$.

Using (10.10) with $f(x) = \exp(x) = e$ yields

$$I \quad -\exp - \quad \overline{\quad} \quad + -\exp \quad - + \quad \overline{\quad} \quad = \quad \bar{e}\cosh \quad \overline{\quad} .$$

On rounding to six decimal digits, $I \quad 1.717896$. The exact value of the integral is $I = e \quad 1 = 1.718282$, rounding to six decimal digits.

10.4 Error estimation for Gauss quadrature

The next theorem provides a bound on the error that has been committed by approximating the integral on the left-hand side of (10.6) by the quadrature rule on the right.

Theorem 10.1 *Suppose that* w *is a weight function, defined, integrable, continuous and positive on* (a, b)*, and that* f *is defined and continuous on* $[a, b]$*; suppose further that* f *has a continuous derivative of order* $2n + 2$ *on* $[a, b]$*,* $n \quad 0$*. Then, there exists a number* η *in* (a, b) *such that*

$$\int w(x) f(x) dx \qquad W \; f(x \;) = K \quad f \qquad (\eta) , \tag{10.18}$$

and

$$K \;= \frac{1}{(2n+2)!} \int w(x)[\pi \qquad (x)] \; dx .$$

Consequently, the integration formula (10.6), (10.7) will give the exact result for every polynomial of degree $2n + 1$.

Proof Recalling the definition of the Hermite interpolation polynomial p for the function f and using Theorem 6.4, we have

$$\begin{aligned} \int_a^b w(x)f(x)\,dx - W_k f(x_k) &= \int_a^b w(x)(f(x) - p(x))\,dx \\ &= \int_a^b w(x)\frac{f^{(2n+2)}(\xi(x))}{(2n+2)!}[\pi(x)]^2\,dx. \end{aligned} \tag{10.19}$$

However, by the Integral Mean Value Theorem, Theorem A.6, the last term is equal to

$$\frac{f^{(2n+2)}(\eta)}{(2n+2)!}\int_a^b w(x)[\pi(x)]^2\,dx,$$

for some $\eta \in (a,b)$, and hence the desired error bound. □

Note that, by virtue of Theorem 10.1, the Gauss quadrature rule gives the exact value of the integral when f is a polynomial of degree $2n+1$ or less, which is the highest possible degree that one can hope for with the $2n+2$ free parameters consisting of the quadrature weights W_k, $k = 0, \ldots, n$, and the quadrature points x_k, $k = 0, \ldots, n$.

A different approach leads to a proof of convergence of the Gauss formulae $G_n(f)$, defined in (10.6), (10.7), as $n \to \infty$.

Theorem 10.2 *Suppose that the weight function w is defined, positive, continuous and integrable on the open interval (a,b). Suppose also that the function f is continuous on the closed interval $[a,b]$. Then,*

$$\lim_{n\to\infty} G_n(f) = \int_a^b w(x)f(x)\,dx.$$

Proof If we choose any positive real number ε then, since f is continuous on $[a,b]$, the Weierstrass Theorem (Theorem 8.1) shows that there is a polynomial p such that

$$|f(x) - p(x)| \leq \varepsilon \quad \text{for all } x \in [a,b]. \tag{10.20}$$

Let N be the degree of this polynomial, and write p as p_N.

Thus we deduce that

$$\begin{aligned} \int_a^b w(x)f(x)\,dx - G_n(f) &= \int_a^b w(x)[f(x) - p_N(x)]\,dx \\ &\quad + \int_a^b w(x)p_N(x)\,dx - G_n(p_N) \\ &\quad + G_n(p_N) - G_n(f). \end{aligned} \tag{10.21}$$

Consider the first term on the right of this equality; it follows from (10.20) that

$$\int w(x)[f(x) - p(x)]dx < \varepsilon W,$$

where

$$W = \int w(x)dx.$$

For the last term on the right of (10.21),

$$\begin{aligned} & \sum (f) - \sum (p) \Big| W [f(x) - p(x)] \\ & < \varepsilon W \\ & = \varepsilon \int w(x)dx \\ & = \varepsilon W, \end{aligned} \tag{10.22}$$

where we have used the fact that all the quadrature weights W are positive (see (10.7)), and that a Gauss quadrature rule integrates a constant function exactly. Now for the middle term in (10.21), if we define N to be the integer part of $-N$, we see that when $n > N$ the quadrature formula is exact for all polynomials of degree $2N + 1$ or less, and hence for the polynomial p (given that $N < 2N + 1 < 2n + 1$). Therefore,

$$\int w(x)p(x)dx - \sum (p) = 0 \quad \text{if } n > N.$$

Putting these three terms together, we see that

$$\int w(x)f(x)dx - \sum (f) < \varepsilon W + 0 + \varepsilon W \quad \text{if } n > N.$$

Finally, given any positive number ε, we define $\varepsilon = \varepsilon/(2W)$ and find the corresponding value of $N = N(\varepsilon)$ to deduce that

$$\int w(x)f(x)dx - \sum (f) < \varepsilon \quad \text{if } n > N,$$

which is what we were required to prove. □

The interest of this theorem is mainly theoretical, as it gives no indication of how rapidly the error tends to zero. However, it does show the importance of the fact that the weights w are positive. Much of the above proof would apply with little change to the Newton–Cotes formulae of Section 7.2. We saw there that for the formulae of order 1 and 2, the trapezium rule and Simpson's rule, the weights are positive. However, when $n > 8$ some of the weights in the Newton–Cotes formula of order n become negative. In this case we have $\sum w = (b - a)$, but we find that $\sum |w| \to \infty$ as $n \to \infty$, so the proof breaks down. Stronger conditions must be imposed on the function f to ensure that the Newton–Cotes formula converges to the required integral. (See the example in Section 7.4.)

10.5 Composite Gauss formulae

It is often useful to define composite Gauss formulae, just as we did for the trapezium rule and Simpson's rule in Section 7.5. Let us suppose, for the sake of simplicity, that $w(x) \equiv 1$. We divide the range $[a, b]$ into m subintervals $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, m$, $m \geq 2$, each of width $h = (b - a)/m$, and write

$$\int_a^b f(x)\,dx = \sum_{j=1}^{m} \int_{x_{j-1}}^{x_j} f(x)\,dx,$$

where

$$x_j = a + jh, \qquad j = 0, 1, \ldots, m.$$

We then map each of the subintervals $[x_{j-1}, x_j]$, $j = 1, 2, \ldots, m$, onto the reference interval $[-1, 1]$ by the change of variable

$$x = \tfrac{1}{2}(x_{j-1} + x_j) + \tfrac{1}{2}ht, \qquad t \in [-1, 1],$$

giving

$$\int_{x_{j-1}}^{x_j} f(x)\,dx = \tfrac{1}{2}h \int_{-1}^{1} g_j(t)\,dt = \tfrac{1}{2}h\, I_j,$$

where

$$g_j(t) = f\left(\tfrac{1}{2}(x_{j-1} + x_j) + \tfrac{1}{2}ht\right) \quad \text{and} \quad I_j = \int_{-1}^{1} g_j(t)\,dt.$$

The composite Gauss quadrature rule is then obtained by applying

the same Gauss formula to each of the integrals I_j. This gives

$$\int_a^b f(x)\,dx \approx \frac{h}{2}\sum_{k=0}^{n} W_k\, g(\xi_k) = \frac{h}{2}\sum_{k=0}^{n} W_k\, f\left(\tfrac{1}{2}(x_{j-1}+x_j) + \tfrac{1}{2}h\xi_k\right), \tag{10.23}$$

where ξ_k are the quadrature points in $(-1,1)$ and W_k are the associated weights for $k = 0,\ldots,n$ with $n \geq 0$.

An expression for the error of this composite formula is obtained, as in Section 7.5, by adding the expressions (10.18) for the errors in the integrals I_j. The result is

$$E = C\,\frac{(b-a)}{2}\,\frac{h^{2n+2}}{(2n+2)!}\,f^{(2n+2)}(\eta) \tag{10.24}$$

where $\eta \in (a,b)$ and

$$C = \int_{-1}^{1} [\pi_{n+1}(t)]^2\,dt\,.$$

Definition 10.1 *The* **composite midpoint rule** *is the composite Gauss formula with* $w(x) \equiv 1$ *and* $n = 0$ *defined by*

$$\int_a^b f(x)\,dx \approx h\sum_{j=1}^{m} f(a+(j-\tfrac{1}{2})h)\,. \tag{10.25}$$

This follows from the fact that when $n = 0$ there is one quadrature point $\xi_0 = 0$ in $(-1,1)$, which is at the midpoint of the interval, and the corresponding quadrature weight W_0 is equal to the length of the interval $(-1,1)$, *i.e.*, $W_0 = 2$. It follows from (10.24) with $n = 0$ and

$$C = \int_{-1}^{1} t^2\,dt = \tfrac{2}{3}$$

that the error in the composite midpoint rule is

$$E = \frac{(b-a)}{24}\,h^2 f''(\eta)\,,$$

where $\eta \in (a,b)$, provided that the function f has a continuous second derivative on $[a,b]$.

10.6 Radau and Lobatto quadrature

We have now discussed two types of quadrature formulae, which have the same form, $\sum w_i f(x_i)$. In the Newton–Cotes formulae the (equally spaced) quadrature points x_i are given, and we were able to find the weights w_i so that the result was exact for polynomials of degree n. By allowing the quadrature points as well as the weights to be freely chosen, we constructed Gauss quadrature formulae which were exact for polynomials of degree $2n+1$. There are also many possible formulae of mixed type, where some, but not all, of the quadrature points are given, and the rest can be freely chosen. We might expect that each quadrature point which is fixed will reduce the degree of polynomial for which such a formula is exact by 1, from the maximum degree of $2n+1$.

It is often useful to be able to fix one of the endpoints of the interval as one of the quadrature points. As an example, suppose we prescribe that $x_1 = a$. Let p_{2n-1} be an arbitrary polynomial of degree $2n$, and write

$$p_{2n-1}(x) = (x-a)q_{2n-2}(x) + r,$$

where the quotient q_{2n-2} is a polynomial of degree $2n-2$ and the remainder r is a constant. The integral of $w p_{2n-1}$ is then

$$\int_a^b w(x)p_{2n-1}(x)\,dx = \int_a^b (x-a)w(x)q_{2n-2}(x)\,dx + r\int_a^b w(x)\,dx.$$

We can now construct the usual Gauss quadrature formula for the interval $[a,b]$ with the modified weight function $(x-a)w(x)$, giving n quadrature points and n weights x_i', W_i', $i = 1,\dots,n$. This formula will be exact for all polynomials q of degree $2n-3$. Provided that the weight function w satisfies the standard conditions on (a,b), the modified weight function does also; in particular it is clearly positive on (a,b). This gives

$$\begin{aligned}\int_a^b w(x)p_{2n-1}(x)\,dx &= \sum_{i=2}^{n} W_i' q_{2n-2}(x_i') + r\int_a^b w(x)\,dx \\ &= \sum_{i=2}^{n} \frac{W_i'}{x_i'-a}p_{2n-1}(x_i') \\ &\quad + r\left[\int_a^b w(x)\,dx - \sum_{i=2}^{n}\frac{W_i'}{x_i'-a}\right]. \qquad (10.26)\end{aligned}$$

The fact that $r = p$ (a) then leads us to consider the quadrature rule

$$\int w(x)f(x)dx \quad W\ f(a) + \quad W\ f(x\ ), \tag{10.27}$$

where

$$\begin{aligned} W \quad &= \quad W^{\square}/(x^{\square} \quad a), \quad k = 1, \ldots, n, \\ W \quad &= \quad w(x)dx \quad W\ . \end{aligned} \tag{10.28}$$

By construction, this formula is exact for all polynomials of degree $2n$. It is obvious that $W\ > 0$ for $k = 1, \ldots, n$. We leave it as an exercise to show that $W\ > 0$ also (see Exercise 5).

With only trivial changes it is easy to see how to construct a similar formula where instead of fixing $x\ = a$ we fix $x\ = b$. These are known as **Radau quadrature formulae**. We leave it as an exercise to construct the formula corresponding to fixing both $x\ = a$ and $x\ = b$, which is known as a **Lobatto quadrature formula**; as might be expected, this is exact for all polynomials of degree $2n \quad 1$ (see Exercise 7).

The formal process could evidently be generalised to allow for fixing one of the quadrature points at an internal point c, where $a < c < b$. However, this leads to the difficulty that the modified weight function

$$w^{\square}: x \quad (x \quad c)w(x)$$

is not positive over the whole interval (a, b); hence we can no longer be sure that it is possible to construct a system of orthogonal polynomials, or, even if we can, that these polynomials will have all their zeros real and distinct and lying in $[a, b]$. In general, therefore, such quadrature formulae may not exist.

10.7 Note

For a detailed guide to the literature on Gauss quadrature rules and its connection to the theory of orthogonal polynomials, we refer to the books cited in the Notes at the end of Chapter 7.

Exercises

10.1 Determine the quadrature points and weights for the weight function $w: x \quad \ln x$ on the interval $(0, 1)$, for $n = 0$ and $n = 1$.

10.2 The weights in the Gauss quadrature formula are given by (10.7), which is

$$W_r = \int_{-1}^{1} w(x)[L_r(x)]^2\,dx.$$

Show that W_r can also be calculated from

$$W_r = \int_{-1}^{1} w(x)L_r(x)\,dx.$$

(This is a simpler way of calculating W_r than (10.7); the importance of (10.7) is that it shows that the weights are all positive.)

10.3 Suppose that f has a continuous second derivative on $[0,1]$. Show that there is a point ξ in $(0,1)$ such that

$$\int_0^1 xf(x)dx = \tfrac{1}{3}f(\tfrac{1}{3}) + \tfrac{1}{9}f''(\xi).$$

10.4 Let $n \geq 0$. Write down the quadrature points x_j, $j = 0,\ldots,n$, for the weight function $w(x) = (1-x)^{-\frac{1}{2}}$ on the interval $(-1,1)$.

By induction, or otherwise, show that for positive integer values of n,

$$\cos(2j+1)\vartheta = \frac{\sin(2n+2)\vartheta}{2\sin\vartheta},$$

unless ϑ is a multiple of π. What is the value of the sum when ϑ is a multiple of π?

Deduce that

$$T_k(x_j) = \int_{-1}^{1} (1-x)^{-\frac{1}{2}} T_k(x)\,dx, \qquad k = 1,\ldots,n,$$

and show that

$$T_k(x_j) = \frac{n+1}{\pi}\int_{-1}^{1} (1-x)^{-\frac{1}{2}} T_k(x)\,dx,$$

where T_k is the Chebyshev polynomial of degree n.

Deduce that the weights of the quadrature formula with weight function $w(x) = (1-x)^{-\frac{1}{2}}$ on the interval $(-1,1)$ are

$$W_j = \frac{\pi}{n+1}, \quad j = 0,\ldots,n.$$

10.5 In the notation for the construction of the Radau quadrature formula in Section 10.6, show that $W_0 > 0$.

10.6 The **Laguerre polynomials** L_j, $j = 0, 1, 2, \ldots$, are the orthogonal polynomials associated with the weight function $w: x \mapsto e^{-x}$ on the semi-infinite interval $(0, \infty)$, with L_j of exact degree j. (See Exercise 5.9.) Show that

$$\int_0^\infty e^{-x}[L_j(x) - L_j^{\square}(x)]p(x)dx = 0$$

when p is any polynomial of degree less than j.

In the Radau formula

$$\int_0^\infty e^{-x} p(x)dx = W_0 p(0) + \sum_{k=1}^{n} W_k p(x_k),$$

where one of the quadrature points is fixed at $x = 0$, show that the other quadrature points x_k, $k = 1, \ldots, n$, are the zeros of the polynomial $L_n - L_n^{\square}$. Deduce that

$$\int_0^\infty e^{-x} p(x)dx = \tfrac{1}{2}p(0) + \tfrac{1}{2}p(2).$$

10.7 Let $n \geq 2$. Show that a polynomial p_{2n-1} of degree $2n - 1$ can be written

$$p_{2n-1}(x) = (x - a)(b - x)q_{2n-3}(x) + r(x - a) + s(b - x),$$

where q_{2n-3} is a polynomial of degree $2n - 3$, and r and s are constants. Hence construct the Lobatto quadrature formula

$$\int_a^b w(x)f(x)dx \approx W_0 f(a) + \sum_{k=1}^{n-1} W_k f(x_k) + W_n f(b),$$

which is exact when f is any polynomial of degree $2n - 1$. Show that all the weights W_k, $k = 0, 1, \ldots, n$, are positive.

10.8 Construct the Lobatto quadrature formula

$$\int_{-1}^{1} f(x) \approx A_{-1} f(-1) + A_0 f(x_0) + A_1 f(1)$$

for the interval $(-1, 1)$ with weight function $w(x) \equiv 1$, and with $n = 2$; write down and solve four equations to determine x_0, A_{-1}, A_0 and A_1.

1 , : B@ % #.* $ $/ , 9 , C # % #..? $ $/ , 9 , 4D

10.9 Write T for the composite trapezium rule (7.15), S for the composite Simpson rule (7.17) and M for the composite midpoint rule (10.25), each with m subintervals. Show that

$$\mathsf{M} = 2\mathsf{I} - \mathsf{I}, \quad \mathsf{S} = \frac{4\mathsf{I} - \mathsf{I}}{3}, \quad \mathsf{S} = \frac{2\mathsf{M} + \mathsf{I}}{3}.$$

11

Piecewise polynomial approximation

11.1 Introduction

Up to now, the focus of our discussion has been the question of approximation of a given function f, defined on an interval $[a, b]$, by a polynomial on that interval either through Lagrange interpolation or Hermite interpolation, or by seeking the polynomial of best approximation (in the -norm or 2-norm). Each of these constructions was *global* in nature, in the sense that the approximation was defined by the same analytical expression on the whole interval $[a, b]$. An alternative and more flexible way of approximating a function f is to divide the interval $[a, b]$ into a number of subintervals and to look for a piecewise approximation by polynomials of low degree. Such piecewise-polynomial approximations are called **splines**, and the endpoints of the subintervals are known as the **knots**.

More specifically, a spline of degree n, n 1, is a function which is a polynomial of degree n or less in each subinterval and has a prescribed degree of smoothness. We shall expect the spline to be at least continuous, and usually also to have continuous derivatives of order up to k for some k, 0 $k < n$. Clearly, if we require the derivative of order n to be continuous everywhere the spline is just a single polynomial, since if two polynomials have the same value and the same derivatives of every order up to n at a knot, then they must be the same polynomial. An important class of splines have degree n, with continuous derivatives of order up to and including n 1, but as we shall see later, lower degrees of smoothness are sometimes considered.

To give a flavour of the theory of splines, we concentrate here on two simple cases: linear splines and cubic splines.

11.2 Linear interpolating splines

Definition 11.1 *Suppose that* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$. *Further, let* $K = \{x_0, \ldots, x_m\}$ *be a subset of* $[a, b]$, *with* $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$, $m \geq 2$. *The* **linear spline** $s_\#$, *interpolating* f *at the points* $x_0, \ldots, x_m$, *is defined by*

$$s_\#(x) = \frac{x - x_{i-1}}{x_i - x_{i-1}} f(x_i) + \frac{x - x_i}{x_{i-1} - x_i} f(x_{i-1}),$$
$$x \in [x_{i-1}, x_i], \quad i = 1, 2, \ldots, m. \tag{11.1}$$

The points x_i, $i = 0, 1, \ldots, m$, *are the* **knots** *of the spline, and* K *is referred to as the* **set of knots**.

As the function $s_\#$ interpolates the function f at the knots, i.e., $s_\#(x_i) = f(x_i)$, $i = 0, 1, \ldots, m$, and over each interval $[x_{i-1}, x_i]$, for $i = 0, 1, \ldots, m$, the function $s_\#$ is a linear polynomial (and therefore continuous), we conclude that $s_\#$ is a continuous piecewise linear function on the interval $[a, b]$.

Given a set of knots $K = \{x_0, \ldots, x_m\}$, we shall use the notation $h_i = x_i - x_{i-1}$, and let $h = \max h_i$. Also, for a positive integer k, we denote by $C^k[a, b]$ the set of all real-valued functions, defined and continuous on the closed interval $[a, b]$, such that all derivatives, up to and including order k, are defined and continuous on $[a, b]$.

In order to highlight the accuracy of interpolation by linear splines we state the following error bound in the ∞-norm over the interval $[a, b]$.

Theorem 11.1 *Suppose that* $f \in C^2[a, b]$ *and let* $s_\#$ *be the linear spline that interpolates* f *at the knots* $a = x_0 < x_1 < \cdots < x_{m-1} < x_m = b$; *then, the following error bound holds:*

$$\|f - s_\#\|_\infty \leq \frac{1}{8} h^2 \|f''\|_\infty,$$

where $h = \max h_i = \max (x_i - x_{i-1})$, *and* $\|\cdot\|_\infty$ *denotes the* ∞*-norm over* $[a, b]$, *defined in (8.1).*

Proof Consider a subinterval $[x_{i-1}, x_i]$, $1 \leq i \leq m$. According to Theorem 6.2, applied on the interval $[x_{i-1}, x_i]$,

$$f(x) - s_\#(x) = \frac{1}{2} f''(\xi)(x - x_{i-1})(x - x_i), \qquad x \in [x_{i-1}, x_i],$$

where $\xi = \xi(x)$ (x_{i-1}, x_i). Thus,

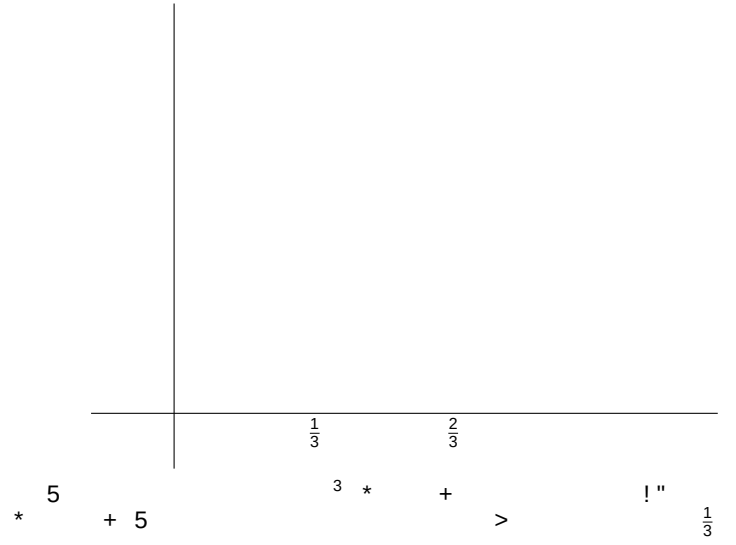
$$f(x) - s_{\#}(x) \leq \frac{1}{8}h^2 \max_{(\square)} f''(\zeta) .$$

Hence,

$$f(x) - s_{\#}(x) \leq \frac{1}{8}h^2 \|f''\|_\infty ,$$

for each $x \in [x_{i-1}, x_i]$ and each $i = 1, 2, \ldots, m$. This gives the required error bound. □

Figure 11.1 shows a typical example: a linear spline approximation to the function $f : x \mapsto e^{-x}$ over the interval $[0, 1]$, using two internal knots, $x_1 = \frac{1}{3}$, $x_2 = \frac{2}{3}$, together with the endpoints of the interval, $x_0 = 0$ and $x_3 = 1$.



We conclude this section with a result that provides a characterisation of linear splines from the viewpoint of the calculus of variations.

A subset A of the real line is said to have **measure zero** if it can be contained in a countable union of open intervals of arbitrarily small total length; in other words, for every $\varepsilon > 0$ there exists a sequence of open intervals (a_i, b_i), $i = 1, 2, 3, \ldots$, such that

$$A \subseteq \bigcup_{i=1}^{\infty} (a_i, b_i) \quad \text{and} \quad \sum_{i=1}^{\infty} (b_i - a_i) < \varepsilon .$$

In particular, any finite or countable set A has measure zero. For example, the set of all rational numbers is countable, and therefore it has measure zero. Trivially, the empty set has measure zero.

Suppose that B is a subset of . We shall say that a certain property = (x) holds for **almost every** x in B, if there exists a set A B of measure zero such that (x) holds for *all* x B A.

A real-valued function v defined on the interval [a, b] is said to be **absolutely continuous** on [a, b] if it has finite derivative v′(ξ) at almost every point ξ in [a, b], v′ is (Lebesgue-) integrable on [a, b], and

$$\int_a^x v'(\xi)\,d\xi = v(x) - v(a), \qquad a \le x \le b.$$

Example 11.1 *Any* v C [a, b] *is absolutely continuous on the interval* [a, b]. *The function* x x −(a + b) *is absolutely continuous on* [a, b], *but it does not belong to* C [a, b] *as it is not differentiable at* x = −(a+b).

Let us denote by H (a, b) the set of all absolutely continuous functions v defined on [a, b] such that v′ L (a, b), *i.e.*,

$$\|v'\| = \int_a^b |v'(\xi)|\,d\xi < .$$

We observe in passing that any function v H (a, b) is uniformly continuous on the closed interval [a, b]. This follows by noting that, for any pair of points x, y [a, b],

$$\begin{aligned} |v(x) - v(y)| &= \left|\int_y^x v'(\xi)\,d\xi\right| \\ &\le \int_y^x |v'(\xi)|\,d\xi \\ &\le |x - y|^{-} \|v'\| . \end{aligned}$$

In the transition from the first line to the second we used the Cauchy–Schwarz inequality.

If k 1, we shall denote by H (a, b) the set of all v H (a, b) such that v is absolutely continuous on [a, b] and v L (a, b). The set H (a, b) is called a **Sobolev space** of index k. We observe that

$$C[a,b] \quad H^1(a,b)$$

for any k 1, with strict inclusion. For example, any linear spline on

$[a, b]$ belongs to $H(a, b)$, but not to $C[a, b]$ unless it is a linear function over the *whole of* the interval $[a, b]$.

Example 11.2 *Let* $\alpha > 1/2$*; the function* $x \mapsto x^{\alpha}$ *then belongs to* $H(0, 1)$*, although it only belongs to* $C[0, 1]$ *if* $\alpha \geq 1$.

As a second example, consider the function $x \mapsto x \ln x$ *which belongs to* $H(0, 1)$*, but not to* $C[0, 1]$.

The variational characterisation of linear splines stated in the next theorem expresses the fact that, among all functions $v \in H(a, b)$ which interpolate a given continuous function f at a fixed set of knots in $[a, b]$, the linear spline $s_{\#}$ that interpolates f at these knots is the 'flattest', in the sense that its 'average slope' $s_{\#}'$ is smallest.

Theorem 11.2 *Suppose that* $s_{\#}$ *is the linear spline that interpolates* $f \in C[a, b]$ *at the knots* $a = x_0 < x_1 < \cdots < x_m = b$*. Then, for any function* v *in* $H(a, b)$ *that also interpolates* f *at these knots,*

$$\int_a^b (s_{\#}')^2 \leq \int_a^b (v')^2 .$$

Proof Let us observe that

$$\int_a^b (v')^2 = \int_a^b (v' - s_{\#}')^2 \, dx + \int_a^b (s_{\#}')^2 \, dx + 2 \int_a^b (v' - s_{\#}') s_{\#}' \, dx . \tag{11.2}$$

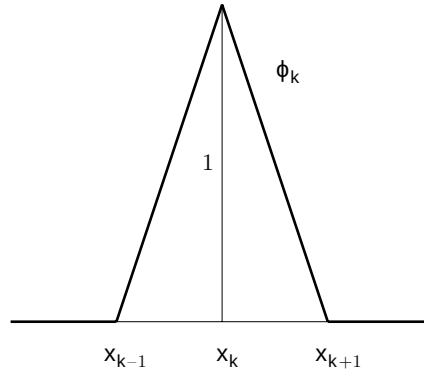
We shall now use integration by parts to show that the last integral is equal to 0; the desired inequality will then follow by noting that the first term on the right-hand side is nonnegative and it is equal to 0 if, and only if, $v = s_{\#}$. Clearly,

$$\int_a^b (v' - s_{\#}') s_{\#}' \, dx = \int_{x_0}^{x_m} (v' - s_{\#}') s_{\#}' \, dx$$

$$= \sum_{k=1}^{m} \left[(v(x_k) - s_{\#}(x_k)) s_{\#}'(x_k) - (v(x_{k-1}) - s_{\#}(x_{k-1})) s_{\#}'(x_{k-1}+) \right.$$

$$\left. - \int_{x_{k-1}}^{x_k} (v(x) - s_{\#}(x)) s_{\#}''(x) \, dx \right] . \tag{11.3}$$

Now $v(x_i) - s_{\#}(x_i) = f(x_i) - f(x_i) = 0$ for $i = 0, 1, \ldots, m$ and, since $s_{\#}$ is a linear polynomial over each of the open intervals (x_{k-1}, x_k), $k =$



5 , * + +

$1, 2, \ldots, m$, it follows that $s_{\#}''$ is identically 0 on each of these intervals. Thus, the expression in the square bracket in (11.3) is equal to 0 for each $k = 1, 2, \ldots, m$. □

Sobolev spaces play an important role in approximation theory. We shall encounter them again in Chapter 14 which is devoted to the approximation of solutions to differential equations by piecewise polynomial functions.

11.3 Basis functions for the linear spline

Suppose that $s_{\#}$ is a linear spline with knots x_i, $i = 0, 1, \ldots, m$, interpolating the function $f \in C[a, b]$. Instead of specifying the value of $s_{\#}$ on each subinterval $[x_{i-1}, x_i]$, $i = 1, 2, \ldots, m$, we can express $s_{\#}$ as a linear combination of suitable 'basis functions' ϕ_i as follows:

$$s_{\#}(x) = \sum_{i=0}^{m} \phi_i(x) f(x_i), \qquad x \in [a, b].$$

Here, we require that each ϕ_i is itself a linear spline which vanishes at every knot except x_i, and $\phi_i(x_i) = 1$. The function ϕ_i is often known as the **linear basis spline** or **hat function**, and is depicted in Figure 11.2.

The formal definition of ϕ is as follows:

$$\phi(x) = \begin{cases} 0 & \text{if } x \quad x_{-}, \\ (x \quad x_{-})/h & \text{if } x_{-} \quad x \quad x, \\ (x \quad x)/h & \text{if } x \quad x \quad x, \\ 0 & \text{if } x \quad x, \end{cases}$$

for $k = 1, \ldots, m \quad 1$, and with

$$\phi(x) = \begin{cases} (x \quad x)/h & \text{if } a = x \quad x \quad x, \\ 0 & \text{if } x \quad x \end{cases} \tag{11.4}$$

and

$$\phi(x) = \begin{cases} 0 & \text{if } x \quad x_{-}, \\ (x \quad x_{-})/h & \text{if } x_{-} \quad x \quad x \quad = b. \end{cases}$$

11.4 Cubic splines

Suppose that $f \quad C[a,b]$ and let $K = x, \ldots, x$ be a set of $m+1$ knots in the interval $[a,b]$, $a = x < x < \quad < x = b$. Consider the set $'$ of all functions $s \quad C[a,b]$ such that

$s(x) = f(x)$, $i = 0, 1, \ldots, m$,
s is a cubic polynomial on $[x_{-}, x]$, $i = 1, 2, \ldots, m$.

Any element of $'$ is referred to as an **interpolating cubic spline**. We note that, unlike linear splines which are uniquely determined by the interpolating conditions, there is more than one interpolating cubic spline $s \quad C[a,b]$ that satisfies the two conditions stated above; indeed, there are $4m$ coefficients of cubic polynomials (four on each subinterval $[x_{-}, x]$, $i = 1, 2, \ldots, m$), and only $m+1$ interpolating conditions and $3(m \quad 1)$ continuity conditions; since s belongs to $C[a,b]$, this means that s, s' and s'' are continuous at the internal knots $x, \ldots, x_{-}$. Hence, we have a total of $4m \quad 2$ conditions for the $4m$ unknown coefficients. Depending on the choice of the remaining two conditions we can construct various interpolating cubic splines.

An important class of cubic splines is singled out by the following definition.

Definition 11.2 *The* **natural cubic spline**, *denoted by* s, *is the element of the set* $'$ *satisfying the end conditions*

$$s''(x) = s''(x \quad) = 0.$$

We shall prove that this definition is correct in the sense that the two additional conditions in Definition 11.2 uniquely determine s : this will be done by describing an algorithm for constructing s .

Construction of the natural cubic spline. Let us begin by defining $\sigma = s''(x)$, $i = 0, 1, \ldots, m$, and noting that s'' is a linear function on each subinterval $[x_{-}, x]$. Therefore, s'' can be expressed as

$$s''(x) = \frac{x \quad x}{h}\sigma_{-} + x \quad x_{-}$$

Theorem 11.3 *Let* s *be the natural cubic spline that interpolates a function* $f \in C[a,b]$ *at the knots* $a = x_0 < x_1 < \cdots < x_m = b$. *Then, for any function* v *in* $H^2(a,b)$ *that also interpolates* f *at the knots,*

$$\|s''\|_2 \leq \|v''\|_2 .$$

The proof is analogous to that of Theorem 11.2 and is left as an exercise.

The *smoothest interpolation property* expressed by Theorem 11.3 is the source of the name *spline*. A spline is a flexible thin curve-drawing aid, made of wood, metal or acrylic. Assuming that its shape is given by the equation $y = v(x)$, $x \in [a,b]$, and is constrained by requiring that it passes through a finite set of prescribed points in the plane, v will take on a shape which minimises the strain energy

$$E(v) = \int \frac{v''(x)}{(1 + v'(x)^2)^{3/2}} dx$$

over all functions v which are constrained in the same way. If the function v is slowly varying, *i.e.*, $\max_{x \in [a,b]} |v'(x)| \ll 1$, this energy-minimisation property is very similar to the result in Theorem 11.3.

11.5 Hermite cubic splines

In the previous section we took $f \in C[a,b]$ and demanded that s belonged to $C^1[a,b]$; here we shall strengthen our requirements on the smoothness of the function that we wish to interpolate and assume that $f \in C^1[a,b]$; simultaneously, we shall relax the smoothness requirements on the associated spline approximation s by demanding that $s \in C[a,b]$ only.

Let $K = \{x_0, \ldots, x_m\}$ be a set of knots in the interval $[a,b]$ with $a = x_0 < x_1 < \cdots < x_m = b$ and $m \geq 2$. We define the **Hermite cubic spline** as a function $s \in C^1[a,b]$ such that

$s(x_i) = f(x_i)$, $s'(x_i) = f'(x_i)$ for $i = 0, 1, \ldots, m$,
s is a cubic polynomial on $[x_{i-1}, x_i]$ for $i = 1, 2, \ldots, m$.

Writing the spline s on the interval $[x_{i-1}, x_i]$ as

$$s(x) = c_{1,i} + c_{2,i}(x - x_{i-1}) + c_{3,i}(x - x_{i-1})^2 + c_{4,i}(x - x_{i-1})^3, \quad x \in [x_{i-1}, x_i], \quad (11.8)$$

we find that $c_{i-1} = f(x_{i-1})$, $c_i = f'(x_{i-1})$, and

$$\begin{aligned} c_{i-1} &= 3\frac{f(x_i) - f(x_{i-1})}{h} - \frac{f'(x_i) + 2f'(x_{i-1})}{h}, \\ c_i &= \frac{f'(x_i) + f'(x_{i-1})}{h} - 2\frac{f(x_i) - f(x_{i-1})}{h}. \end{aligned} \tag{11.9}$$

Note that the Hermite cubic spline only has a continuous first derivative at the knots, and therefore it is *not* an interpolating cubic spline in the sense of Section 11.4.

Unlike natural cubic splines, the coefficients of a Hermite cubic spline on each subinterval can be written down explicitly without the need to solve a tridiagonal system.

Concerning the size of the interpolation error, we have the following result.

Theorem 11.4 *Let* $f \in C[a,b]$*, and let* s *be the Hermite cubic spline that interpolates* f *at the knots* $a = x_0 < x_1 < \cdots < x_n = b$*; then, the following error bound holds:*

$$\|f - s\|_\infty \le \frac{1}{384} h^4 \|f'''\|_\infty,$$

where $f''' = f^{(3)}$ *is the fourth derivative of* f *with respect to its argument,* $h = \max h_i = \max (x_i - x_{i-1})$*, and* $\|\cdot\|_\infty$ *denotes the* ∞*-norm on the interval* $[a,b]$*.*

The proof is analogous to that of Theorem 11.1, except that Theorem 6.4 is used instead of Theorem 6.2.

Both the linear spline and the Hermite cubic spline are local approximations; the value of the spline at a point x between two knots x_{i-1} and x_i depends only on the values of the function and its derivative at these two knots. On the other hand, the natural cubic interpolating spline is a global approximation and, in this respect, it is more typical of a generic spline: a change in just one of the values at a knot, $f(x_i)$, will alter the right-hand side of the system of equations (11.7), so the values of all the quantities σ_i will change. Thus, the spline will change throughout the whole interval $[x_0, x_n]$. We conclude this section with an example.

Example 11.3 *Figure 11.3 shows the Hermite cubic spline approximation to the function* $f: x \to 1/(1+x)$*, using four equally spaced knots in the interval* $[0,5]$*.*

from the standard properties of binomial coefficients. Hence Q is a polynomial in x of degree $n - r - 1$, and the result follows by induction. Finally, this shows that Q is a polynomial of degree 0, and is therefore constant on . Thus, by the same argument, Q is identically 0 on . □

Theorem 11.5 *For each* $n \geq 1$*, the function* S *defined by*

$$S_n(x) = \sum_{k=0}^{n+1} (-1)^k \binom{n+1}{k} (x - kh)_+^n$$

is a spline of degree n *with equally spaced knots* kh*,* $k = 0, 1, \ldots, n+1$*. It has a continuous derivative of order* $n - 1$ *and is identically 0 outside the interval* $(0, (n+1)h)$*.*

Proof The function S_n is clearly a spline as stated, and $S_n(x)$ is identically 0 for $x \leq 0$. When $x \geq (n+1)h$ the arguments $x - kh$, $k = 0, 1, \ldots, n+1$, of the positive parts are all nonnegative, so that

$$S_n(x) = \sum_{k=0}^{n+1} (-1)^k \binom{n+1}{k} (x - kh)^n,$$

and this is identically zero by Lemma 11.1. □

Taking $n = 1$ we find that

$$S_1(x) = x_+ - 2(x - h)_+ + (x - 2h)_+ .$$

After normalisation by $1/h$ so as to have a maximum value of 1, and shifting $x = 0$ to $x = x_{-1}$, this yields a representation of the linear hat function ϕ_k from (11.4) in the form

$$\phi_k(x) = \frac{1}{h} S_1(x - x_{-1}),$$

which, for $1 \leq k \leq n$, is nonzero over two consecutive intervals: $(x_{-1}, x_1]$ and $[x_1, x_2)$.

In the same way we obtain a basis function for the cubic spline by taking $n = 3$:

$$S_3(x) = x_+^3 - 4(x - h)_+^3 + 6(x - 2h)_+^3 - 4(x - 3h)_+^3 + (x - 4h)_+^3 .$$

Normalising so as to have a maximum value of 1 and shifting $x = 0$ to $x = x_{-1}$, we get

$$\psi_k(x) = \frac{1}{4h} S_3(x - x_{-1}).$$

proximation in Theorem 11.4, but not for the natural cubic spline. The analysis of the error in the natural cubic spline approximation is quite complicated; Powell gives full details in his book.

The following are classical texts on the theory of splines.

) " / " , *The Theory of Splines and Their Applications*, Mathematics in Science and Engineering, 38, Academic Press, New York, 1967.

! , *A Practical Guide to Splines*, Revised Edition, Springer Applied Mathematical Sciences, 27, Springer, New York, 2001.

& - , *Spline Functions: Basic Theory,* John Wiley & Sons, New York, 1981.

The variational characterisations of splines stated in Sections 11.1 and 11.3 stem from the work of J.C. Holladay, Smoothest curve approximation, *Math. Comput.* **11**, 233–243, 1957.

Our definition of the Sobolev space H (a, b) in Section 11.1, based on the concept of absolute continuity, is specific to functions of a single variable. More generally, for functions of several real variables one needs to invoke the theory of weak differentiability or the theory of distributions to give a rigorous definition of the Sobolev space H (Ω) with Ω ; alternatively, one can define H (Ω) by completion of the set of smooth functions in a suitable norm. For the sake of simplicity of exposition we have chosen to avoid such general approaches.

Exercises

11.1 An interpolating spline of degree n is required to have continuous derivatives of order up to and including n 1 at the knots. How many additional conditions are required to specify the spline uniquely?

11.2 (i) Suppose that f is a polynomial of degree 1. Show that the linear spline s# which interpolates f at the knots x for i = 0, 1, . . . , m is identical to f, so that s# f.

(ii) Suppose that f is a polynomial of degree 3. Show that the Hermite cubic spline s$ which interpolates f at the knots x , i = 0, 1, . . . , m, is identical to f, so that s$ f.

(iii) Suppose that f is a polynomial of degree 3. Show that the natural cubic spline s which interpolates f at the knots x , i = 0, 1, . . . , m, is not in general identical to f.

11.3 Suppose that the natural cubic spline s interpolates the function $f: x \quad x$ on the interval $[0,1]$, the knots being equally spaced, so that $x = ih$, $i = 0,1,\ldots,m$, with $h = 1/m$, $m \quad 2$. Write down the equations which determine the quantities σ. If the two additional conditions are $\sigma = \sigma = 0$, show that these equations are not satisfied by $\sigma = f''(x)$, $i = 1,\ldots,m \quad 1$, so that s and f are not identical. If, however, these two additional conditions are replaced by $\sigma = f''(0)$, $\sigma = f''(1)$, show that $\sigma = f''(x)$, $i = 0,1,\ldots,m$, and deduce that s and f are identical.

11.4 A linear spline on the interval $[0,1]$ is expressed in terms of the basis functions as

$$s(x) = \quad \alpha \phi (x).$$

Instead of being required to interpolate the function f at the knots, the spline s is required to minimise $f \quad s$. Show that the coefficients α satisfy the system of equations

$$A = ,$$

where the elements of the matrix A are

$$A = \int \phi (x)\phi (x)dx$$

and the elements of are

$$b = \int f(x)\phi (x)dx.$$

Now suppose that the knots are equally spaced, so that $x = kh$, $k = 0,1,\ldots,m$, where $h = 1/m$, $m \quad 2$. Show that the matrix A is tridiagonal, with $A = -h$ for $i = 1,\ldots,m \quad 1$, and determine the other nonzero elements of A. Show also that A has the properties required for the use of the Thomas algorithm described in Section 3.3.

11.5 In the notation of Exercise 4, suppose that $f(x) = x$. Verify that the system of equations is satisfied by $\alpha = kh$, so that $s = f$.

Now suppose that $f(x) = x$. Verify that the equations are satisfied by $\alpha = (kh) + Ch$, where C is a constant to be determined. Deduce that $s(x) = f(x) + Ch$.

11.6 In the notation of Theorem 11.5, the spline basis function S of degree n is defined by

$$S \quad (x) = \quad (\quad 1) \quad \begin{matrix} n+1 \\ k \end{matrix} \quad (x \quad kh) \quad .$$

Explain why, for any value of a,

$$(x \quad a) \ \ (x \quad a) = (x \quad a) \quad .$$

Show that

$$xS \quad (x) + [(n+2)h \quad x]S \quad (x \quad h) = S \quad (x) \, .$$

Hence show by induction that $S \quad (x) \quad 0$ for all x.

11.7 Use the result of Exercise 6 to show by induction that each basis function $S \quad$ is symmetric; that is,

$$S \quad (p+x) = S \quad (p \quad x)$$

for all x, where $p = -(n+1)h$.

12

Initial value problems for ODEs

12.1 Introduction

Ordinary differential equations frequently occur in mathematical models that arise in many branches of science, engineering and economics. Unfortunately it is seldom that these equations have solutions which can be expressed in closed form, so it is common to seek approximate solutions by means of numerical methods. Nowadays this can usually be achieved very inexpensively to high accuracy and with a reliable bound on the error between the analytical solution and its numerical approximation. In this section we shall be concerned with the construction and the analysis of numerical methods for first-order differential equations of the form

$$y' = f(x, y) \tag{12.1}$$

for the real-valued function y of the real variable x, where y' $\frac{dy}{dx}$ and f is a given real-valued function of two real variables. In order to select a particular integral from the infinite family of solution curves that constitute the general solution to (12.1), the differential equation will be considered in tandem with an **initial condition**: given two real numbers x_0 and y_0, we seek a solution to (12.1) for $x > x_0$ such that

$$y(x_0) = y_0 \ . \tag{12.2}$$

The differential equation (12.1) together with the initial condition (12.2) is called an **initial value problem**.

If you believe that any initial value problem of the form (12.1), (12.2) possesses a unique solution, take a look at the following example.

Proof We define a sequence of functions $(y_n)_{n=0}^{\infty}$ by

$$y_0(x) = y_*,$$
$$y_n(x) = y_* + \int_{x_*}^{x} f(s, y_{n-1}(s))\,ds, \quad n = 1, 2, \ldots. \tag{12.4}$$

Since f is continuous on D, it is clear that each function y_n is continuous on $[x_*, X_0]$. Further, since

$$y_{n+1}(x) = y_* + \int_{x_*}^{x} f(s, y_n(s))\,ds,$$

it follows by subtraction that

$$y_{n+1}(x) - y_n(x) = \int_{x_*}^{x} [f(s, y_n(s)) - f(s, y_{n-1}(s))]\,ds. \tag{12.5}$$

We now proceed by induction, and assume that, for some positive value of n,

$$|y_n(x) - y_{n-1}(x)| \le \frac{K}{L}\frac{[L|x - x_*|]^n}{n!}, \quad x_* \le x \le X_0, \tag{12.6}$$

and that

$$|y_j(x) - y_{j-1}(x)| \le \frac{K}{L}\frac{[L|x - x_*|]^j}{j!},$$
$$x_* \le x \le X_0, \quad j = 1, \ldots, n. \tag{12.7}$$

Trivially, the hypotheses of the theorem and (12.4) imply that (12.6) and (12.7) hold for $n = 1$.

Now, (12.7) and (12.3) yield that

$$|y_j(x) - y_{j-1}(x)| \le \frac{K}{L}\, e^{L|x - x_*|} \le \frac{K}{L}\, e^{L(X_0 - x_*)} = C,$$
$$x_* \le x \le X_0, \quad j = 1, \ldots, n.$$

Therefore $(x, y_{n-1}(x)) \in D$ and $(x, y_n(x)) \in D$ for all $x \in [x_*, X_0]$. Hence, using (12.5), the Lipschitz condition and (12.6),

$$|y_{n+1}(x) - y_n(x)| \le L \int_{x_*}^{x} \frac{K}{L}\frac{[L|s - x_*|]^n}{n!}\,ds$$
$$= \frac{K}{L}\frac{[L|x - x_*|]^{n+1}}{(n+1)!}, \tag{12.8}$$

for all $x \quad [x\ , X_0\ ]$. Moreover, using (12.8) and (12.7),

$$y \quad (x) \quad y \qquad y \quad (x) \quad y\ (x) + y\ (x) \quad y$$
$$\frac{K}{L}\frac{[L(x \quad x\ )]}{(n+1)!} + \frac{K}{L} \qquad \frac{[L(x \quad x\ )]}{j!}$$
$$= \quad \frac{K}{L} \qquad \frac{[L(x \quad x\ )]}{(j+1)!}\,, \tag{12.9}$$

for all $x \quad [x\ , X_0\ ]$. Thus, (12.6) and (12.7) hold with n replaced by $n+1$, and hence, by induction, they hold for all positive integers n.

Since the infinite series $\quad^{\infty}\ (c\ /j!)$ converges (to $e^{+} \quad 1$) for any value of $c \qquad$, and for $c = L(X_0 \qquad x\ )$ in particular, it follows from (12.6) that the infinite series

$$\overset{\infty}{} \quad [y\ (x) \quad y\ _{-}\ (x)]$$

converges absolutely and uniformly for $x \quad [x\ , X_0\ ]$. However,

$$y\ + \quad [y\ (x) \quad y\ _{-}\ (x)] = y\ (x)\,,$$

showing that the sequence of continuous functions $(y\ )$ converges to a limit, uniformly on $[x\ , X_0\ ]$, and hence that the limit itself is a continuous function. Calling this limit y, we see from (12.4) that

$$\begin{aligned} y(x) &= \lim_{\to\infty} y \qquad (x) \\ &= y\ + \lim_{\to\infty} \qquad f\,(s, y\ (s))ds\,, \\ &= y\ + \qquad \lim_{\to\infty} f\,(s, y\ (s))ds\,, \\ &= y\ + \qquad f\,(s, y(s))ds\,, \end{aligned} \tag{12.10}$$

where we used the uniform convergence of the sequence of functions $(y\ )$ in the transition from line two to line three to interchange the order of the limit process and integration, and the continuity of the function f in the transition from line three to line four. As $s \qquad f\,(s, y(s))$ is a continuous function of s on the interval $[x\ , X_0\ ]$, its integral over the interval $[x\ , x]$ is a continuously differentiable function of x. Hence, by

(12.10), y is a continuously differentiable function of x on [x , X_0]; *i.e.*,
y

As a very simple example, consider the linear equation

$$y' = py + q, \tag{12.12}$$

where p and q are constants. Then, $L = |p|$, independently of C, and $K = |py + q|$. Hence, for any interval $[x, X_0]$, the conditions are satisfied by choosing C sufficiently large; therefore, the initial value problem has a unique continuously differentiable solution, defined for all $x \in [x, \infty)$.

Now, consider another example

$$y' = y, \quad y(0) = 1.$$

Here for any interval $[0, X_0]$ we have $K = 1$. Choosing any positive value of C we find that

$$|u - v| = |u + v| \, |u - v| \leq L|u - v| \, |u, v|,$$

where $L = 2(1 + C)$. We therefore now require the condition

$$C \geq \frac{1}{2(1+C)} e^{(\cdot)^2} - 1.$$

This is satisfied if

$$X_0 \leq F(C) = \frac{1}{2(1+C)} \ln(1 + 2C + 2C^2),$$

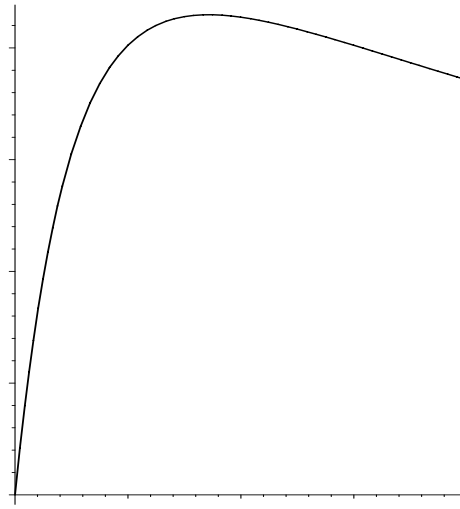
where ln means log. A sketch of the graph of the function F against C shows that F takes its maximum value near $C = 1.714$, and this gives the condition $X_0 \leq 0.43$ (see Figure 12.1).

Thus, we are *unable* to prove the existence of the solution over the infinite interval $[0, \infty)$. This is correct, of course, as the unique solution of the initial value problem is

$$y(x) = \frac{1}{1 - x}, \quad 0 \leq x < 1,$$

and this is not continuous, let alone continuously differentiable, on any interval $[0, X_0]$ with $X_0 \geq 1$. The conditions of Picard's Theorem, which are sufficient but not necessary for the existence and the uniqueness of the solution, have given a rather more restrictive bound on the size of the interval over which the solution exists.

The method of proof of Picard's Theorem also suggests a possible technique for constructing approximations to the solution, by determining the functions y_n from (12.4). In practice it may be impossible, or very difficult, to evaluate the necessary integrals in closed form. We



leave it as an exercise (see Exercise 3) to show that for the simple linear equation (12.12), with initial condition $y(0) = 1$, the function y is the same as the approximation obtained from the exact solution by expanding the exponential function as a power series and retaining the terms up to the one involving x.

In the rest of this chapter we shall consider step-by-step numerical methods for the approximate solution of the initial value problem (12.1), (12.2). We shall suppose throughout that the function f satisfies the conditions of Picard's Theorem. Suppose that the initial value problem (12.1), (12.2) is to be solved on the interval $[x, X_0]$. We divide this interval by the **mesh points** $x_n = x + nh$, $n = 0, 1, \ldots, N$, where $h = (X_0 - x)/N$ and N is a positive integer. The positive real number h is called the **step size** or **mesh size**. For each n we seek a numerical approximation y_n to $y(x_n)$, the value of the analytical solution at the mesh point x_n; these values y_n are calculated in succession, for $n = 1, 2, \ldots, N$.

12.2 One-step methods

A one-step method expresses y_{n+1} in terms of the previous value y_n; later on we shall consider k-step methods, where y_{n+k} is expressed in terms of the k previous values $y_{n+k-1}, \ldots, y_n$, where $k \geq 2$. The simplest example of a one-step method for the numerical solution of the initial value problem (12.1), (12.2) is Euler's method.

Euler's method. Given that $y(x_0) = y_0$, let us suppose that we have already calculated y_n, up to some n, $0 \leq n \leq N-1$, $N \geq 1$; we define

$$y_{n+1} = y_n + hf(x_n, y_n).$$

Thus, taking in succession $n = 0, 1, \ldots, N-1$, one step at a time, the approximate values y_n at the mesh points x_n can be easily obtained. This numerical method is known as **Euler's method**.

In order to motivate the definition of Euler's method, let us observe that on expanding $y(x_{n+1}) = y(x_n + h)$ into a Taylor series about x_n, retaining only the first two terms, and writing $y'(x_n) = f(x_n, y(x_n))$, we have that

$$y(x_n + h) = y(x_n) + hf(x_n, y(x_n)) + \mathcal{O}(h^2).$$

After replacing $y(x_n)$ and $y(x_n + h)$ by their numerical approximations, denoted by y_n and y_{n+1}, respectively, and discarding the $\mathcal{O}(h^2)$ term, we arrive at Euler's method.

More generally, a one-step method may be written in the form

$$y_{n+1} = y_n + h\Phi(x_n, y_n; h), \qquad n = 0, 1, \ldots, N-1, \quad y(x_0) = y_0, \tag{12.13}$$

where $\Phi(\cdot, \cdot; \cdot)$ is a continuous function of its variables. For example, in the case of Euler's method, $\Phi(x_n, y_n; h) = f(x_n, y_n)$. More intricate examples of one-step methods will be discussed below.

In order to assess the accuracy of the numerical method (12.13), we define the **global error**, e_n, by

$$e_n = y(x_n) - y_n.$$

We also need the concept of **truncation error**, T_n, defined by

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \Phi(x_n, y(x_n); h). \tag{12.14}$$

The next theorem provides a bound on the magnitude of the global error in terms of the truncation error.

Theorem 12.2

Assuming that $y \in C[x_n, x_{n+1}]$, *i.e.*, that y is a twice continuously differentiable function of x on $[x_n, x_{n+1}]$, and expanding $y(x_{n+1})$ about the point x_n into a Taylor series with remainder (see Theorem A.4), we have that

$$y(x_{n+1}) = y(x_n) + hy'(x_n) + \frac{h^2}{2!}y''(\xi_n), \qquad x_n < \xi_n < x_{n+1}.$$

Substituting this expansion into (12.18) gives

$$T_n = \frac{1}{2}hy''(\xi_n).$$

Let $M = \max_{x_0 \le x \le x_{n+1}} |y''(\zeta)|$. Then, $|T_n| \le T$, $n = 0, 1, \ldots, N-1$, where $T = \frac{1}{2}hM$. Inserting this into (12.16) and noting that for Euler's method $\Phi(x_n, y_n; h) = f(x_n, y_n)$ and therefore $L_\Phi = L$ where L is the Lipschitz constant for f, we have that

$$|e_{n+1}| \le \frac{1}{2}M\,\frac{e^{L(x_{n+1}-x_0)} - 1}{L}\,h^2, \quad n = 0, 1, \ldots, N. \tag{12.19}$$

Let us highlight the practical relevance of our error analysis by focusing on a particular example.

Example 12.2 *Let us consider the initial value problem* $y' = \tan^{-1} y$, $y(0) = y_0$, *where* y_0 *is a given real number. In order to find an upper bound on the global error* $e_n = y(x_n) - y_n$, *where* y_n *is the Euler approximation to* $y(x_n)$, *we need to determine the constants* L *and* M *in the inequality (12.19).*

Here $f(x, y) = \tan^{-1} y$; so, by the Mean Value Theorem (Theorem A.3),

$$f(x, u) - f(x, v) = \frac{\partial f}{\partial y}(x, \eta)\,(u - v) = \frac{\partial f}{\partial y}(x, \eta)\,|u - v|,$$

where η lies between u and v. In our case

$$\frac{\partial f}{\partial y}(x, y) = (1 + y^2)^{-1} \le 1,$$

and therefore $L = 1$. To find M we need to obtain a bound on y'' (without actually solving the initial value problem!). This is easily achieved by differentiating both sides of the differential equation with respect to the variable x:

$$y'' = \frac{d}{dx}(\tan^{-1} y) = (1 + y^2)^{-1}\,\frac{dy}{dx} = (1 + y^2)^{-1} \tan^{-1} y.$$

Therefore $y''(x)$ $M = -\pi$. Inserting the values of L and M into (12.19) and noting that $x = 0$, we have

$$e \quad -\pi\,(e \quad 1)\,h\,, \quad n = 0, 1, \ldots, N\,.$$

Thus, given a tolerance 012, specified beforehand, we can ensure that the error between the (unknown) analytical solution and its numerical approximation does not exceed this tolerance by choosing a positive step size h such that

$$h \quad \frac{4}{\pi(e^2 \quad 1)}\,012.$$

For such h we shall have $y(x\ ) \quad y = e \quad 012$, for $n = 0, 1, \ldots, N$, as required. Thus, at least in principle, we can calculate the numerical solution to arbitrarily high accuracy by choosing a sufficiently small step size h.

A numerical experiment shows that this error estimate is rather pessimistic. Taking, for example, $y = 1$ and $X_0 = 1$, our bound implies that the tolerance $012 = 0.01$ will be achieved with $h \quad 0.0074$; hence, it would appear that we need $N \quad 135$. In fact, using $N = 27$ gives a result from Euler's method which is just within this tolerance, so the error estimate has predicted the use of a step size which is five times smaller than is actually required.

Example 12.3 *As a more typical practical example, consider the problem*

$$y' = y \ + g(x)\,, \quad y(0) = 2\,, \tag{12.20}$$

where

$$g(x) = \frac{x \quad 6x \ + 12x \quad 14x + 9}{(1 + x)}\,,$$

is so chosen that the solution is known, and is

$$y(x) = \frac{(1 \quad x)(2 \quad x)}{1 + x}\,.$$

The results of some numerical calculations on the interval $x \quad [0, 1.6]$ are shown in Figure 12.2. They use step sizes 0.2, 0.1 and 0.05, and show how halving the step size gives a reduction of the error also by a factor of roughly 2, in agreement with the error bound (12.19).

that the one-step method (12.13) is consistent if, and only if,

$$\Phi(x, y; 0) \quad f(x, y). \tag{12.21}$$

This condition is sometimes taken as the definition of consistency. We shall henceforth always assume that (12.21) holds.

Now, we are ready to state a convergence theorem for the general one-step method (12.13).

Theorem 12.3 *Suppose that the initial value problem (12.1), (12.2) satisfies the conditions of Picard's Theorem, and also that its approximation generated from (12.13) when* $h \quad h$ *lies in the region* D*. Assume further that the function* $\Phi(\ , \ ; \ )$ *is continuous on* $D \quad [0, h\ ]$*, and satisfies the consistency condition (12.21) and the Lipschitz condition*

$$\Phi(x, u; h) \quad \Phi(x, v; h) \quad L_{\%}\, u \quad v \qquad \text{on } D \quad [0, h\ ]. \tag{12.22}$$

Then, if successive approximation sequences $(y\ )$*, generated by using the mesh points* $x \ = x\ + nh$*,* $n = 1, 2, \ldots, N$*, are obtained from (12.13) with successively smaller values of* h*, each* h *less than* h *, we have convergence of the numerical solution to the solution of the initial value problem in the sense that*

$$\lim_{\to\ \infty} y \ = y(x) \qquad \text{as} \quad x \qquad x \quad [x\ , X_0\ ] \text{ when } h \quad 0 \text{ and } n \qquad .$$

Proof Suppose that $h = (X_0 \quad x\ )/N$, where N is a positive integer. We shall assume that N is sufficiently large so that $h \quad h$. Since $y(x\ ) = y$ and therefore $e \ = 0$, Theorem 12.2 implies that

$$y(x\ ) \quad y \qquad \frac{e^{1 \quad 2 \quad -} \qquad 1}{L_{\%}} \quad \max_{\leq \ \leq \ -} T \quad , \qquad n = 1, 2, \ldots, N. \tag{12.23}$$

From the consistency condition (12.21) we have

$$T \quad = \quad \frac{y(x \quad) \quad y(x\ )}{h} \quad f(x\ , y(x\ )) + (\Phi(x\ , y(x\ ); 0) \quad \Phi(x\ , y(x\ ); h)). \tag{12.24}$$

According to the Mean Value Theorem, Theorem A.3, the expression in the first bracket is equal to $y^{\prime}(\xi\ ) \quad y^{\prime}(x\ )$, where $\xi \qquad [x\ , x \quad]$. By Picard's Theorem, $y^{\prime}$ is continuous on the closed interval $[x\ , X_0\ ]$; therefore, it is uniformly continuous on this interval. Hence, for each $\varepsilon > 0$ there exists $h\ (\varepsilon)$ such that

$$y^{\prime}(\xi\ ) \quad y^{\prime}(x\ ) \qquad -\varepsilon \qquad \text{for } h < h\ (\varepsilon), \quad n = 0, 1, \ldots, N \quad 1.$$

Also, since $\Phi(\ ,\ ;\ )$ is a continuous function on the closed set $\mathsf{D} \quad [0, \mathsf{h}$

for any pair of points $(x_n, y(x_n))$, $(x_{n+1}, y(x_{n+1}))$ *on the solution curve.*

12.4 An implicit one-step method

A one-step method with second-order accuracy is the **trapezium rule method**

$$y_{n+1} = y_n + \tfrac{1}{2}h[f(x_n, y_n) + f(x_{n+1}, y_{n+1})]. \tag{12.26}$$

This method is easily motivated by writing

$$y(x_{n+1}) - y(x_n) = \int_{x_n}^{x_{n+1}} y'(x)\,dx,$$

and approximating the integral by the trapezium rule. Since the right-hand side involves the integral of the function $x \to y'(x) = f(x, y(x))$ we see at once from (7.6) that the truncation error

$$T_n = \frac{y(x_{n+1}) - y(x_n)}{h} - \tfrac{1}{2}[f(x_n, y(x_n)) + f(x_{n+1}, y(x_{n+1}))]$$

of the trapezium rule method satisfies the bound

$$|T_n| \le \tfrac{1}{12}h^2 M, \quad \text{where } M = \max_{x_n \le x \le x_{n+1}} |y'''(x)|. \tag{12.27}$$

The important difference between this method and Euler's method is that the value y_{n+1} appears on both sides of (12.26). To calculate y_{n+1} from the known y_n therefore requires the solution of an equation, which will usually be nonlinear. This additional complication means an increase in the amount of computation required, but not usually a very large increase. The equation (12.26) is easily solved for y_{n+1} by Newton's method, assuming that the derivative $\partial f/\partial y$ can be calculated quickly; as a starting point for the Newton iteration the obvious estimate

$$y_n + hf(x_n, y_n),$$

will usually be close, and a couple of iterations will then suffice.

Methods of this type, which require the solution of an equation to determine the new value y_{n+1}, are known as **implicit methods**.

Writing the trapezium rule method in the standard form (12.13) we see that

$$\begin{aligned} h\Phi(x_n, y_n; h) &= \tfrac{1}{2}[f(x_n, y_n) + f(x_{n+1}, y_{n+1})] \\ &= \tfrac{1}{2}[f(x_n, y_n) + f(x_{n+1}, y_n + h\Phi(x_n, y_n; h))]. \end{aligned} \tag{12.28}$$

Hence, the function Φ is also defined in an implicit form.

In order to employ Theorem 12.2 to estimate the error in the trapezium rule method we need a value for the Lipschitz constant $L_\%$. From (12.28) we find that

$$\Phi(x\ ,u;h)\quad \Phi(x\ ,v;h) = -\ f(x\ ,u)\quad f(x\ +h,u+h\Phi(x\ ,u;h)) \\ f(x\ ,v)\quad f(x\ +h,v+h\Phi(x\ ,v;h))\ .$$

Hence,

$$\begin{aligned}&\Phi(x\ ,u;h)\quad \Phi(x\ ,v;h)\\ &\quad -\ f(x\ ,u)\quad f(x\ ,v)\\ &\quad +-\ f(x\ +h,u+h\Phi(x\ ,u;h))\quad f(x\ +h,v+h\Phi(x\ ,v;h))\\ &\quad -L\ u\quad v\\ &\quad +-L\ u+h\Phi(x\ ,u;h)\quad v\quad h\Phi(x\ ,v;h)\\ &\quad -L\ u\quad v\ +-L\ u\quad v\ +-L\ h\,\Phi(x\ ,u;h)\quad \Phi(x\ ,v;h)\ .\end{aligned}$$

This shows that

$$1\quad -hL\quad \Phi(x\ ,u;h)\quad \Phi(x\ ,v;h)\quad L\ u\quad v\ ,$$

and, therefore,

$$L_\%\quad \frac{L}{1\quad -hL}\ ,\qquad \text{provided that}\ -hL\ <1\ .$$

Consequently, (12.16) and (12.27) imply that the global error in the trapezium rule method is $(h\ )$, as h tends to 0.

Figure 12.3 depicts the results of some numerical calculations on the interval $x\quad [0,1.6]$ for the same problem as in Figure 12.2. The step sizes are 0.4 and 0.2, larger than for Euler's method; nevertheless we see a much reduced error in comparison with Euler's method, and also how the reduction in the step size h by a factor of 2 gives a reduction in the error by a factor of about 4, as predicted by our error analysis.

12.5 Runge–Kutta methods

Euler's method is only first-order accurate; nevertheless, it is simple and cheap to implement because, to obtain $y\quad$ from $y\ $, we only require a single evaluation of the function f, at $(x\ ,y\ )$. Runge–Kutta methods aim to achieve higher accuracy by sacrificing the efficiency of Euler's method through re-evaluating $f(\quad ,\quad)$ at points intermediate between

$$y''(x) = f + f\,y' = f + f\,f,$$
$$y'''(x) = f_{xx} + f_{xy}\,f + (f_x + f_y\,f)f + f\,(f_x + f_y\,f),$$

and so on; in these expressions the subscripts x and y denote partial derivatives, and all functions appearing on the right-hand sides are to be evaluated at $(x, y(x))$. We also need to expand $\Phi(x, y(x); h)$ in powers of h, giving (with the same notational conventions as before)

$$\Phi(x, y(x); h) = af + b\,f + \alpha h f_x + \beta h f f_y + \tfrac{1}{2}(\alpha h)^2 f_{xx} + \alpha\beta h^2 f f_{xy} + \tfrac{1}{2}(\beta h)^2 f^2 f_{yy} + O(h^3).$$

Thus, we obtain the truncation error in the form

$$\begin{aligned} T &= \frac{y(x+h) - y(x)}{h} - \Phi(x, y(x); h) \\ &= f + \tfrac{1}{2}h(f_x + f f_y) \\ &\quad + \tfrac{1}{6}h^2[f_{xx} + 2f f_{xy} + f^2 f_{yy} + f_x(f_x + f_y f)] \\ &\quad - af - b[f + \alpha h f_x + \beta h f f_y + \tfrac{1}{2}(\alpha h)^2 f_{xx} \\ &\qquad + \alpha\beta h^2 f f_{xy} + \tfrac{1}{2}(\beta h)^2 f^2 f_{yy}] + O(h^3). \end{aligned}$$

As $1 - a - b = 0$, the term $(1 - a - b)f$ is equal to 0. The coefficient of the term in h is

$$\tfrac{1}{2}(f_x + f f_y) - b\alpha f_x - b\beta f f_y$$

which vanishes for all functions f provided that

$$b\alpha = b\beta = \tfrac{1}{2}.$$

The method is therefore second-order accurate if

$$\beta = \alpha, \quad a = 1 - \frac{1}{2\alpha}, \quad b = \frac{1}{2\alpha}, \qquad \alpha \neq 0,$$

showing that there is a one-parameter family of second-order methods of this form, parametrised by $\alpha \neq 0$. The truncation error of the method then becomes

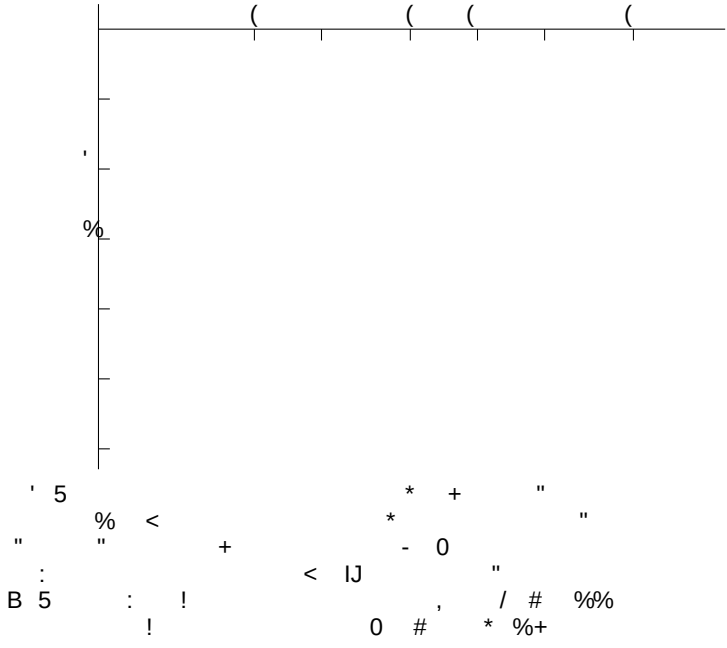
$$T = h^2\left(\tfrac{1}{6} - \tfrac{1}{4\alpha}\right)(f_x + f f_y) + \left(\tfrac{1}{6} - \tfrac{1}{4\alpha}\right) f f_y + \tfrac{1}{6}(f_y f_y + f f_y^2) + O(h^3). \tag{12.32}$$

Evidently there is no choice of the free parameter α which will make this method third-order accurate for all functions f; this can be seen, for example, by considering the initial value problem $y' = y$, $y(0) = 1$, and noting that in this case (12.32), with $f(x, y) = y$, yields

$$T = -h\,y(x) + O(h^3) = -h\,e^x + O(h^3).$$

Two examples of second-order Runge–Kutta methods of the form (12.29)–(12.31) are the modified Euler method and the improved Euler method.

(a)



To illustrate the behaviour of the one-step methods which we have discussed, Figure 12.4 shows the errors in the calculation of $y(1.6)$, where $y(x)$ is the solution to the problem (12.20) on the interval $[0, 1.6]$. The horizontal axis indicates N, the number of equally spaced mesh points used in the interval $(0, 1.6]$, on a logarithmic scale, and the vertical axis shows $\ln e = \ln |y(1.6) - y|$. The three methods employed are Euler's method, the trapezium rule method, and the classical Runge–Kutta method (12.33). The three lines show clearly the improved accuracy of the higher-order methods, and the rate at which the accuracy improves as N increases.

12.6 Linear multistep methods

While Runge–Kutta methods give an improvement over Euler's method in terms of accuracy, this is achieved by investing additional computational effort; in fact, Runge–Kutta methods require more evaluations of $f(x, y)$ than would seem necessary. For example, the fourth-order method involves four function evaluations per step. For comparison, by considering three consecutive points x_{n-1}, $x_n = x_{n-1} + h$, $x_{n+1} = x_{n-1} + 2h$, integrating the differential equation between x_{n-1} and x_{n+1},

yields

$$y(x_{n+1}) = y(x_{n-1}) + \int_{x_{n-1}}^{x_{n+1}} f(x, y(x))\,dx,$$

and applying Simpson's rule to approximate the integral on the right-hand side then leads to the method

$$y_{n+1} = y_{n-1} + \frac{1}{3}h\left[f(x_{n-1}, y_{n-1}) + 4f(x_n, y_n) + f(x_{n+1}, y_{n+1})\right], \tag{12.34}$$

requiring only three function evaluations per step. In contrast with the one-step methods considered in the previous section where only a single value y_n was required to compute the next approximation y_{n+1}, here we need *two* preceding values, y_n and y_{n-1}, to be able to calculate y_{n+1}, and therefore (12.34) is *not* a one-step method.

In this section we consider a class of methods of the type (12.34) for the numerical solution of the initial value problem (12.1), (12.2), called **linear multistep methods**.

Given a sequence of equally spaced mesh points (x_n) with step size h, we consider the general **linear k-step method**

$$\alpha_k y_{n+k} = h \sum_{j=0}^{k} \beta_j f(x_{n+j}, y_{n+j}), \tag{12.35}$$

where the coefficients $\alpha_0, \ldots, \alpha_k$ and $\beta_0, \ldots, \beta_k$ are real constants. In order to avoid degenerate cases, we shall assume that $\alpha_k \neq 0$ and that α_0 and β_0 are not both equal to 0. If $\beta_k = 0$, then y_{n+k} is obtained explicitly from previous values of y_j and $f(x_j, y_j)$, and the k-step method is then said to be **explicit**. On the other hand, if $\beta_k \neq 0$, then y_{n+k} appears not only on the left-hand side but also on the right, within $f(x_{n+k}, y_{n+k})$; due to this implicit dependence on y_{n+k} the method is then called **implicit**. The method (12.35) is called *linear* because it involves only linear combinations of the y_{n+j} and the $f(x_{n+j}, y_{n+j})$, $j = 0, 1, \ldots, k$; for the sake of notational simplicity, henceforth we shall often write f_{n+j} instead of $f(x_{n+j}, y_{n+j})$.

Example 12.4 *We have already seen an example of a linear two-step method in (12.34); here we present further examples of linear multistep methods.*

(a) Euler's method is a trivial case: it is an explicit linear one-step

method. The **implicit Euler method**

$$y_{n+1} = y_n + hf(x_{n+1}, y_{n+1}) \tag{12.36}$$

is an implicit linear one-step method. Another trivial example is the **trapezium rule method**, given by

$$y_{n+1} = y_n + \frac{1}{2}h(f_{n+1} + f_n);$$

it, too, is an implicit linear one-step method.

(b) The **Adams–Bashforth method**

$$y_{n+4} = y_{n+3} + \frac{1}{24}h(55f_{n+3} - 59f_{n+2} + 37f_{n+1} - 9f_n)$$

is an example of an explicit linear four-step method, while the **Adams–Moulton method**

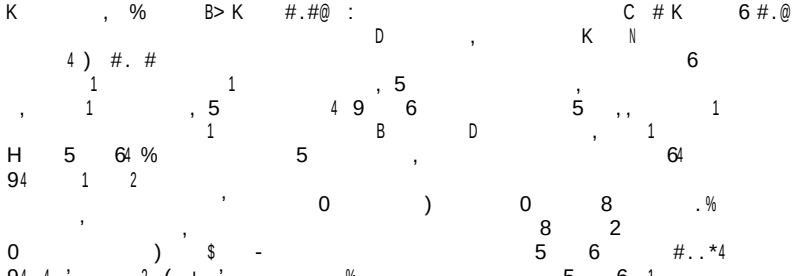
$$y_{n+3} = y_{n+2} + \frac{1}{24}h(9f_{n+3} + 19f_{n+2} - 5f_{n+1} + f_n)$$

is an implicit linear three-step method.

There are systematic ways of generating linear multistep methods, but these constructions will not be discussed here. Instead, we turn our attention to the analysis of linear multistep methods and introduce the concepts of *(zero-) stability*, *consistency* and *convergence*. The significance of these properties cannot be overemphasised: the failure of any of the three will render the linear multistep method practically useless.

12.7 Zero-stability

As is clear from (12.35) we need k starting values, $y_0, \ldots, y_{k-1}$, before we can apply a linear k-step method to the initial value problem (12.1), (12.2): of these, y_0 is given by the initial condition (12.2), but the others,



y ,...,y _ , have to be computed by other means: say, by using a

Before stating the main theorem of this section, we recall a classical result from the theory of kth-order linear recurrence relations.

Lemma 12.1 *Consider the* k*th-order homogeneous linear recurrence relation*

$$\alpha\ y \quad + \quad + \alpha\ y \quad + \alpha\ y \ = 0, \qquad n = 0, 1, 2, \dots, \qquad (12.38)$$

with $\alpha \ = 0$, $\alpha \ = 0$, $\alpha \quad , \ j = 0, 1, \dots, k$, *and the corresponding* **characteristic polynomial**

$$\rho(z) = \alpha\ z \ + \quad + \alpha\ z + \alpha\ .$$

Let $z\ ,\ 1 \quad r \quad ,\ \quad k$, *be the distinct roots of the polynomial* ρ, *and let* $m \quad 1$ *denote the multiplicity of* $z\ $, *with* $m\ + \quad + m\ = k$. *If a sequence* $(y\ )$ *of complex numbers satisfies (12.38), then*

$$y\ = \quad p\ (n) z\ , \qquad \textit{for all } n \quad 0, \qquad (12.39)$$

where $p\ (\ )$ *is a polynomial in* n *of degree* $m \quad 1,\ 1 \quad r \quad$. *In particular, if all roots are simple, that is* $m\ = 1,\ 1 \quad r \quad k$, *then the* $p\ ,\ r = 1, \dots, k$, *are constants.*

Proof We give a sketch of the proof. Let us first consider the case when all of the (distinct) roots $z\ , z\ , \dots, z$ are simple. As, by assumption, $\alpha\ = 0$, none of the roots is equal to 0. It is then easy to verify by direct substitution that, since $\rho(z\ ) = 0$, $r = 1, 2, \dots, k$, each of the sequences $(y\ ) = (z\ )$, $r = 1, 2, \dots, k$, satisfies (12.38).

In order to prove that any solution $(y\ )$ of (12.38) can be expressed as a linear combination of the sequences $(z\ ), (z\ ), \dots, (z\ )$, it suffices to show that these k sequences are linearly independent. To do so, let us suppose that

$$C\ z\ + C\ z\ + \quad + C\ z\ = 0, \qquad \text{for all } n = 0, 1, 2, \dots.$$

Then, in particular,

$$\begin{array}{ccccccc} C & + C & + & + C & = & 0, \\ C\ z & + C\ z & + & + C\ z & = & 0, \\ \dots & \dots & \dots & \dots & \dots & \\ C\ z^{-} & + C\ z^{-} & + & + C\ z^{-} & = & 0. \end{array}$$

[1] 9 1 8 4 #*C # 1 4 0 ,) 5 '
:) # . / 6 ! " #@? 4

The matrix of this system of k simultaneous linear equations for the k unknowns $C_1, C_2, \ldots, C_k$ has the determinant

$$\Delta = \begin{vmatrix} 1 & 1 & \ldots & 1 \\ z_1 & z_2 & \ldots & z_k \\ \ldots & \ldots & \ldots & \ldots \\ z_1^{k-1} & z_2^{k-1} & \ldots & z_k^{k-1} \end{vmatrix},$$

known as the Vandermonde determinant, and $\Delta = \prod_{1 \le i < j \le k} (z_j - z_i)$. Since the roots are distinct, $\Delta \neq 0$, so the matrix of the system is nonsingular. Therefore $C_1 = C_2 = \ldots = C_k = 0$ is the unique solution, which then means that the sequences $(z_1^n), (z_2^n), \ldots, (z_k^n)$ are linearly independent.

Now, suppose that (y_n) is any solution of (12.38); as $\Delta \neq 0$, there exists a unique set of k constants, $C_1, C_2, \ldots, C_k$, such that

$$y_m = C_1 z_1^m + C_2 z_2^m + \ldots + C_k z_k^m, \qquad m = 0, 1, \ldots, k-1. \tag{12.40}$$

Substituting these equalities into (12.38) for $n = 0$, we conclude that

$$\begin{aligned} 0 &= \alpha_0 y_0 + \alpha_1 (C_1 z_1^{-1} + \ldots + C_k z_k^{-1}) + \\ &\quad + \alpha_2 (C_1 z_1 + \ldots + C_k z_k) \\ &= \alpha_0 y_0 + C_1 (\rho(z_1) - \alpha_0 z_1) + \ldots + C_k (\rho(z_k) - \alpha_0 z_k) \\ &= \alpha_0 (y_0 - (C_1 z_1 + \ldots + C_k z_k)). \end{aligned}$$

As $\alpha_0 \neq 0$, it follows that

$$y_0 = C_1 z_1 + \ldots + C_k z_k,$$

which, together with (12.40), proves (12.39) for $0 \le n \le k$ in the case of simple roots. Next, we select $n = 1$ in (12.38) and proceed in the same manner as in the case of $n = 0$ discussed above to show that (12.39) holds for $0 \le n \le k+1$. Continuing in the same way, we deduce by induction that (12.39) holds for all $n \ge 0$.

In the case when $\rho(z)$ has repeated roots, the proof is similar, except that instead of (z_r^n), $r = 1, 2, \ldots, n$, the following k sequences are used:

$$\begin{aligned} &(z_r^n), \\ &(nz_r^n), \\ &\ldots\ldots\ldots \\ &(n(n-1)\ldots(n-m+2)z_r^n), \quad r = 1, 2, \ldots, \ell. \end{aligned} \tag{12.41}$$

These can be shown to satisfy (12.38) by direct substitution on noting that $\rho(z_r) = \rho'(z_r) = \ldots = \rho^{(m-1)}(z_r) = 0$, given that z_r is a root of

$\rho(z)$ of multiplicity m , $r = 1, 2, \ldots,$. The linear independence of the sequences (12.41) follows as before, except instead of $\prod_{3} (z \quad z\ )$, the value of the corresponding determinant is now

$$* = \prod_{\leq 3 \leq} (z \quad z\ ) \qquad \prod (m \quad 1)!!$$

where $0!! = 1$, $m!! = m!\,(m \quad 1)! \ldots 1!$ for $m = 1, 2, \ldots$. As the roots $z\ , z\ , \ldots, z$ are distinct, we have that $* \ = 0$, and therefore the sequences (12.41) are linearly independent. The rest of the argument is identical as in the case of simple roots. □

Now, we are ready to state the main result of this section.

Theorem 12.4 (Root Condition) *A linear multistep method is zero-stable for any initial value problem of the form (12.1), (12.2), where* f *satisfies the hypotheses of Picard's Theorem, if, and only if, all roots of the first characteristic polynomial of the method are inside the closed unit disc in the complex plane, with any which lie on the unit circle being simple.*

The algebraic stability condition contained in this theorem, namely that *the roots of the first characteristic polynomial lie in the closed unit disc and those on the unit circle are simple*, is often called the **Root Condition.**

Proof of theorem Consider the method (12.35), applied to $y' = 0$:

$$\alpha\ y \quad + \quad + \alpha\ y \quad + \alpha\ y\ = 0\,. \tag{12.42}$$

According to Lemma 12.1, every solution of this kth-order linear recurrence relation has the form

$$y\ = \quad p\ (n) z\ , \tag{12.43}$$

where z is a root, of multiplicity $m \quad 1$, of the first characteristic polynomial ρ of the method, and the polynomial p has degree $m \quad 1$, $1 \quad r \quad ,\ \quad k$. Clearly, if $z \ > 1$ for some r, then there are starting values $y\ , y\ , \ldots, y\ _-$ for which the corresponding solution grows like

1 , , 8 QQ
B D > * # 1 B D ? 1
5 4

z , and if $z = 1$ and the multiplicity is $m > 1$, then there is a solution growing like n^{-} . In either case there are solutions that grow unboundedly as n , *i.e.*, as $h \to 0$ with nh fixed. Considering starting values $y_0, y_1, \ldots, y_{-}$ which give rise to such an unbounded solution (y_n), and starting values $z_0 = z_1 = \cdots = z_{-} = 0$ for which the corresponding solution of (12.42) is (z_n) with $z_n = 0$ for all n, we see that (12.37) cannot hold. To summarise, if the Root Condition is violated, then the method is not zero-stable.

The proof that the Root Condition is sufficient for zero-stability is long and technical, and will be omitted here. For details, the interested reader is referred to Theorem 3.1 on page 353 of W. Gautschi, *Numerical Analysis: an Introduction*, Birkhäuser, Boston, MA, 1997. □

Example 12.5 *We shall explore the zero-stability of the methods from Example 12.4 using the Root Condition.*

(a) The Euler method and the implicit Euler method have first characteristic polynomial $\rho(z) = z - 1$ with simple root $z = 1$, so both methods are zero-stable. The same is true of the trapezium rule method.

(b) The Adams–Bashforth and Adams–Moulton methods considered in Example 12.4 have first characteristic polynomials, respectively, $\rho(z) = z(z - 1)$ and $\rho(z) = z(z - 1)$. These have multiple root $z = 0$ and simple root $z = 1$, and therefore both methods are zero-stable.

(c) The three-step method

$$11y_{n+} + 27y_{n+} - 27y_{n+} - 11y_n = 3h(f_{n+} + 9f_{n+} + 9f_{n+} + f_n) \qquad (12.44)$$

is *not* zero-stable. Indeed, the corresponding first characteristic polynomial $\rho(z) = 11z^{} + 27z^{} - 27z - 11$ has roots at $z = 1$, $z \approx -0.32$, $z = -3.14$, so $z > 1$.

(d) The first characteristic polynomial of the three-step method

$$y_{n+} + y_{n+} - y_{n+} - y_n = 2h(f_{n+} + f_n)$$

is $\rho(z) = z^{} + z^{} - z - 1 = (z + 1)(z - 1)$, which has roots $z = -1$, $z = 1$. The first of these is a double root lying on the unit circle; therefore, the method is *not* zero-stable.

12.8 Consistency

In this section we consider the accuracy of the linear k-step method (12.35). For this purpose, as in the case of one-step methods, we introduce the notion of truncation error. Thus, suppose that y is a solution to the ordinary differential equation (12.1). The truncation error of (12.35) is then defined as follows:

$$T = \frac{[\alpha\, y(x \quad) \quad h\beta\, f(x \quad, y(x \quad))]}{h \qquad \beta}. \tag{12.45}$$

Of course, the definition requires implicitly that $\sigma(1) = \qquad \beta = 0$. Again, as in the case of one-step methods, the truncation error can be thought of as the residual that is obtained by inserting the solution of the differential equation into the formula (12.35) and scaling this residual appropriately (in this case dividing through by $h \qquad \beta$), so that T resembles $y' \quad f(x, y(x))$.

Definition 12.4 *The numerical method (12.35) is said to be* **consistent** *with the differential equation (12.1) if the truncation error defined by (12.45) is such that for any* $\varepsilon > 0$ *there exists an* $h(\varepsilon)$ *for which*

$$T \quad < \varepsilon \quad \textit{for } 0 < h < h(\varepsilon),$$

and any $k + 1$ *points* $(x \ , y(x \ )), \ldots, (x \quad , y(x \quad))$ *on any solution curve in* D *of the initial value problem (12.1), (12.2).*

Now, let us suppose that the solution to the differential equation is sufficiently smooth, and let us expand the expressions $y(x \quad)$ and $f(x \quad , y(x \quad)) = y'(x \quad)$ into Taylor series about the point x . On substituting these expansions into the numerator in (12.45) we obtain

$$T = \frac{1}{h\sigma(1)} \; C\, y(x \ ) + C\, hy'(x \ ) + C\, h\, y''(x \ ) + \tag{12.46}$$

where

$$\begin{aligned} C \quad &= \quad \alpha \ , \\ C \quad &= \quad j\alpha \qquad \beta \ , \\ C \quad &= \quad -\alpha \qquad j\beta \ , \\ C \quad &= \quad -\alpha \qquad \frac{\qquad}{\_}\beta \ . \end{aligned} \tag{12.47}$$

For consistency we need that, as $h \to 0$ and $n \to \infty$ with $x_n \to x \in [x_0, X_0]$, the truncation error T tends to 0. This requires that $C_0 = 0$ and $C_1 = 0$ in (12.46). In terms of the characteristic polynomials this consistency requirement can be restated in compact form as

$$\rho(1) = 0 \quad \text{and} \quad \rho'(1) = \sigma(1) \,(= 0)\,.$$

Let us observe that, according to this condition, if a linear multistep method is consistent, then it has a *simple* root on the unit circle at $z = 1$; thus, the Root Condition is not violated by this root.

Definition 12.5 *The numerical method (12.35) is said to have* **order of accuracy** p*, if* p *is the largest positive integer such that, for any sufficiently smooth solution curve in* D *of the initial value problem (12.1), (12.2), there exist constants* K *and* h *such that*

$$T \le K h^{p+1} \quad \textit{for } 0 < h \le h\,,$$

for any $k + 1$ *points* $(x_0, y(x_0)), \ldots, (x_k, y(x_k))$ *on the solution curve.*

Thus, we deduce from (12.46) that the method is of order of accuracy p if, and only if,

$$C_0 = C_1 = \cdots = C_p = 0 \quad \text{and} \quad C_{p+1} \neq 0\,.$$

In this case,

$$T_n = \frac{C_{p+1}}{\sigma(1)} h^{p+1} y^{(p+1)}(x_n) + O(h^{p+2})\,.$$

The number $C_{p+1}/\sigma(1)$ is called the **error constant** of the method.

Example 12.6 *Let us determine all values of the real parameter* b*,* $b \neq 0$*, for which the linear multistep method*

$$y_{n+2} + (2b - 3)(y_{n+1} - y_n) - y_n = hb(f_{n+1} + f_n)$$

is zero-stable. We shall show that there exists a value of b *for which the order of the method is* 4*, and that if the method is zero-stable for some value of* b*, then its order cannot exceed* 2*.*

According to the Root Condition, this linear multistep method is zero-stable if, and only if, all roots of its first characteristic polynomial

$$\rho(z) = z^2 + (2b - 3)(z - 1) - 1$$

belong to the closed unit disc, and those on the unit circle are simple.

Clearly, $\rho(1) = 0$; upon dividing $\rho(z)$ by $z - 1$ we see that $\rho(z)$ can be written in the following factorised form:

$$\rho(z) = (z - 1)\rho_1(z), \qquad \text{where} \quad \rho_1(z) = z - 2(1 - b)z + 1.$$

Thus, the method is zero-stable if, and only if, all roots of the polynomial $\rho_1(z)$ belong to the closed unit disc, and those on the unit circle are simple and differ from 1. Suppose that the method is zero-stable. It then follows that $b = 0$ and $b = 2$, since these values of b correspond to double roots of $\rho_1(z)$ on the unit circle, respectively, $z = 1$ and $z = -1$. Further, since the product of the two roots of $\rho_1(z)$ is equal to 1, both have modulus less than or equal to 1, and neither of them is equal to -1, it follows that they must both be strictly complex; hence the discriminant of the quadratic polynomial $\rho_1(z)$ must be negative. That is, $4(1 - b)^2 - 4 < 0$. In other words, $b \in (0, 2)$.

Conversely, suppose that $b \in (0, 2)$. Then, the roots of $\rho(z)$ are

$$z = 1, \qquad z = 1 - b + i\sqrt{1 - (b - 1)^2}.$$

Since $|z| = 1$, $|z| = 1$ and $z \neq z$, all roots of $\rho(z)$ lie on the unit circle and they are simple. Hence the method is zero-stable. To summarise, the method is zero-stable if, and only if, $b \in (0, 2)$.

In order to analyse the order of accuracy of the method, we note that, upon Taylor series expansion, its truncation error can be written in the form

$$T = \frac{1}{\sigma(1)}\left[1 - \frac{b}{6}\right]h\,y'''(x) + \frac{1}{4}(6 - b)h^2 y''(x) + \frac{1}{120}(150 - 23b)h^3 y'''(x) + O(h^4),$$

where $\sigma(1) = 2b = 0$. If $b = 6$, then $T = O(h^4)$ and so the method is of order 4. As $b = 6$ does not belong to the interval $(0, 2)$, we deduce that the method is *not* zero-stable for $b = 6$.

Since zero-stability requires $b \in (0, 2)$, in which case $1 - b/6 \neq 0$, it follows that if the method is zero-stable, then $T = O(h^3)$.

12.9 Dahlquist's theorems

An important result connecting the concepts of zero-stability, consistency and convergence of a linear multistep method was proved by the Swedish mathematician Germund Dahlquist.

Theorem 12.5 (Dahlquist's Equivalence Theorem) *For a linear* k*-step method that is consistent with the ordinary differential equation (12.1) where* f *is assumed to satisfy a Lipschitz condition, and with consistent starting values, zero-stability is necessary and sufficient for convergence. Moreover if the solution* y *has continuous derivative of order* p + 1 *and truncation error* (h)*, then the global error of the method,* e = y(x) y *, is also* (h)*.*

The proof of this result is long and technical; for details of the argument, see Theorem 6.3.4 on page 357 of W. Gautschi, *Numerical Analysis: an Introduction*, Birkhäuser, Boston, MA, 1997, or Theorem 5.10 on page 244 of P. Henrici, *Discrete Variable Methods in Ordinary Differential Equations*, Wiley, New York, 1962.

By virtue of Dahlquist's theorem, if a linear multistep method is not zero-stable its global error cannot be made arbitrarily small by taking the mesh size h sufficiently small for any sufficiently accurate initial data. In fact, if the Root Condition is violated, then there exists a solution to the linear multistep method which will grow by an arbitrarily large factor in a fixed interval of x, however accurate the starting conditions are. This result highlights the importance of the concept of zero-stability and indicates its relevance in practical computations.

A second theorem by Dahlquist imposes a restriction on the order of accuracy of a zero-stable linear multistep method.

Theorem 12.6 (Dahlquist's Barrier Theorem) *The order of accuracy of a zero-stable* k*-step method cannot exceed* k + 1 *if* k *is odd, or* k + 2 *if* k *is even.*

A proof of this result will be found in Section 4.2 of Gautschi's book or in Section 5.2-8 of Henrici's book, cited above.

Theorem 12.6 makes it very difficult to choose a 'best' multistep method of a given order. Suppose, for example, that we consider five-step methods. The general five-step method involves 12 parameters, of

1 5 " A # # B$D % A & # , , 5
8 , 5 "0 $ 4

which 11 are independent: the method is obviously unaffected by multiplying all the parameters by a nonzero constant. Now it would be possible to construct a five-step method of order 10, by solving the 11 equations of the form $C_q = 0$, $q = 0, 1, \ldots, 10$, where C_q is given in (12.47). But the Barrier Theorem states that this method would not be zero-stable, and the order of a zero-stable five-step method cannot exceed 6. There is a family of stable five-step methods of order 6, involving 4 free parameters, and there is no obvious way of deciding whether any one of these methods is better than the others.

Example 12.7 (i) *The Barrier Theorem says that when* $k = 1$ *the order of accuracy of a zero-stable method cannot exceed* 2. *The trapezium rule method has order* 2, *and is zero-stable.*

(ii) *The two-step method*

$$y_{n+2} - y_{n+1} = h(-f_{n+2} + 3f_{n+1} - f_n)$$

is zero-stable, as the roots of the first characteristic polynomial, $\rho(z) = z^2 - z$, *are* 1 *and* -1. *A simple calculation shows that its order of accuracy is* 4; *by the Barrier Theorem, this is the highest order which could be achieved by a two-step method.*

(iii) *The three-step method*

$$\begin{aligned} 11y_{n+3} + 27y_{n+2} - 27y_{n+1} - 11y_n \\ = 3h(f_{n+3} + 9f_{n+2} + 9f_{n+1} + f_n) \end{aligned}$$

has order 6. *The Barrier Theorem therefore implies that this method is not zero-stable. We have already shown this in Example 12.5(c) using the Root Condition.*

It is found that all the zero-stable k-step methods of highest possible order are *implicit*, with β_k nonzero.

12.10 Systems of equations

In this section we discuss the application of numerical methods to simultaneous systems of differential equations, which we shall write in the form

$$\frac{dy}{dx} = f(x, y).$$

Here is an m-component vector function of x, and is an m-component vector function of the independent variable x and the vector variable . In component form the system becomes

$$\frac{dy}{dx} = f\ (x, y\ , \ldots, y\ ), \qquad j = 1, 2, \ldots, m\,.$$

The system comprises m simultaneous differential equations. To single out a unique solution we need m side conditions, and we shall suppose that all these conditions are given at the same value of x, and have the form

$$(x\ ) = \quad ,$$

or, in component form,

$$y\ (x\ ) = y\quad , \qquad j = 1, 2, \ldots, m\,,$$

where the values of y are given. This is called an initial value problem for a system of ordinary differential equations; we may also require a solution of the system on an interval $[a, b]$, with r conditions given at one end of the interval and m r conditions at the other end. This constitutes a boundary value problem, and requires different numerical methods which are considered in the next chapter.

All the numerical methods which we have discussed apply without change to systems of differential equations; it is only necessary to realise that we are dealing with vectors. For example, the first stage of the classical Runge–Kutta method (12.33) becomes

$$= \quad (x\ ,\quad);$$

we must evaluate all the elements of the vector before proceeding to the next stage to calculate , and so on.

The most important difference which arises in dealing with a system of differential equations is in the practical use of an *implicit* multistep method. As we have seen, this almost always requires an iterative method for the solution of an equation to determine y . Applying such a method to a system of differential equations now involves the solution of a system of equations, which will usually be nonlinear, to determine the elements of the vector . In real-life problems it is quite common to deal with systems of several hundred differential equations, and it then becomes very important to be sure that the improved efficiency of the implicit method justifies the very considerable extra work in each step of the process.

We shall not discuss the extension of our earlier analysis to deal with

systems of differential equations; in almost all cases we simply need to introduce vector notation, and replace the absolute value of a number by the norm of a vector. For example, in the proof of Theorem 12.2, (12.17) becomes

$$+ hL_{\%} \qquad + h \ \mathbf{T} \quad , \qquad n = 0, 1, \ldots, N \quad 1 ,$$

where is any norm on , with obvious definitions of the global error and the truncation error $\mathbf{T}$. Similarly, Picard's Theorem and its proof, discussed at the beginning of the chapter in the case of a single ordinary differential equation, can be easily extended to an m-component system of differential equations by replacing the absolute value sign with a vector norm on throughout.

12.11 Stiff systems

The phenomenon of stiffness usually appears only in a system of differential equations, but we begin by discussing an almost trivial example of a single equation,

$$y' = \lambda y , \qquad y(0) = y \ ,$$

where λ is a constant. The solution of this equation is evidently $y(x) = y \ \exp(\lambda x)$. When $\lambda < 0$ the absolute value of the solution is exponentially decreasing, so it is sensible to require that the absolute value of our numerical solution also decreases. It is very easy to give expressions for the result of a numerical solution using Euler's method and the implicit Euler method (12.36). They are, respectively,

$$y^{\&} = (1 + h\lambda) \ y \ , \qquad y' = (1 \quad h\lambda)^{-} \ y \ .$$

When $\lambda < 0$ and $h > 0$, we have $(1 \quad h\lambda) > 1$; therefore, the sequence (y') decreases monotonically with increasing n. On the other hand, for $\lambda < 0$ and $h > 0$,

$$1 + h\lambda \ < 1 \qquad \text{if, and only if,} \qquad 0 < h \ \lambda \ < 2 .$$

This gives the restriction $h \ \lambda \ < 2$ on the size of h for which the sequence $(y^{\&})$ decreases monotonically; if h exceeds $2/ \ \lambda$, the numerical solution obtained by Euler's method will oscillate with increasing magnitude with increasing n and fixed $h > 0$, instead of converging to zero as n .

We now consider the same two methods applied to the initial value problem for a system of differential equations of the form

$$' = A \ , \qquad (0) = \ ,$$

where A

We consider the system where A is the 2×2 matrix

$$A = \begin{matrix} 8003 & 1999 \\ 23988 & 6004 \end{matrix},$$

and the initial condition is

$$(0) = \begin{matrix} 1 \\ 4 \end{matrix}.$$

The eigenvalues of A are $\lambda = -7$ and $\lambda = -14000$; the solution of the problem is

$$(x) = \begin{matrix} e^{-} \\ 4e^{-} \end{matrix}.$$

Clearly, $\lim_{\to \infty} (x) = 0$.

The numerical solution uses 12 steps of size $h = 0.004$; the results are shown in Table 12.1. The second column gives the first component of the solution, $y(x) = e^{-}$, the third column shows the result from the implicit Euler method, and the last gives the result of the standard Euler method. The last column is a dramatic example of what happens when the step size h is too large; in this case $h\lambda = 56$. The numerical values given by the implicit Euler method have an error of a few units in the third decimal digit; to get the same accuracy from the Euler method would require a step size about 30 times smaller, and about 30 times as much work.

It is clear that the difficulty in the numerical example is caused by the size of the eigenvalue -14000, but what is important is its size relative to the other eigenvalue. The special constant-coefficient system $y' = Ay$ is said to be **stiff** if all the eigenvalues of A have negative real parts, and if the ratio of the largest of the real parts to the smallest of the real parts is large. Most practical problems are nonlinear, and for such problems it is quite difficult to define precisely what is meant by stiffness. To begin with we may replace the system by a linearised approximation, the first terms of an expansion

$$'(x) = '(x\ ) + \frac{\partial}{\partial x}(x\ , y(x\ ))(x \quad x\ ) + J(x\ )(\ (x) \quad (x\ )) +$$

[illegible]

Table 12.1. *The use of Euler's method and the implicit Euler method to solve a stiff system.*

$_1$* + 7 - -
' $) $)& $) $'% $'% $'($ $ $ $ % $' $($& %$) % ' '(') '& '
&)$$)$' &%)))) $' ')(%)($ '& ")&( )& (( )%$ ') ( ))))% &()

where J is the Jacobian matrix of the function , whose (i,j)-entry is

$$(\mathsf{J}(\mathsf{x}\ ))\quad = \frac{\partial \mathsf{f}}{\partial \mathsf{y}}(\mathsf{x}\ ,\ \ (\mathsf{x}\ ))\,.$$

We can then think of the system as being stiff if the eigenvalues of the matrix $\mathsf{J}(\mathsf{x}\ )$ have negative real parts and if the ratio of the largest of the real parts to the smallest is large. Although this gives some indication of the sort of problems which may cause difficulty, the behaviour of nonlinear systems is much more complicated than this. It is not difficult to construct examples in which all the eigenvalues of the Jacobian matrix have negative real parts, yet the norm of the solution of the differential equation is exponentially increasing as $\mathsf{x}\quad +\quad .$

Even though any classification of nonlinear systems of differential equations into stiff and nonstiff, based only on monitoring the eigenvalues of $\mathsf{J}(\mathsf{x}\ )$, is somewhat simplistic, it does highlight some of the key difficulties. Stiff systems of differential equations arise in many application areas, a typical one being chemical engineering. For example, in parts of an oil refinery there may be a large number of substances undergoing chemical reactions with widely different reaction rates. These reaction rates correspond to the eigenvalues of the Jacobian matrix, and it is not unusual to find the ratio of the largest of the real parts to the smallest to be in excess of 10 . For such problems it is essential to find a numerical method which imposes no restriction on the step size;

Euler's method, which might require the restriction $10 \quad h < 2$, would evidently be quite useless.

Application of the linear multistep method

$$\alpha\, y \quad = h \quad \beta\, f(x \quad, y \quad)$$

to the equation $y' = \lambda y$ leads to the kth-order linear recurrence relation

$$(\alpha \quad \lambda h\beta\,)y \quad = 0\,. \tag{12.48}$$

The characteristic polynomial of the linear recurrence relation (12.48) is

$$\pi(z;\lambda h) = \quad (\alpha \quad \lambda h\beta\,)z \quad .$$

Alternatively, we can write this in terms of the first and second characteristic polynomials of the linear multistep method as

$$\pi(z;\lambda h) = \rho(z) \quad \lambda h\sigma(z)\,.$$

In the present context, the polynomial $\pi(\quad;\lambda h)$ is usually referred to as the **stability polynomial** of the linear multistep method. According to Lemma 12.1, the general solution of the recurrence relation (12.48) can be expressed in terms of the distinct roots z , $1 \quad r \quad$, $\quad k$, of $\pi(\quad;\lambda h)$. Letting m denote the multiplicity of the root z , $1 \quad r \quad$, $m + \quad + m = k$, we have that

$$y \;=\; \quad p\,(n)z \quad , \tag{12.49}$$

where the polynomial $p\,(\quad)$ has degree $m \quad 1$, $1 \quad r \quad$.

Clearly, the roots z are functions of λh. For $\lambda \quad$, with $\mathrm{Re}(\lambda) < 0$, the solution of the model problem

$$y' = \lambda y\,, \qquad y(0) = y \quad ,$$

converges in $\quad$ to 0 as $x \quad$. Thus, we would like to ensure that, when a linear multistep method is applied to this problem, the step size h can be chosen so that the resulting sequence of numerical approximations $(y \quad)$ exhibits an analogous behaviour as $n \quad$, that is, $\lim_{\to\infty} y \quad = 0$. By virtue of (12.49), this can be guaranteed by demanding that each root $z \quad = z\,(\lambda h)$ has modulus less then 1.

Definition 12.6 *A linear multistep method is said to be* **absolutely stable** *for a given value of* λh *if each root* $z = z(\lambda h)$ *of the associated stability polynomial* $\pi(\ ; \lambda h)$ *satisfies* $|z(\lambda h)| < 1$.

Our aim is, therefore, to single out those values of λh for which the linear multistep method is absolutely stable.

Definition 12.7 *The* **region of absolute stability** *of a linear multistep method is the set of all points* λh *in the complex plane for which the method is absolutely stable.*

Ideally, the region of absolute stability of the method should admit all values of λ, $\mathrm{Re}(\lambda) < 0$, so as to ensure that there is no limitation on the size of h, however large $|\lambda|$ may be. This leads us to the next definition.

Definition 12.8 *A linear multistep method is said to be* **A-stable** *if its region of absolute stability contains the negative (left) complex half-plane.*

Unfortunately, the condition of A-stability is extremely demanding. Dahlquist has shown the following results which are collectively known as his **Second Barrier Theorem**:

(i) No *explicit* linear multistep method is A-stable;
(ii) No A-stable linear multistep method can have order greater than 2.
(iii) The second-order A-stable linear multistep method with the smallest error constant is the trapezium rule method.

The trapezium rule method is a one-step method, so the associated stability polynomial has only one root, given by

$$z = \frac{1 + \frac{1}{2}\lambda h}{1 - \frac{1}{2}\lambda h}.$$

Evidently $|z| < 1$ if $\mathrm{Re}(h\lambda) = h\,\mathrm{Re}(\lambda) < 0$, so the trapezium rule method is indeed A-stable.

To construct useful methods of higher order we need to relax the condition of A-stability by requiring that the region of absolute stability should include a large part of the negative half-plane, and certainly that it contains the whole of the negative real axis.

[illegible]

The most efficient methods of this kind in current use are the **Backward Differentiation Formulae**, or BDF methods. These are the linear multistep methods (12.35) in which $\beta = 0$, $0 \quad j \quad k \quad 1$, $k \quad 1$, and $\beta = 0$. Thus,

$$\alpha \; y \quad + \quad + \alpha \; y \; = h\beta \; f \quad .$$

The coefficients are obtained by requiring that the order of accuracy of the method is as high as possible, *i.e.*, by making the coefficients C zero in (12.47) for $j = 0, 1, \ldots, k$. For $k = 1$ this yields the implicit Euler method (BDF1), whose order of accuracy is, of course, 1; the method is A-stable. The choice of $k = 6$ results in the sixth-order, six-step BDF method (BDF6):

$$147y \quad 360y \quad + 450y \quad 400y \quad + 225y \quad 72y \quad + 10y = 60hf \quad . \qquad (12.50)$$

Although the method (12.50) is not A-stable, its region of absolute stability includes the whole of the negative real axis (see Figure 12.5). For the intermediate values, $k = 2, 3, 4, 5$, we have the following kth-order, k-step BDF methods, respectively:

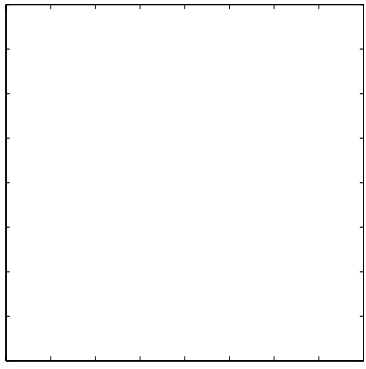
$$\begin{aligned}
3y \quad 4y \quad + y \quad &= 2hf \quad , \\
11y \quad 18y \quad + 9y \quad 2y \quad &= 6hf \quad , \\
25y \quad 48y \quad + 36y \quad 16y \quad + 3y \quad &= 12hf \quad , \\
137y \quad 300y \quad + 300y \quad 200y \quad + 75y \quad 12y \quad &= 60hf \quad ,
\end{aligned}$$

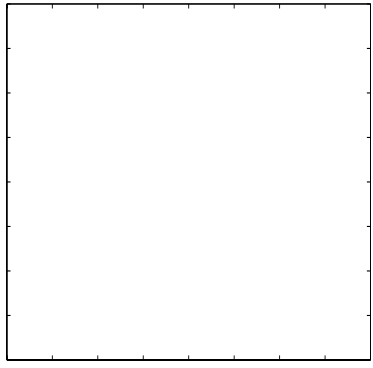
referred to as BDF2, BDF3, BDF4 and BDF5. Their regions of absolute stability are also shown in Figure 12.5. In each case the region of absolute stability includes the negative real axis. Higher-order methods of this type cannot be used, as all BDF methods, with $k > 6$, are zero-unstable.

12.12 Implicit Runge–Kutta methods

For Runge–Kutta methods absolute stability is defined in much the same way as for linear multistep methods; *i.e.*, by applying the method in question to the model problem $y' = \lambda y$, $y(0) = y$, λ , $\mathrm{Re}(\lambda) < 0$, and demanding that the resulting sequence $(y \;)$ converges to 0 as n , with $h\lambda$ held fixed. The set of all values of $h\lambda$ in the complex plane for which the method is absolutely stable is called the region of absolute stability of the Runge–Kutta method.

Classical Runge–Kutta methods are explicit, and are unsuitable for





where

$$k_i = f(x_n + hc_i, y_n + h\sum_{j=1}^{s} a_{ij} k_j), \qquad 1 \le i \le s. \tag{12.51}$$

It is convenient to display the coefficients in a **Butcher tableau**

$$\begin{array}{c|ccc} c_1 & a_{11} & \dots & a_{1s} \\ \vdots & \vdots & \vdots & \vdots \\ c_s & a_{s1} & \dots & a_{ss} \\ \hline & b_1 & \dots & b_s \end{array}$$

The method is then defined by the matrix $A = (a_{ij})_{s \times s}$, of order s, and the two vectors $b = (b_1, \dots, b_s)$ and $c = (c_1, \dots, c_s)$. For example, the classical four-stage Runge–Kutta method is defined by the tableau

$$\begin{array}{c|cccc} 0 & & & & \\ \frac{1}{2} & \frac{1}{2} & & & \\ \frac{1}{2} & 0 & \frac{1}{2} & & \\ 1 & 0 & 0 & 1 & \\ \hline & \frac{1}{6} & \frac{1}{3} & \frac{1}{3} & \frac{1}{6} \end{array}$$

The 4×4 array representing the matrix A for this method, displayed in the upper right quadrant of the tableau, follows the usual notational convention that zero elements after the last nonzero element in each row of the matrix A are omitted.

This is an explicit method, shown by the fact that the matrix A is *strictly lower triangular*, with $a_{ij} = 0$ when $1 \le i \le j \le 4$. Each value k_i can therefore be calculated in sequence, all the quantities on the right-hand side of (12.51) being known.

It is not difficult to construct s-stage implicit methods which are A-stable. For example, this can be done by choosing the coefficients c_i and b_i to be the quadrature points and weights respectively in the Gauss quadrature formula for the evaluation of

$$\int_{-1}^{1} g(x)\,dx \approx \sum_{i=1}^{s} b_i\, g(c_i).$$

The numbers a_{ij} can then be chosen so that the method has order $2s$, and is A-stable.

For example, the array

-(3 3)	-	—(3 2 3)
-(3 + 3)	—(3 + 2 3)	-
	-	-

defines a 2-stage A-stable method of order 4.

However, there is a heavy price to pay for using implicit methods of this kind, as we now have to calculate all the numbers k_i, $i = 1, 2, \ldots, s$, simultaneously, not in succession. For a system of m differential equations an implicit linear multistep method requires the solution of m simultaneous equations at each step; an s-stage implicit Runge–Kutta method requires the solution of sm simultaneous equations. This is a considerable increase in cost, and the general implicit Runge–Kutta methods cannot compete in efficiency with the Backward Differentiation Formulae such as (12.50); their use is almost exclusively limited to stiff systems of ODEs.

The overall computational effort can be somewhat reduced by using **diagonally implicit Runge–Kutta** (or DIRK) methods, in which the matrix A is lower triangular, so that $a_{ij} = 0$ if $j > i$. A further improvement in efficiency is possible by requiring in addition that all the diagonal elements a_{ii} are the same; unfortunately it has proved difficult to construct such methods with order greater than 4.

12.13 Notes

In this chapter we have only been able to introduce some of the basic ideas in what has become a vast area of numerical analysis. In particular we have not discussed the practical implementation of the various methods. The questions of how to choose the step size h to obtain efficiently a prescribed accuracy, and when and how to adjust h during the course of the calculation, are dealt with in the following books.

/ " & $ 3 " , *Solving Ordinary Differential Equations I: Nonstiff Problems,* Second Edition, Springer Series in Computational Mathematics, 8, Springer, Berlin, 1993.

) # , *A First Course in the Numerical Analysis of Differential Equations*, Cambridge University Press, Cambridge, 1996.

, *Numerical Methods for Ordinary Differential Systems,* John Wiley & Sons, Chichester, 1991.

For a study of dynamical systems and their numerical analysis, with focus on long-time behaviour, we refer to

) + &) % , *Dynamical Systems and Numerical Analysis*, Cambridge University Press, Cambridge, 1999.

The numerical solution of stiff initial value problems for systems of ordinary differential equations is discussed in

/ , *Solving Ordinary Differential Equations II: Stiff and Differential-Algebraic Problems*, Springer Series in Computational Mathematics, 14, Springer, Berlin, 1991.

An extensive survey of the theory of Runge–Kutta and linear multistep methods is found in

! , *The Numerical Analysis of Ordinary Differential Equations. Runge–Kutta and General Linear Methods*, Wiley-Interscience, John Wiley & Sons, Chichester, 1987.

Satisfactory theoretical treatment of nonlinear systems of differential equations from the point of view of stiffness requires the development of a genuinely nonlinear stability theory which does not involve the rather dubious idea of defining stiffness through linearisation based on the 'frozen Jacobian matrix'. We close by mentioning just one concept in this direction – that of *algebraic stability*. Given a Runge–Kutta method with Butcher tableau

$$\begin{array}{c|c} & \mathsf{A} \\ \hline & \end{array}$$

we define the matrices

$$\mathsf{B} = \mathrm{diag}(\mathsf{b}_1, \mathsf{b}_2, \ldots, \mathsf{b}_s) \quad \text{and} \quad \mathsf{M} = \mathsf{B}\mathsf{A} + \mathsf{A}^T\mathsf{B} - \mathsf{b}\mathsf{b}^T .$$

The method is said to be **algebraically stable** if the matrices B and M are both positive semidefinite, *i.e.*, $\mathsf{B} \geq 0$ and $\mathsf{M} \geq 0$ for all . Algebraic stability can be seen to ensure that approximations to solutions of nonlinear systems of differential equations exhibit acceptable numerical behaviour. For example, the Gauss–Runge–Kutta methods discussed in the last section are algebraically stable. For further details, see, for example,

, - - , *Stability of Runge–Kutta Methods for Stiff Nonlinear Differential Equations,* North-Holland, Amsterdam, 1984.

Exercises

12.1 Verify that the following functions satisfy a Lipschitz condition on the respective intervals and find the associated Lipschitz constants:

(a) $f(x,y) = 2yx^{-}$, x $[1,$ $)$;
(b) $f(x,y) = e^{-}$ $\tan^{-}$ y , x $[1,$ $)$;
(c) $f(x,y) = 2y(1+y)^{-}$ $(1+e^{-|}$ $|)$, x $($, $)$.

12.2 Suppose that m is a fixed positive integer. Show that the initial value problem

$$y' = y \qquad , \qquad y(0) = 0 ,$$

has infinitely many continuously differentiable solutions. Why does this not contradict Picard's Theorem?

12.3 Write down the solution y of the initial value problem

$$y' = py + q , \quad y(0) = 1 ,$$

where p and q are constants. Suppose that the method in the proof of Picard's Theorem is used to generate the sequence of approximations y (x), $n = 0, 1, 2, \ldots$; show that y (x) is a polynomial of degree n, and consists of the first $n+1$ terms in the series expansion of $y(x)$ in powers of x.

12.4 Show that Euler's method fails to approximate the solution $y(x) = (4x/5)$ of the initial value problem $y' = y$, $y(0) = 0$. Justify your answer.

Consider approximating the same problem with the implicit Euler method. Show that there is a solution of the form y $=$ $(C$ $h)$, n 0, with C $= 0$ and C $= 1$ and C > 1 for all n 2.

12.5 Write down Euler's method for the solution of the problem

$$y' = xe^{-} \qquad 5y , \qquad y(0) = 0$$

on the interval $[0, 1]$ with step size $h = 1/N$. Denoting by y the resulting approximation to $y(1)$, show that y $y(1)$ as N .

12.6 Consider the initial value problem

$$y' = \ln\ln(4+y) , \qquad x \quad [0,1] , \qquad y(0) = 1 ,$$

and the sequence (y_n), $n = 0, 1, \ldots, N$, generated by the Euler method

$$y_{n+1} = y_n + h \ln \ln(4 + y_n), \qquad n = 0, 1, \ldots, N-1, \qquad y_0 = 1,$$

using the mesh points $x_n = nh$, $n = 0, 1, \ldots, N$, with spacing $h = 1/N$.

(i) Let T_n denote the truncation error of Euler's method for this initial value problem at the point $x = x_n$. Show that $|T_n| \leq h/4$.

(ii) Verify that

$$|y(x_{n+1}) - y_{n+1}| \leq (1 + hL)\,|y(x_n) - y_n| + h\,|T_n|$$

for $n = 0, 1, \ldots, N-1$, where $L = 1/(2 \ln 4)$.

(iii) Find a positive integer N, as small as possible, such that

$$\max_{0 \leq n \leq N} |y(x_n) - y_n| \leq 10^{-3}$$

whenever $N \geq N$.

12.7 Define the truncation error T_n of the trapezium rule method

$$y_{n+1} = y_n + \frac{1}{2}h\,(f_{n+1} + f_n)$$

for the numerical solution of $y' = f(x, y)$ with $y(0) = y_0$ given, where $f_n = f(x_n, y_n)$ and $h = x_{n+1} - x_n$.

By integrating by parts the integral

$$\int_{x_n}^{x_{n+1}} (x - x_n)(x - x_{n+1}) y'''(x)\,dx,$$

or otherwise, show that

$$T_n = -\frac{1}{12}h^3 y'''(\xi_n)$$

for some ξ_n in the interval (x_n, x_{n+1}), where y is the solution of the initial value problem.

Suppose that f satisfies the Lipschitz condition

$$|f(x, u) - f(x, v)| \leq L|u - v|$$

for all real x, u, v, where L is a positive constant independent

of x, and that $y'''(x) \leq M$ for some positive constant M independent of x. Show that the global error $e_n = y(x_n) - y_n$ satisfies the inequality

$$|e_{n+1}| \leq |e_n| + \frac{1}{2}h|L(|e_n| + |e_{n+1}|) + \frac{1}{12}h^3 M.$$

For a constant step size $h > 0$ satisfying $hL < 2$, deduce that, if $y_n = y(x_n)$, then

$$|e_{n+1}| \leq \frac{h^3 M}{12L}\left[\frac{1 + \tfrac{1}{2}hL}{1 - hL}\right] - \frac{1}{12}h^3 M.$$

12.8 Show that the one-step method defined by

$$y_{n+1} = y_n + \tfrac{1}{2}h(k_1 + k_2),$$

where

$$k_1 = f(x_n, y_n), \quad k_2 = f(x_n + h, y_n + hk_1)$$

is consistent and has truncation error

$$T_n = -\tfrac{1}{12}h^2 f_{yy}(f_x + f_y f) - \tfrac{1}{6}(f_{xx} + 2f_{xy} f + f_{yy} f^2) + O(h^3).$$

12.9 When the classical fourth-order Runge–Kutta method is applied to the differential equation $y' = \lambda y$, where λ is a constant, show that

$$y_{n+1} = (1 + h\lambda + \tfrac{1}{2}h^2\lambda^2 + \tfrac{1}{6}h^3\lambda^3 + \tfrac{1}{24}h^4\lambda^4)y_n.$$

Compare this with the Taylor series expansion of $y(x_n)$ $= y(x_n + h)$ about the point $x = x_n$.

12.10 Consider the one-step method

$$y_{n+1} = y_n + \alpha h f(x_n, y_n) + \beta h f(x_n + \gamma h, y_n + \gamma h f(x_n, y_n)),$$

where α, β and γ are real parameters and $h > 0$. Show that the method is consistent if, and only if, $\alpha + \beta = 1$. Show also that the order of the method cannot exceed 2.

Suppose that a second-order method of the above form is applied to the initial value problem $y' = -\lambda y$, $y(0) = 1$, where λ is a positive real number. Show that the sequence $(y_n)_{n \geq 0}$ is bounded if, and only if, $h \leq 2/\lambda$. Show further that, for such λ,

$$y(x_n) - y_n \leq \frac{1}{6}\lambda^3 h^2 x_n e^{\lambda x_n}, \quad n \geq 0.$$

12.11 Find the values of α and β so that the three-step method

$$y \quad + \alpha(y \quad y \quad) \quad y \quad = h\beta(f \quad + f \quad)$$

has order of accuracy 4, and show that the resulting method is *not* zero-stable.

12.12 Consider approximating the initial value problem $y' = f(x, y)$, $y(0) = y$ by the linear multistep method

$$y \quad + by \_ \; + ay \_ \; = hf(x \; , y \;)$$

on the regular mesh $x \; = nh$ where a and b are constants.

(i) For a certain (unique) choice of a and b, this method is consistent. Find these values of a and b and verify that the order of accuracy is 1.

(ii) Although the method is consistent for the choice of a and b from part (i), the numerical solution it generates will not, in general, converge to the solution of the initial value problem as $h \quad 0$, because the method is not zero-stable. Show that the method is not zero-stable for these a and b, and describe quantitatively what the unstable solutions will look like for small h.

12.13 Given that α is a positive real number, consider the linear two-step method

$$y \qquad \alpha y \; = \frac{h}{3}[f(x \quad , y \quad) + 4f(x \quad , y \quad) + f(x \; , y \;)],$$

on the mesh $x : x \; = x \; + nh, n = 1, 2, \ldots, N$ of spacing h, $h > 0$. Determine the set of all α such that the method is zero-stable. Find α such that the order of accuracy is as high as possible; is the method convergent for this value of α?

12.14 Which of the following linear multistep methods for the solution of the initial value problem $y' = f(x, y)$, $y(0)$ given, are zero-stable?

(a) $y \qquad y \; = hf \;$,
(b) $y \quad + y \quad 2y \_ \; = h(f \quad + f \; + f \_ \;)$,
(c) $y \qquad y \_ \; = -h(f \quad + 4f \; + f \_ \;)$,
(d) $y \qquad y \; = -h(3f \quad f \_ \;)$,
(e) $y \qquad y \; = —h(5f \quad + 8f \quad f \_ \;)$.

For the methods under (a) and (c) explore absolute stability when applied to the differential equation $y' = \lambda y$ with $\lambda < 0$.

$-(3 \quad 3)$	$- \quad —(3 \quad 2 \quad 3)$
$-(3+ \quad 3)$	$—(3+2 \quad 3) \quad -$
	$- \quad -$

deduce that $\mathsf{y} \quad = \mathsf{R}(\lambda\mathsf{h})\mathsf{y}$, where

$$\mathsf{R}(\lambda\mathsf{h}) = \frac{1 + -\lambda\mathsf{h} + —\lambda \ \mathsf{h}}{1 \quad -\lambda\mathsf{h} + —\lambda \ \mathsf{h}}.$$

By writing $\mathsf{R}(\mathsf{z})$ in the factorised form $(\mathsf{z}+\mathsf{p})(\mathsf{z}+\mathsf{q})/(\mathsf{z} \quad \mathsf{p})(\mathsf{z} \quad \mathsf{q})$, deduce that this Runge–Kutta method is A-stable.

13

Boundary value problems for ODEs

13.1 Introduction

In the previous chapter we discussed numerical methods for initial value problems in which all the associated side conditions for a system of differential equations are prescribed at the same point. Now we go on to consider problems where these conditions specify values at more than one point. Typically we require the solution on an interval $[a, b]$, and some conditions are given at a, and the rest at b, although more complicated situations are possible, involving three or more points.

We shall begin with the simplest case, of a second-order equation with one condition given at a and one at b. This problem is sufficient to introduce the basic ideas, and is of a type which arises quite often in practice.

We then go on to discuss the shooting method for the solution of more general problems.

13.2 A model problem

The simplest two-point boundary problem involves the second-order differential equation

$$y'' + r(x)y = f(x), \qquad a < x < b, \tag{13.1}$$

with the boundary conditions

$$y(a) = A, \quad y(b) = B, \tag{13.2}$$

where A and B are given real numbers. We shall assume that r and f are given real-valued functions, defined and continuous on the bounded closed interval $[a, b]$ of the real line, and that

$$r(x) \quad 0, \quad a \quad x \quad b.$$

The reason for this condition will appear later, in Theorem 13.4.

We shall construct a numerical approximation to the solution on a uniform mesh of points

$$x_j = a + jh, \quad j = 0, 1, \ldots, n, \quad h = (b - a)/n, \quad n \geq 2,$$

so that $x_0 = a$, $x_n = b$. The second derivative is approximated using the second central difference defined below.

Definition 13.1 *The central difference* δy *of* y *is defined by*

$$\delta y(x_j) = y(x_j + \tfrac{1}{2}h) - y(x_j - \tfrac{1}{2}h).$$

Higher-order differences are defined recursively by

$$\delta^k y(x_j) = \delta[\delta^{k-1} y(x_j)] = \delta^{k-1} y(x_j + \tfrac{1}{2}h) - \delta^{k-1} y(x_j - \tfrac{1}{2}h).$$

In particular, the second central difference may be written

$$\begin{aligned} \delta^2 y(x_j) &= \delta y(x_j + \tfrac{1}{2}h) - \delta y(x_j - \tfrac{1}{2}h) \\ &= y(x_j + h) - 2y(x_j) + y(x_j - h). \end{aligned}$$

Theorem 13.1 *(i) Suppose that* $y \in C^4[x_j - h, x_j + h]$, *i.e., that* y *has continuous fourth derivative on the interval* $[x_j - h, x_j + h]$. *Then, there exists a number* ξ *in* $(x_j - h, x_j + h)$ *such that*

$$\frac{\delta^2 y(x_j)}{h^2} = y''(x_j) + \frac{1}{12}h^2 y^{(4)}(\xi).$$

(ii) Suppose that $y \in C^5[x_j - h, x_j + h]$; *then, there exists a number* η *in* $(x_j - h, x_j + h)$ *such that*

$$\frac{\delta^2 y(x_j)}{h^2} = y''(x_j) + \frac{1}{12}h^2 y^{(4)}(x_j) + \frac{1}{360}h^4 y^{(6)}(\eta). \tag{13.3}$$

Proof (i) Taylor's Theorem shows that there exist numbers ξ_- and ξ_+ in the intervals $(x_j - h, x_j)$ and $(x_j, x_j + h)$, respectively, such that

$$\begin{aligned} y(x_j - h) &= y(x_j) - hy'(x_j) + \tfrac{1}{2}h^2 y''(x_j) - \tfrac{1}{6}h^3 y'''(x_j) + \tfrac{1}{24}h^4 y^{(4)}(\xi_-), \\ y(x_j + h) &= y(x_j) + hy'(x_j) + \tfrac{1}{2}h^2 y''(x_j) + \tfrac{1}{6}h^3 y'''(x_j) + \tfrac{1}{24}h^4 y^{(4)}(\xi_+). \end{aligned} \tag{13.4}$$

Since $y^{(4)}$ is continuous on $[x_j - h, x_j + h]$, there is a number ξ in (ξ_-, ξ_+), and thus also in $(x_j - h, x_j + h)$, such that

$$\tfrac{1}{2}(y^{(4)}(\xi_-) + y^{(4)}(\xi_+)) = y^{(4)}(\xi).$$

The required result is now obtained by adding the two equations (13.4) and dividing by h.

(ii) The proof is completely analogous, and is left to the reader as an exercise. (See Exercise 1.) □

We can now use the central difference approximation to construct the numerical solution. Writing Y_j for the numerical approximation to $y(x_j)$, we approximate the differential equation by

$$\frac{\delta^2 Y_j}{h^2} + r_j Y_j = f_j, \quad j = 1, 2, \ldots, n-1, \tag{13.5}$$

where we have used the notation $r_j = r(x_j)$, $f_j = f(x_j)$. Now, (13.5) is a system of $n-1$ linear algebraic equations for the $n-1$ unknowns Y_j, $j = 1, 2, \ldots, n-1$, with the boundary conditions specifying the values of Y_0 and Y_n,

$$Y_0 = A, \qquad Y_n = B. \tag{13.6}$$

The system may be written in matrix form as

$$M\mathbf{Y} = \mathbf{g},$$

where $\mathbf{Y}$, $\mathbf{g}$ and, for $n \geq 4$, the matrix M $(n-2) \times (n-2)$ is tridiagonal. Here $\mathbf{Y} = (Y_1, \ldots, Y_{n-1})^T$, the nonzero elements of M are

$$M_{1,1} = 2/h^2 + r_1, \qquad M_{1,2} = M_{2,1} = -1/h^2, \tag{13.7}$$

and the elements of the column vector $\mathbf{g}$ on the right-hand side are

$$g_1 = f_1 + A/h^2, \quad g_{n-1} = f_{n-1} + B/h^2, \quad g_j = f_j, \quad j = 2, 3, \ldots, n-2.$$

Note how the known boundary values Y_0 and Y_n have been transferred to the right-hand side, and appear in the first and last elements of $\mathbf{g}$. The solution of this system is very easy, using the algorithm for tridiagonal matrices described in Section 3.3. Using the fact that $r(x) \geq 0$, we see that the off-diagonal elements of M are negative, the diagonal elements are positive, and in each row the diagonal element is at least as large as the sum of absolute values of the off-diagonal elements. Theorem 3.4 implies that no row interchanges are needed in the calculation, and that the matrix M is nonsingular. The calculation is therefore very straightforward and efficient, and requires very little computational time, even for a mesh which may contain several hundred points.

Now, $y(x)$ and Y satisfy

$$\begin{aligned} L(y(x)) &= f + T, \qquad j = 1, 2, \ldots, n-1, \\ L(Y) &= f, \qquad\quad j = 1, 2, \ldots, n-1, \end{aligned}$$

from the definition of truncation error and (13.5); hence, by subtraction,

$$L(e) = T, \qquad j = 1, 2, \ldots, n-1,$$

with the boundary conditions $e_0 = e_n = 0$. We must now use the bound on T to derive a bound on the error e. This will be achieved by means of the following theorem.

Theorem 13.3 (**Maximum Principle**) *Let* a_j, b_j, c_j, $j = 0, 1, \ldots, n$, *be positive real numbers such that* $b_j \geq a_j + c_j$, *and suppose that* u_j, $j = 0, 1, \ldots, n$, *are real numbers such that*

$$a_j u_{j-1} - b_j u_j + c_j u_{j+1} \leq 0, \qquad j = 1, 2, \ldots, n-1.$$

Then, $u_j \leq K$, $j = 0, 1, \ldots, n$, *where* $K = \max(u_0, u_n, 0)$.

Proof Let $u_m = \max(u_0, u_1, \ldots, u_n)$; then if $m = 0$, $m = n$, or $u_m \leq 0$ the result is trivial. Suppose then that $1 \leq m \leq n-1$, and that $u_m > 0$. Since u_m is the maximum of the u_j, we know that

$$u_m \geq u_{m-1}, \qquad u_m \geq u_{m+1}.$$

Hence

$$\begin{aligned} b_m u_m &\leq a_m u_{m-1} + c_m u_{m+1} \\ &\leq a_m u_m + c_m u_m \\ &\leq b_m u_m, \end{aligned}$$

since $u_m > 0$. This means that equality holds throughout, so that $u_{m-1} = u_m = u_{m+1}$. We can then apply the same argument to both u_{m-1} and u_{m+1}, continuing until we find that either $u_0 = u_m$ or $u_n = u_m$. Thus, in this case $u_0 = u_m = \max(u_0, u_n, \ldots, u_n)$, as required. □

Theorem 13.4 *Suppose that the solution* y *of the boundary value problem (13.1), (13.2) has a continuous fourth derivative on* $[a, b]$, *and that* Y_j, $j = 0, 1, \ldots, n$, *is the solution of the central difference approximation (13.5), (13.6). Then,*

$$\max_{a \leq x \leq b} |y(x) - Y(x)| \leq \tfrac{1}{96} h^4 (b - a) M. \tag{13.10}$$

Proof Let $e_j = y(x_j) - Y_j$. We have already seen that $L(e_j) = T_j$, $j = 1, 2, \ldots, n-1$. Defining

$$\phi_j = C(1/2)(2j-n)h - n h^2, \qquad j = 0, 1, \ldots, n, \tag{13.11}$$

where C is a constant, we see that

$$\begin{aligned} L(\phi_j) &= C(1/2)(2j-2-n) - 2(2j-n) + (2j+2-n) + r_j\phi_j \\ &= 8C + r_j\phi_j, \qquad j = 1, 2, \ldots, n-1. \end{aligned}$$

Hence

$$L(e_j + \phi_j) = T_j - 8C + r_j\phi_j, \qquad j = 1, 2, \ldots, n-1.$$

If we choose $C = T/8$ with $T = -h^2 M$, we see that $L(e_j + \phi_j) \le 0$, since $T_j \le T$, $r_j \ge 0$ and $\phi_j \ge 0$, and L satisfies the conditions of the Maximum Principle. Now,

$$e_0 + \phi_0 = e_n + \phi_n = 0,$$

so that, according to Theorem 13.3, $e_j + \phi_j \le 0$ for $j = 0, 1, \ldots, n$. However, $Cn h - \phi_j \le 0$, so we have the result

$$e_j \le Cn h - \phi_j = -(b-a) T = -h^2 (b-a) M, \qquad j = 0, 1, \ldots, n.$$

By applying the same argument to $L(-e_j + \phi_j)$ we find that

$$-e_j \le -h^2 (b-a) M, \qquad j = 0, 1, \ldots, n.$$

Combining these upper bounds for e_j and $-e_j$ gives the required result. □

The function ϕ_j defined by (13.11) is called a **comparison function**. An alternative proof of Theorem 13.4, based on the properties of monotone matrices, can be given by using the result in Exercise 2. Notice that the condition $r(x) \ge 0$ is used in the application of the Maximum Principle in the above proof.

This theorem shows that, provided the solution y has a continuous fourth derivative, the numerical method is **convergent**, that is

$$\max_{a \le x \le b} |y(x_j) - Y_j| \to 0 \quad \text{as } n \to \infty$$

(or, equivalently, as $h = (b-a)/n \to 0$). This means that we can obtain any required accuracy by choosing n sufficiently large.

Note that the approximation to $y''(x)$ at $x = x$ may be written

$$\frac{-[\delta y(x + -h) + \delta y(x \quad -h)]}{h}.$$

For $j = 1, 2, \ldots, n \quad 1$, we define the truncation error T as in Definition 13.2. In addition, since we shall now also incur an error in the approximation of the boundary condition at $x = a$, we define

$$T = \frac{2(1 + \alpha h)}{h} + r \quad y(0) \quad \frac{2}{h} y(h) \quad f + \frac{2}{h} A.$$

The aim of our next result is to quantify the size of the truncation error in terms of the mesh size h.

Theorem 13.6 *Suppose that the solution y to the boundary value problem (13.1), (13.2) has a continuous fourth derivative on the closed interval $[a \quad h, b]$. Then, the truncation error of the central difference approximation to (13.1) with boundary conditions (13.12) may be written*

$$\begin{aligned} T &= \quad -h\, y\ (\xi), \qquad j = 1, 2, \ldots, n \quad 1, \\ T &= \quad -h\, y\ (\xi) \quad -h y''''(\chi), \end{aligned}$$

for some value of ξ in the interval $(x_- , x \quad)$, $1 \quad j \quad n \quad 1$, and some value χ in the interval (x_- , x) where $x_- = a \quad h$.

Proof For $j = 1, 2, \ldots, n \quad 1$, this is the same result as in Theorem 13.2. When $j = 0$, we find that

$$\begin{aligned} T &= \frac{2(1 + \alpha h)}{h} + r \quad y(0) \quad \frac{2}{h} y(h) \quad f + \frac{2}{h} A \\ &= \frac{y(h) \quad 2y(0) + y(\quad h)}{h} + r(0) y(0) \quad f(0) \\ &\quad \frac{2}{h} \quad \frac{y(h) \quad y(\quad h)}{2h} \quad \alpha y(0) \quad A \\ &= \quad -h\, y\ (\xi) \quad \frac{}{*} -h\, y''''(\chi), \end{aligned}$$

where we have used Theorem 13.5. □

Theorem 13.7 *Suppose that the solution y of (13.1) with the boundary conditions (13.12) has a continuous fourth derivative on the interval $[a \quad h, b]$; then, the numerical solution obtained from the central difference*

approximation satisfies

$$\max_{\leq\ \leq} \mathsf{y}(\mathsf{x}\ ) \quad \mathsf{Y} \qquad \mathsf{h}^{\ 1} — (\mathfrak{b} \quad \mathsf{a})\ \mathsf{M}\ + \text{-}(\mathfrak{b} \quad \mathsf{a})\mathsf{M}^{\ 2}\ .$$

Proof The proof is very similar to that of Theorem 13.4, but requires the use of a more complicated comparison function ϕ . Let us define

$$\mathsf{L}^{\square}(\mathsf{u}\ ) \quad = \qquad \frac{\delta\ \mathsf{u}}{\mathsf{h}} + \mathsf{r}\ \mathsf{u}$$

In particular,

$$e + \phi \quad \max e + \phi , 0 . \tag{13.15}$$

However, $L^{\square}(e + \phi) \quad 0$; thus, by the definition of $L^{\square}(e + \phi)$,

$$e + \phi \quad \frac{2}{2(1+\alpha h) + h\ r} (e + \phi) .$$

On writing $\delta = 2/(2(1+\alpha h) + h\ r)$ and noting that, since $\alpha > 0$ and $r \quad 0$, we have $0 < \delta < 1$, it follows that

$$e + \phi \quad \delta(e + \phi) . \tag{13.16}$$

Inserting this inequality into the left-hand side of (13.15), we find that

$$e + \phi \quad \max \delta(e + \phi), 0 .$$

If $e + \phi$ were positive, this inequality and the fact that $0 < \delta < 1$ would imply $e + \phi \quad 0$, leading to a contradiction. Therefore, $e + \phi \quad 0$. Returning with this information to (13.14), we conclude that $e + \phi \quad 0$ for $j = 0, 1, \ldots, n$, and the rest of the proof then follows as in the proof of Theorem 13.1. □

13.5 The general self-adjoint problem

The general self-adjoint boundary value problem is

$$\frac{d}{dx} \quad p(x)\frac{dy}{dx} \quad + r(x)y = f(x) , \qquad a < x < b , \tag{13.17}$$

where r and f are real-valued functions, defined and continuous on $[a, b]$, p is a real-valued continuously differentiable function on $[a, b]$, $r(x) \quad 0$ and $p(x) \quad c > 0$. We shall consider only the case where the boundary conditions prescribe the values of y at each end,

$$y(a) = A , \qquad y(b) = B . \tag{13.18}$$

The central difference approximation to the equation (13.17) may be written

$$\frac{\delta(p\ \delta Y)}{h} + r\ Y \ = f \ , \qquad j = 1, 2, \ldots, n \quad 1 ,$$

or, in detail,

$$\frac{p \quad (Y \quad Y) \quad p_{-} \quad (Y \quad Y_{-})}{h} + r\ Y \ = f \ , \tag{13.19}$$

for $j = 1, 2, \ldots, n$, and is supplemented by the boundary conditions

$$Y_0 = A, \qquad Y_n = B. \tag{13.20}$$

It is easy to see that this represents a system of linear equations for the unknowns $Y_0, Y_1, \ldots, Y_{n-1}$, and that the matrix of the system is tridiagonal and diagonally dominant, just as it was in the special case (13.1), which corresponds to $p(x) \equiv 1$. The solution of the system is therefore a very simple matter.

Next, we consider the error analysis of the difference scheme (13.19), (13.20). We begin by quantifying the size of the truncation error

$$T_j = \frac{\delta(p_j \delta y(x_j))}{h} + r_j y(x_j) - f_j, \qquad j = 1, 2, \ldots, n-1,$$

in terms of the mesh size h.

Lemma 13.1 *Suppose that* $p \in C^1[a,b]$ *and* $y \in C^4[a,b]$. *The truncation error* T_j *of the central difference approximation (13.19) then satisfies*

$$T_j - T(x_j) = -\frac{h^2}{12} \max_{x \in [a,b]} \left| (py')''(x) + p(x) y''''(x) \right| + 2p_j y''''(\xi_j) \Big|,$$

for $j = 1, 2, \ldots, n-1$.

Proof By expanding in Taylor series as we have done before, we find that

$$\begin{aligned} p_{j+\frac{1}{2}}[y(x_{j+1}) - y(x_j)] &= p_{j+\frac{1}{2}}[hy'_j + \tfrac{1}{2}h^2 y''(x_j) + \tfrac{1}{6}h^3 y'''(\xi_j)], \\ p_{j-\frac{1}{2}}[y(x_j) - y(x_{j-1})] &= p_{j-\frac{1}{2}}[hy'_j - \tfrac{1}{2}h^2 y''(x_j) + \tfrac{1}{6}h^3 y'''(\xi_j)], \end{aligned}$$

where $\xi_j \in (x_j, x_{j+1})$ and $\xi_j \in (x_{j-1}, x_j)$. The first term in the difference of these expressions gives, in the same way,

$$h[p_{j+\frac{1}{2}} y'(x_{j+1}) - p_{j-\frac{1}{2}} y'(x_{j-1})] = h[h(py')'(x_j) + \tfrac{1}{6}h^2 (py')'''(\xi_j)]$$

where $\xi_j \in (x_{j-1}, x_{j+1})$. For the other term we can write

$$\begin{aligned} -h^2 &\left[p_{j+\frac{1}{2}} y'''(\xi_j) - p_{j-\frac{1}{2}} y'''(\xi_j) \right] \\ &= -h^2 \left[\frac{1}{h}(p_{j+\frac{1}{2}} - p_{j-\frac{1}{2}}) y'''(\xi_j) + p_{j-\frac{1}{2}} \left[\frac{1}{2} y'''(\xi_j) - y'''(\xi_j) \right] \right] \\ &\quad - h^2 \left[hp'(\xi_j) y'''(\xi_j) + p_{j-\frac{1}{2}} \cdot 2hy''''(\xi_j) \right], \end{aligned}$$

since $|\xi_j - \xi_j| < 2h$. Here, $\xi_j \in (x_{j-1}, x_{j+1})$ and ξ_j lies between ξ_j and ξ_j. The required bound follows immediately. □

that Y_j, $j = 0, 1, \ldots, n$, *is the solution of the central difference approximation (13.19), (13.20). Then, with* T *as in Lemma 13.1,*

$$\max_{a \le x \le b} |y(x_j) - Y_j| \le \frac{1}{12} h^2 T. \tag{13.21}$$

Proof The proof of this theorem follows that of Theorem 13.4, using the bound from Lemma 13.1 on the truncation error and the comparison function ϕ from Lemma 13.2. The details are left as an exercise. □

13.6 The Sturm–Liouville eigenvalue problem

Suppose that r is a real-valued function, defined and continuous on the closed interval $[a, b]$, p is a real-valued function, defined and continuously differentiable on $[a, b]$, and $r(x) \ge 0$, $p(x) \ge c > 0$ for all $x \in [a, b]$. The differential equation

$$\frac{d}{dx}\left(p(x)\frac{dy}{dx}\right) + r(x)y = \lambda y, \quad a < x < b, \tag{13.22}$$

with homogeneous boundary conditions $y(a) = y(b) = 0$, has only the trivial solution $y \equiv 0$, except for an infinite sequence of positive *eigenvalues* $\lambda = \lambda_m$, $m = 1, 2, \ldots$. We shall now consider a numerical method for finding these eigenvalues and the corresponding *eigenfunctions*, $y_m(x)$, $m = 1, 2, \ldots$.

In the simple case where $p(x) \equiv 1$ and $r(x) \equiv 0$ the solution to this problem is, of course, $\lambda_m = [m\pi/(b - a)]^2$, $y_m(x) = A \sin m\pi t$, $m = 1, 2, \ldots$, where A is a nonzero constant and $t = (x - a)/(b - a)$.

Using the same finite difference approximation as in the previous section, we obtain the equations

$$\frac{p_j(Y_{j+1} - Y_j) - p_{j-1}(Y_j - Y_{j-1})}{h^2} + r_j Y_j = \Lambda Y_j, \quad j = 1, 2, \ldots, n - 1.$$

Together with the boundary conditions $Y_0 = Y_n = 0$, this shows that Λ is an eigenvalue of a symmetric tridiagonal matrix M whose entries are

$$M_{jj} = \frac{p_j + p_{j-1}}{h^2} + r_j, \quad 1 \le j \le n - 1,$$

$$M_{j-1} = -\frac{p_{j-1}}{h^2}, \quad 2 \le j \le n, \quad M_{j,j+1} = -\frac{p_j}{h^2}, \quad 1 \le j \le n - 1,$$

and the approximate function values Y are the elements of the corresponding eigenvector. This algebraic eigenvalue problem is easily solved by the method described in Chapter 5.

The boundary value problems which we have discussed so far have all had a unique solution. The eigenvalue problem (13.22) has an infinite number of solutions, and the mesh used in the numerical computation has to be chosen to adequately represent the eigenfunctions required – the computation can obviously only find a finite number of them. The matrix M has $n \quad 1$ eigenvalues and eigenvectors and, as we shall see, it will normally give a good approximation to the first few eigenvalues, $\lambda , \lambda , \ldots$, and a much less accurate approximation to $\lambda \_ .$

To analyse the error in the eigenvalue we proceed as before, by defining the truncation error

$$T = \frac{p \quad (y \quad y) \quad p\_ \quad (y \quad y\_)}{h} + r\, y \quad \lambda y ,$$

$$j = 1, 2, \ldots, n \quad 1,$$

where $y = y(x)$. These equations can now be written

$$\begin{aligned} (M \quad \Lambda I)\mathbf{Y} &= \mathbf{0}, \\ (M \quad \lambda I) &= \mathbf{T}, \end{aligned}$$

where

$$\begin{aligned} \mathbf{Y} &= (Y , \ldots, Y\_) , \\ &= (y , \ldots, y\_) , \\ \mathbf{T} &= (T , \ldots, T\_) . \end{aligned}$$

Theorem 5.15 of Chapter 5 applies to this problem, and shows that one of the eigenvalues, Λ , of the matrix M satisfies

$$\lambda \quad \Lambda \quad \mathbf{T} \quad / \quad . \tag{13.23}$$

In the simpler case where $p(x) \quad 1$ and $r(x) \quad 0$ the truncation error is

$$T = \quad -h\, y\ (\xi), \qquad \xi \quad (x\_ , x \quad),$$

so the numerical method has evaluated the eigenvalue with error less than

$$-h \quad \bar{} \quad [y\ (\xi)] \quad \bar{} \quad [y(x)] \quad \bar{} \quad .$$

Since the mth eigenfunction y is given by

$$y(x) = y \quad (x) = \sin(m\pi(x \quad a)/(b \quad a)), \qquad x \quad (a,b),$$

we see that

$$y \ (x) = \frac{m\pi}{b \quad a} \quad y(x), \qquad x \quad (a,b).$$

This shows that, for example, the error in the tenth eigenvalue, corresponding to $m = 10$, is likely to be about 10 times larger than the error in the first eigenvalue; more generally, to evaluate higher eigenvalues of the equation will require the use of a smaller interval h.

13.7 The shooting method

The methods we have described for the linear boundary value problem may be extended to nonlinear differential equations. We shall not discuss how this is done; instead, we shall describe an alternative approach, called the **shooting method**. We shall consider the nonlinear model problem

$$y'' = f(x,y), \quad a < x < b, \qquad y(a) = A, \quad y(b) = B,$$

where we assume that the function $f(x,y)$ is continuous and differentiable, and that

$$\frac{\partial f}{\partial y}(x,y) \quad 0, \qquad a < x < b, \quad y \qquad .$$

The central idea of the method is to replace the boundary value problem under consideration by an initial value problem of the form

$$y'' = f(x,y), \quad a < x \quad b, \qquad y(a) = A, \quad y'(a) = t,$$

where t is to be chosen in such a way that $y(b) = B$. This can be thought of as a problem of trying to determine the angle of inclination $\tan^{-} t$ of a loaded gun, so that, when shot from height A at the point $x = a$, the bullet hits the target placed at height B at the point $x = b$. Hence the name, shooting method.

Once the boundary value problem has been transformed into such an 'equivalent' initial value problem, any of the methods for the numerical solution of initial value problems discussed in Chapter 12 can be applied to find a numerical solution. Thus, in particular, the costly exercise of solving a large system of nonlinear equations, arising from a direct finite

difference approximation of the nonlinear boundary value problem, can be completely avoided.

If we write

$$y'(a) = t\,,$$

a numerical solution of the differential equation with the initial conditions $y(a) = A, y'(a) = t$ can be obtained by any of the methods of Chapter 12. This solution will depend on t, and we may write it as $y(x;t)$. In particular the value at $x = b$ will be a function of t,

$$y(b;t) = \psi(t)\,. \tag{13.24}$$

The solution of the nonlinear boundary value problem therefore reduces to the determination of the value of t for which the boundary condition at $x = b$ is also satisfied, *i.e.*,

$$\psi(t) \quad B = 0\,.$$

There are a number of well-known methods for the solution of equations of this form; Newton's method is an obvious example. Generally, we shall not, of course, have a closed form expression for the function $\psi(t)$, in general, but this is not necessary; all that is needed is a numerical algorithm to calculate the value of $\psi(t)$ for a given value of t, and this we have. To use Newton's method we shall also need to be able to calculate the value of $\psi'(t)$, and this is easily done.

The function $y(x;t)$ is defined, for all t, as the solution of the initial value problem

$$y''(x;t) = f(x, y(x;t))\,, \qquad y(a;t) = A\,, \quad y'(a;t) = t\,, \tag{13.25}$$

where $'$ and $''$ indicate differentiation with respect to the variable x. We can differentiate these throughout with respect to t, giving

$$\frac{\partial}{\partial t}y''(x;t) = \frac{\partial f}{\partial y}(x, y(x;t))\frac{\partial y}{\partial t}(x;t)\,, \quad \frac{\partial y}{\partial t}(a;t) = 0\,, \quad \frac{\partial y'}{\partial t}(a;t) = 1\,.$$

Writing

$$w(x,t) = \frac{\partial y}{\partial t}(x;t)\,,$$

and interchanging the order of differentiation, we find that $w(x;t)$ may be obtained as the solution of the initial value problem

$$w''(x;t) = w(x;t)\,\frac{\partial f}{\partial y}(x, y(x;t))\,, \qquad w(a;t) = 0\,, \quad w'(a;t) = 1\,. \tag{13.26}$$

By virtue of (13.24), the required derivative is then given by

$$\psi'(t) = w(b, t)\,.$$

To implement this method, it is convenient to solve the two initial value problems, (13.25) and (13.26), in tandem, by writing them as a system of four simultaneous first-order differential equations:

Proof Suppose that the solution of the system of differential equations with $t = t$ is $u(x;t)$, $i = 1,2,3,4$, and the corresponding numerical solution is $v(t)$, $i = 1,2,3,4$, $j = 1,2,\ldots,n$; then

$$u(x;t) \quad v(t) \quad C(t)h .$$

Moreover $v(t) \quad B \quad \varepsilon$, so that

$$u(b;t) \quad B \quad u(b;t) \quad v(t) + v(t) \quad B \quad C(t)h + \varepsilon . \tag{13.28}$$

Let us write $\eta(x;t) = y(x) \quad u(x;t)$; by subtraction we see that

$$\begin{aligned} \eta''(x;t) &= y''(x) \quad u''(x;t) \\ &= f(x,y(x)) \quad f(x,u(x,t)) \\ &= \eta(x;t)\frac{\partial f}{\partial y}(x,\xi(x;t)) , \end{aligned}$$

where $\xi(x;t)$ lies between $u(x;t)$ and $y(x)$.

Suppose that $\eta'(a;t) > 0$; since $\eta(a;t) = 0$, there is some interval to the right of a in which $\eta(x;t) > 0$. Then, either $\eta(x;t) > 0$ for the whole of $(a,b]$, or there is a value c such that $a < c < b$ and $\eta(c;t) = 0$. In the latter case, $\eta'(x;t)$ must vanish at some point $x = d$ between a and c. However, in the interval $[a,d]$, $\eta(x;t) > 0$ and $\partial f/\partial y \quad 0$, so that $\eta''(x;t) > 0$. Consequently, in the interval $[a,d]$, $\eta'(x;t) > \eta'(a;t) > 0$, and we have a contradiction. Thus, $\eta(x;t) > 0$ for all $a < x \quad b$. It then follows that $\eta''(x;t)$, and hence also $\eta'(x;t)$ are positive on the whole interval $[a,b]$, which means that $x \quad \eta(x;t)$ is monotonic increasing on $[a,b]$. If we had begun with the assumption that $\eta'(a;t) < 0$ an analogous argument shows that $x \quad \eta(x;t)$ would have been monotonic decreasing on $[a,b]$. It is left to the reader to discuss the trivial case when $\eta'(a,t) = 0$.

In any case,

$$\eta(x;t) \quad \eta(b;t) , \quad a \quad x \quad b ,$$

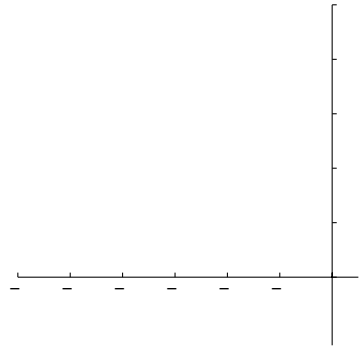
and therefore, since $y(b) = B$ and recalling (13.28),

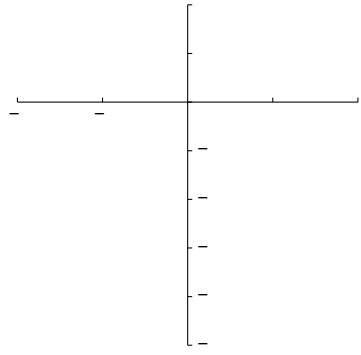
$$y(x) \quad u(x;t) \quad B \quad u(b;t) \quad C(t)h + \varepsilon .$$

Thus, finally,

$$y(x) \quad v(t) \quad y(x) \quad u(x;t) + u(x;t) \quad v(t) \quad C(t)h + \varepsilon + C(t)h , \quad j = 1,2,\ldots,n ,$$

and hence the desired bound. □





& 5 , B , *& $+

13.8 Notes

The following books are standard texts on the subject of numerical approximation of boundary value problems:

! , , *Numerical Methods for Two-Point Boundary Value Problems,* Reprint of the 1968 original published by Blaisdell, Dover, New York, 1992.

! , , *Numerical Solution of Two-Point Boundary Value Problems*, SIAM, Philadelphia, fourth printing, 1990.

A more recent survey of the subject is found in

0 +) " % + + + 4 % % " *Numerical Solution of Boundary Value Problems for Ordinary Differential Equations,* Corrected reprint of the 1988 original, Classics in Applied Mathematics, 13, SIAM, Philadelphia, 1995.

In practical implementations of the shooting method into mathematical software (see, for example, Appendix A in the Ascher *et al.* book), the interval [a, b] is subdivided into smaller intervals on each of which the shooting method is applied with appropriately chosen initial values. The 'initial' conditions on the subintervals are then simultaneously adjusted in order to satisfy the boundary conditions and appropriate continuity conditions at the points of the subdivision. From the practical viewpoint, this extension of the basic shooting method considered in this chapter is extremely important: the various difficulties which may arise in the implementation of the basic method (such as, for example, growth of the

solution to the initial value problem over the interval $[a, b]$, leading to loss of accuracy in the solution of the equation $\psi(t) = B$) are discussed, for example, in Section 2.4 of the 1992 book by Keller.

Sturm–Liouville problems originated in a paper of Jacques Charles François Sturm: Sur les équations différentielles linéaires du second ordre, *J. Math. Pures Appl.* **1**, 106–186, 1836, in Joseph Liouville's newly founded journal. Sturm's paper was followed by a series of articles by Sturm and Liouville in subsequent volumes of the journal. They examined general linear second-order differential equations, the properties of their eigenvalues, the behaviour of the eigenfunctions and the series expansion of arbitrary functions in terms of these eigenfunctions. An extensive survey of the theory and numerical analysis of Sturm–Liouville problems can be found in

$ " *Numerical Solution for Sturm–Liouville Problems,* Oxford University Press Monographs in Numerical Analysis, Clarendon Press, Oxford, 1993.

See also Section 11.3, page 478, of the Ascher *et al.* book cited above.

Exercises

13.1 Suppose that $y \quad C\ [x \quad h, x + h]$; show that there exists a real number η in $(x \quad h, x + h)$ such that

$$\frac{\delta\ y(x)}{h} = y^{\square\square}(x) + —h\ y\ \ (x) + ——h\ y\ \ (\eta)\,.$$

13.2 Use Theorem 3.6 to show that the matrix M in (13.7) is monotone. Use the result of Exercise 4 to show that $M^{-} \quad \infty \quad -$.

13.3 On the interval $[a, b]$ the differential equation

$$y^{\square\square} + f(x)y = g(x)$$

is approximated by

$$\frac{\delta\ y}{h} + \beta_{-}\ y_{-} + \beta\ y + \beta\ y \quad = \beta_{-}\ g_{-} + \beta\ g + \beta\ g \quad ,$$

where β_{-}, β and β are constants. Assuming that the solution y has the appropriate number of continuous derivatives, show that the truncation error of this approximation may be written as follows:

(i) if $\beta_{-1} + \beta_0 + \beta_1 = 1$, then

$$T = (\beta_{-1} + \beta_0 + \beta_1)y''(x) + Z \quad h,$$

where $Z \quad (\beta_{-1} + \beta_1)M$;

(ii) if $\beta_{-1} + \beta_0 + \beta_1 = 1$ and $\beta_{-1} = \beta$, then

$$T = (\beta \quad \beta_{-1})hy'''(x) + Z \quad h ,$$

where $Z \quad [-(\beta_{-1} + \beta_1) + -]M$;

(iii) if $\beta_{-1} + \beta_0 + \beta_1 = 1$, $\beta = \beta_{-1}$ and $\beta = -$, then

$$T = (\beta \quad -)h\, y\ (x) + Z \quad h ,$$

where $Z \quad [-\beta + --]M$;

(iv) if $\beta_{-1} = \beta = -$ and $\beta = -$, then

$$T = --h\, y\ (x) + Z \quad h ,$$

where $Z \quad ---M$.

13.4 The approximation of Exercise 3 is used, with the values $\beta = \beta_{-1} = 1/12$, $\beta = 5/6$. Use Taylor's Theorem with integral remainder (Appendix, Theorem A.5) to show that the truncation error of this approximation may be written

$$T = \int_{-*}^{*} G(s)y\ (x + s)\, ds,$$

where

$$G(s) = (h \quad s)\ /5! \quad -h\ (h \quad s)\ /3!, \quad 0 \quad s \quad h,$$

with a similar expression for $h \quad s \quad 0$. Show that $G(s)$ 0 for all $s \quad [h, h]$, and hence use the Integral Mean Value Theorem to show that the truncation error can be expressed as

$$T = \frac{h}{240}y\ (\xi)$$

for some value of ξ in $(x \quad h, x + h)$.

13.5 Suppose that the solution of (13.1), (13.2) has a continuous sixth derivative on $[a, b]$, and that Y is the solution of the approximation used in Exercise 4. Show that

$$y(x) \quad Y \qquad \text{—}h\ (b \quad a)\ M\ , \qquad j = 0, \ldots, n\,,$$

provided that

$$h\ r(x) \quad 12\,, \qquad j = 1, \ldots, n \quad 1\,.$$

13.6 Complete the proof of Theorem 13.7.

13.7 Show that the solution of the boundary value problem

$$y'' + a\ y = 0\,, \quad y(\ \ 1) = 1\,, \qquad y(1) = 1\,,$$

is

$$y(x) = \frac{\cosh ax}{\cosh a}\,.$$

Use the identity

$$\cosh(x + h) + \cosh(x \quad h) = 2\cosh x\ \cosh h$$

to verify that the solution of the difference approximation (13.5) to this problem is

$$Y \ = \frac{\cosh \vartheta x}{\cosh \vartheta}\,,$$

where

$$\vartheta = (1/h)\cosh^{-}\ (1 + -a\ h\ )\,.$$

By expanding in Taylor series, show that

$$Y \ = y(x\ ) + \text{—}h\ a\ (\cosh ax \sinh a \quad x \sinh ax \cosh a)/(\cosh a)$$
$$+ \quad (h\ )\,.$$

Verify that this result is consistent with Theorem 13.4 when h is small.

13.8 Carry out a similar analysis as in Exercise 7 for the boundary value problem

$$y'' \quad a\ y = 0\,, \qquad y(0) = 0\,, \quad y(1) = 1\,,$$

and explain why in this case Theorem 13.4 cannot be used. What restriction is required on the value of a?

13.9 The eigenvalue problem

$$y'' = \lambda y, \qquad y(0) = y(1) = 0,$$

is approximated by

$$\frac{Y_{j-1} - 2Y_j + Y_{j+1}}{h^2} = \mu Y_j, \quad 1 \le j \le n-1, \quad Y_0 = Y_n = 0.$$

Show that the differential equation has solution $y = \sin m\pi x$, $\lambda = m^2\pi^2$ for any positive integer m. Show also that the difference approximation has solution $Y_j = \sin m\pi x_j$, $j = 0, 1, \ldots, n$, and give an expression for the corresponding value of μ. Use the fact that

$$1 - \cos\theta = \tfrac{1}{2}\theta^2 - \tfrac{1}{24}\xi\theta^4, \quad 0 < \xi < 1,$$

to show that $\lambda - \mu \le m^4\pi^4 h^2/12$, and compare with the bound given by (13.23).

14

The finite element method

14.1 Introduction: the model problem

In Chapter 13 we explored finite difference methods for the numerical solution of two-point boundary value problems. The present chapter is devoted to the foundations of the theory of finite element methods. For the sake of simplicity the exposition will be, at least initially, confined to the second-order ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + r(x)u = f(x), \quad a < x < b, \tag{14.1}$$

where $p \in C^1[a,b]$, $r \in C[a,b]$, $f \in L^2(a,b)$ and $p(x) \geq c > 0$, $r(x) \geq 0$ for all $x \in [a,b]$, subject to the boundary conditions

$$u(a) = A, \quad u(b) = B. \tag{14.2}$$

Later on in the chapter, in Section 14.5, we shall also consider the ordinary differential equation

$$\frac{\mathrm{d}}{\mathrm{d}x}\left(p(x)\frac{\mathrm{d}u}{\mathrm{d}x}\right) + q(x)\frac{\mathrm{d}u}{\mathrm{d}x} + r(x)u = f(x), \quad a < x < b, \tag{14.3}$$

subject to the boundary conditions (14.2). Indeed, much of the material discussed here can be extended to partial differential equations; for pointers to the relevant literature we refer to the Notes at the end of the chapter.

The finite element method was proposed in a paper by Richard Courant in the early 1940s, although the historical roots of the method can be traced back to earlier work by Galerkin in 1915; unfortunately, the relevance of Courant's article was not recognised at the time and the idea was forgotten. In the early 1950s the method was rediscovered by engineers, but its systematic mathematical analysis began only a decade later. Since then, the finite element method has been developed into one of the most general and powerful techniques for the numerical solution of differential equations which is widely used in engineering design and analysis.

Unlike finite difference schemes which seek to approximate the unknown analytical solution to a differential equation at a finite number of selected points, the grid points or mesh points in the computational domain, the finite element method supplies an approximation to the analytical solution in the form of a piecewise polynomial function, defined over the entire computational domain. For example, in the case of the boundary value problem (14.1), (14.2), the simplest finite element method uses a linear spline, defined over the interval $[a, b]$, to approximate the analytical solution u.

We shall consider two techniques for the construction of finite element approximations: the **Rayleigh–Ritz principle** and the **Galerkin principle**. In the case of the boundary value problem (14.1), (14.2) the approximations which stem from these two principles will be seen to coincide. We note, however, that since the Rayleigh–Ritz principle relies on the fact that the boundary value problem under consideration can be restated as a variational problem involving the minimisation of a certain quadratic functional over a function space, its use is restricted to *symmetric* boundary value problems, such as (14.1), (14.2) where (14.1) does not contain a first-derivative term; for example, the Rayleigh–Ritz principle is not applicable to (14.3), (14.2) unless $q(x) \equiv 0$. The precise sense in which the word *symmetric* is to be interpreted here will be clar-

14.2 Rayleigh–Ritz and Galerkin principles

The Rayleigh–Ritz principle relies on converting the boundary value problem (14.1), (14.2) into a variational problem involving the minimisation of a certain quadratic functional over a function space.

Let us define the quadratic functional $+ : \mathrm{H}_{\&}(\mathsf{a},\mathsf{b})$ by

$$+(\mathsf{w}) = - \int_{\mathsf{a}}^{\mathsf{b}} [\mathsf{p}(\mathsf{x})(\mathsf{w}^{\square}) + \mathsf{r}(\mathsf{x})\mathsf{w}\,]\mathrm{d}\mathsf{x} \int_{\mathsf{a}}^{\mathsf{b}} \mathsf{f}(\mathsf{x})\mathsf{w}(\mathsf{x})\mathrm{d}\mathsf{x}$$

where w $\mathrm{H}_{\&}(\mathsf{a},\mathsf{b})$, and consider the following *variational problem*:

(RR) find u $\mathrm{H}_{\&}(\mathsf{a},\mathsf{b})$ such that $+(\mathsf{u}) = \min_{\square\$} +(\mathsf{w})$,

which we shall henceforth refer to as the **Rayleigh–Ritz principle**. For the sake of notational simplicity we define

$$(\,(\mathsf{w},\mathsf{v}) = \int_{\mathsf{a}}^{\mathsf{b}} [\mathsf{p}(\mathsf{x})\,\mathsf{w}^{\square}(\mathsf{x})\mathsf{v}^{\square}(\mathsf{x}) + \mathsf{r}(\mathsf{x})\,\mathsf{w}(\mathsf{x})\mathsf{v}(\mathsf{x})]\mathrm{d}\mathsf{x}$$

and recall from Chapter 9 the definition of inner product on $\mathrm{L}\ (\mathsf{a},\mathsf{b})$:

$$\langle \mathsf{w},\mathsf{v}\rangle = \int_{\mathsf{a}}^{\mathsf{b}} \mathsf{w}(\mathsf{x})\mathsf{v}(\mathsf{x})\mathrm{d}\mathsf{x}\,.$$

Using these, we can rewrite $+(\mathsf{w})$ as follows:

$$+(\mathsf{w}) = -\,(\,(\mathsf{w},\mathsf{w}) \quad \langle \mathsf{f},\mathsf{w}\rangle\,, \quad \mathsf{w} \quad \mathrm{H}_{\&}(\mathsf{a},\mathsf{b})\,. \tag{14.4}$$

The mapping $(\,: \mathrm{H}\ (\mathsf{a},\mathsf{b}) \quad \mathrm{H}\ (\mathsf{a},\mathsf{b})$ is a **bilinear functional** in the following sense:

$$\begin{aligned}
&(\,(\lambda\ \mathsf{w}\ + \lambda\ \mathsf{w}\ ,\mathsf{v}) = \lambda\ \ (\,(\mathsf{w}\ ,\mathsf{v}) + \lambda\ \ (\,(\mathsf{w}\ ,\mathsf{v})\\
&\quad \text{for all } \lambda\ ,\lambda \qquad \text{and all } \mathsf{w}\ ,\mathsf{w}\ ,\mathsf{v} \quad \mathrm{H}\ (\mathsf{a},\mathsf{b});\\
&(\,(\mathsf{w},\mu\ \mathsf{v}\ + \mu\ \mathsf{v}\ ) = \mu\ \ (\,(\mathsf{w},\mathsf{v}\ ) + \mu\ \ (\,(\mathsf{w},\mathsf{v}\ )\\
&\quad \text{for all } \mu\ ,\mu \qquad \text{and all } \mathsf{w},\mathsf{v}\ ,\mathsf{v} \quad \mathrm{H}\ (\mathsf{a},\mathsf{b}).
\end{aligned}$$

We note, in addition, that the bilinear functional $(\,(\ \ ,\ \ )$ is **symmetric**, in that

$$(\,(\mathsf{w},\mathsf{v}) = (\,(\mathsf{v},\mathsf{w}) \qquad \mathsf{w},\mathsf{v} \quad \mathrm{H}\ (\mathsf{a},\mathsf{b})\,. \tag{14.5}$$

Our next result provides an equivalent characterisation of the Rayleigh–Ritz principle; it relies on the fact that the bilinear functional $(\,(\ \ ,\ \ )$ is symmetric in the sense of (14.5).

Theorem 14.1 *A function* u *in* $H_{\&}(a,b)$ *minimises* + ( ) *over* $H_{\&}(a,b)$ *if, and only if,*

$$(G) \qquad ((u,v) = f,v! \qquad v \quad H\ (a,b). \tag{14.6}$$

This identity will be referred to as the **Galerkin principle**.

Proof of theorem Suppose that u $H_{\&}(a,b)$ minimises + ( ) over $H_{\&}(a,b)$; that is, + (u) + (w) for all w $H_{\&}(a,b)$. Noting that w = u + λv belongs to $H_{\&}(a,b)$ for all λ and all v H (a, b), we deduce that

$$\begin{aligned} + (u) \qquad + (u + \lambda v) &= - ((u + \lambda v, u + \lambda v) \qquad f, u + \lambda v! \\ &= \quad + (u) + \lambda[((u,v) \qquad f,v!] + -\lambda \ \ ((v,v) \end{aligned} \tag{14.7}$$

for all v H (a, b) and all λ . Here, in the transition from the first line to the second we made use of the fact that ((u, v) = ((v, u) for all v in H (a, b), which follows from (14.5). Now, (14.7) implies that

$$-\lambda \ \ ((v,v) \quad \lambda[((u,v) \qquad f,v!]$$

for all v H (a, b) and all λ . Let us suppose that λ > 0, divide both sides of the last inequality by λ and pass to the limit λ 0 to deduce that

$$0 \quad ((u,v) \qquad f,v! \qquad v \quad H\ (a,b)\,. \tag{14.8}$$

On replacing v by v in (14.8), we have that also

$$0 \quad ((u,v) \qquad f,v! \qquad v \quad H\ (a,b)\,. \tag{14.9}$$

We conclude from (14.8) and (14.9) that

$$((u,v) = f,v! \qquad v \quad H\ (a,b)\,, \tag{14.10}$$

as required.

Conversely, if u $H_{\&}(a,b)$ is such that ((u, v) = f, v! for all v in H (a, b), then

$$+ (u + \lambda v) = + (u) + \lambda\,[((u,v) \qquad f,v!] + -\lambda \ \ ((v,v) \quad + (u)$$

for all v H (a, b) and all λ ; therefore, u minimises + ( ) over $H_{\&}(a,b)$. □

Thus we have shown that, as long as ((,) is a symmetric bilinear functional, u $H_{\&}(a,b)$ satisfies the Rayleigh–Ritz principle if, and only if, it satisfies the Galerkin principle. Our next task is to explain the

relationship between (RR) and (G) on the one-hand and (14.1), (14.2) on the other. Since in the case of a symmetric bilinear functional $((\ , \ )$ the principles (RR) and (G) are equivalent, it is sufficient to clarify the connection between (G), for example, and the boundary value problem (14.1), (14.2).

We begin with the following definition.

Definition 14.3 *If a function* $u \quad H_{\&}(a,b)$ *satisfies the Galerkin principle (14.6), it is called a* **weak solution** *to the boundary value problem (14.1), (14.2), and the Galerkin principle is referred to as the* **weak formulation** *of the boundary value problem (14.1), (14.2).*

Let us justify this terminology. Suppose that $u \quad H\ (a,b)$ " $H_{\&}(a,b)$ is a solution to the boundary value problem (14.1), (14.2). Then,

$$\frac{d}{dx}\quad p(x)\frac{du}{dx}\quad + r(x)u = f(x)\,, \tag{14.11}$$

for almost every $x \quad (a,b)$ (see the discussion prior to Example 11.1 for a definition of **almost every**). Multiplying this equality by an arbitrary function $v \quad H\ (a,b)$, and integrating over (a,b), we conclude that

$$\int \frac{d}{dx}\quad p(x)\frac{du}{dx}\quad v\,dx + \int\quad r(x)uv\,dx = \int\quad f(x)v(x)\,dx\,.$$

On integration by parts in the first term on the left-hand side,

$$\int \frac{d}{dx}\quad p(x)\frac{du}{dx}\quad v\,dx = \quad p(x)\frac{du}{dx}v\quad + \int\quad p(x)\frac{du}{dx}\frac{dv}{dx}\,dx\,.$$

Since, by hypothesis, $v(a) = 0$ and $v(b) = 0$, it follows that

$$\int p(x)\frac{du}{dx}\frac{dv}{dx}\,dx + \int\quad r(x)uv\,dx = \int\quad f(x)v(x)\,dx$$

for all $v \quad H\ (a,b)$. Thus, we have shown the following result.

Theorem 14.2 *If* $u \quad H\ (a,b)$ " $H_{\&}(a,b)$ *is a solution to the boundary value problem (14.1), (14.2), then* u *is a weak solution to this problem; that is,*

$$(\,(u,v) = \ f,v!\qquad v \quad H\ (a,b)\,. \tag{14.12}$$

The converse implication, namely that any weak solution $u \quad H_{\&}(a,b)$ of (14.1), (14.2) belongs to $H\ (a,b)$ " $H_{\&}(a,b)$ and solves (14.1), (14.2) in the usual (pointwise) sense, is not true in general, unless the weak

the problem by minimising $+(\quad)$ over a finite-dimensional subset $S_{\&}^{*}$ of $H_{\&}(a,b)$, instead.

A simple way of constructing $S_{\&}^{*}$ is to choose any function $\psi \quad H_{\&}(a,b)$, for example,

$$\psi(x) = \frac{B \quad A}{b \quad a}(x \quad a) + A \tag{14.13}$$

and a finite set of linearly independent functions $\phi \ , j = 1, \ldots, n \quad 1$, in $H\ (a,b)$ for $n \quad 2$, and then define

$$S_{\&}^{*} \quad = \quad v^{*} \quad H_{\&}(a,b)\colon v^{*}(x) = \psi(x) + \quad \bar{} \quad v\,\phi\ (x)\,,$$

$$\text{where } (v\ , \ldots, v\ \_\ ) \qquad \bar{} \quad .$$

We consider the following approximation of problem (RR):

$$(\mathrm{RR})^{*} \quad \text{find } u^{*} \quad S_{\&}^{*} \text{ such that } +(u^{*}) = \min_{\cdot\ \ \square 5} \ +(w^{*})\,.$$

Our next result is a finite-dimensional analogue of Theorem 14.1.

Theorem 14.4 *A function* $u^{*} \quad S_{\&}^{*}$ *minimises* $+(\quad)$ *over* $S_{\&}^{*}$ *if, and only if,*

$$(\mathrm{G})^{*} \qquad (\,(u^{*}, v^{*}) = \ f, v^{*}\,! \qquad v^{*} \quad S^{*}\,. \tag{14.14}$$

Here,

$$S^{*} \quad = \quad v^{*} \quad H\ (a,b)\colon v^{*}(x) = \quad \bar{} \quad v\,\phi\ (x)\,,$$

$$\text{where } (v\ , \ldots, v\ \_\ ) \qquad \bar{} \quad .$$

The problem $(\mathrm{G})^{*}$ can be thought of as an approximation to the Galerkin principle (G), and is therefore referred to as the **Galerkin method**. For a similar reason, $(\mathrm{RR})^{*}$ is called the Rayleigh–Ritz method, or just **Ritz method**. Thus, in complete analogy with the equivalence of (RR) and (G) formulated in Theorem 14.1, Theorem 14.4 now expresses the equivalence of $(\mathrm{RR})^{*}$ and $(\mathrm{G})^{*}$, the approximations to (RR) and (G), respectively. Of course, as in the case of (RR) and (G), the equivalence of $(\mathrm{RR})^{*}$ and $(\mathrm{G})^{*}$ relies on the assumption that the bilinear functional $(\,(\quad,\quad)$ is symmetric. The proof is identical to that of Theorem 14.1, and is left as an exercise.

Theorem 14.4 provides no information about the existence and uniqueness of u^{*} that minimises $+(\quad)$ over $S_{\&}^{*}$ (or, equivalently, of the existence

and uniqueness of u^* that satisfies (14.14)). This question is settled by our next result.

Theorem 14.5 *There exists a unique function* $u^* \in S_\Delta^*$ *that minimises* $+(\ )$ *over* S_Δ^**; this* u^* *is called the* **Ritz approximation** *to* u*. Equivalently, there exists a unique function* $u^* \in S_\Delta^*$ *that satisfies (14.14); this* u^* *is called the* **Galerkin approximation** *to* u*. The Ritz and Galerkin approximations to* u *coincide.*

Proof We shall prove the second of these two equivalent statements: we shall show that there exists a unique $u^* \in S_\Delta^*$ that satisfies (14.14). The proof of uniqueness of $u^* \in S_\Delta^*$ is analogous to the proof of Theorem 14.3, with u, $\tilde{u}$, $H_\Delta(a,b)$ and $H(a,b)$, replaced by u^*, $\hat{u}^*$, S_Δ^* and S^*, respectively. Since S_Δ^* is finite-dimensional, the uniqueness of u^* satisfying (14.14) implies its existence. □

Having shown the existence and uniqueness of u^* minimising $+(\ )$ over S_Δ^* (or, equivalently, satisfying (14.14)), we adopt the following definition.

Definition 14.4 *The functions* ϕ_i, $i = 1, 2, \ldots, n-1$, *appearing in the definitions of* S_Δ^* *and* S^* *are called the* **Galerkin basis functions**.

Since any function $v^* \in S^*$ can be represented as a linear combination of the Galerkin basis functions ϕ_i, $1 \le i \le n-1$, it is clear that (14.14) is equivalent to

$$(u^*, \phi_i) = (f, \phi_i), \quad 1 \le i \le n-1. \tag{14.15}$$

As u^* belongs to S_Δ^*, it can be expressed in terms of ψ and the Galerkin basis functions as

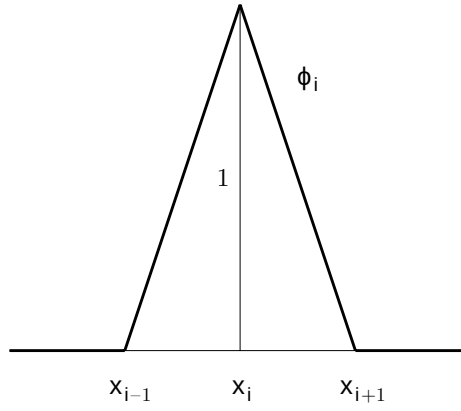
$$u^*(x) = \psi(x) + \sum_{j=1}^{n-1} u_j \phi_j(x),$$

where u_j, $j = 1, \ldots, n-1$, are to be determined. On substituting this expansion of u^* into (14.15), we arrive at the following system of simultaneous linear equations:

$$\sum_{j=1}^{n-1} M_{ij} u_j = b_i, \quad 1 \le i \le n-1, \tag{14.16}$$

where

$$\mathsf{M} = (\phi, \phi), \quad \mathsf{b} =$$



' C , + 4

For the finite element method the important property of the basis functions ϕ , 1 i n 1, is that they have *local* support, being nonzero only in one pair of adjacent intervals, (x – , x] and [x , x). This means that, in the matrix M ,

$$M = 0 \quad \text{if } i \; j > 1 .$$

The matrix M is, therefore, symmetric, positive definite and tridiagonal, and the associated system of linear equations can be solved very efficiently by the methods of Section 3.3, the most efficient algorithm being LU decomposition, without any use of symmetry. The fact that M is positive definite means that no interchanges are necessary.

The function ψ in (14.13), which is included in the definition of $S^*_{\&}$ to ensure that u^* satisfies the boundary conditions at $x = a$ and $x = b$, is then given by

$$\psi(x) = A\phi \ (x) + B\phi \ (x) ,$$

which is also piecewise linear; clearly, $\psi(a) = A$ and $\psi(b) = B$. Here, ϕ and ϕ are defined by setting, respectively, $i = 0$ and $i = n$ in (14.19) and restricting the resulting functions to the interval $[a, b] = [x , x]$. In (14.17) we see that the term ($(\psi, \phi$) is nonzero only for $i = 1$ and $i = n$ 1.

Before attempting to solve the system of linear equations we must, of course, first compute the elements of the matrix M , and the quantities on the right-hand side, b , $i = 1, \ldots, n$ 1; see (14.16) and (14.17). The

matrix elements are obtained from

$$M = ((\phi , \phi)) = \int p(x)\phi'(x)\phi'(x)dx + \int r(x)\phi (x)\phi (x)dx,$$

with $1 \quad i, j \quad n \quad 1$. We have written this as the sum of two terms, as the matrix M is often written in this way as the sum of two matrices which, for historical reasons, are often known as the **stiffness matrix** and the **mass matrix**, respectively. The terms M are very simple; in fact in the first integral the derivatives ϕ' and ϕ' are piecewise constant functions over $[a, b]$.

It may be possible to compute these integrals analytically, but more generally some form of numerical quadrature will be necessary. It is then easy to show that if we use certain types of quadrature formulae we shall be led to the same system of equations as in the finite difference method of Section 13.5. Consider the particularly simple case where the mesh points are equally spaced, so that $x = a + jh$, $j = 0, 1, \ldots, n$, $h = (b \quad a)/n$. If we then approximate the integrals involved in the stiffness matrix by the midpoint rule (see Chapter 10), we obtain

$$\int p(x)\phi'_{-}(x)\phi'(x)dx = (1/h) \int p(x)dx$$

$$p_{-} \quad /h,$$

where $p_{-} = p(x \quad h/2)$, and similarly for the other integrals involved. For the integrals in the mass matrix we use the trapezium rule, and then

$$\int r(x)\phi_{-}(x)\phi (x)dx \quad 0,$$

since ϕ is zero at x_{-} and ϕ_{-} is zero at x. In the same way

$$\int r(x)[\phi (x)]\, dx \quad -hr,$$

where $r = r(x)$, since ϕ is zero at one end of the interval and unity at the other. The other part of the integral is, similarly,

$$\int r(x)[\phi (x)]\, dx \quad -hr. \tag{14.20}$$

Assuming that $f \quad C[a, b]$, approximating the integral on the right-hand side by the trapezium rule in the same way, and putting all the parts together, equation (14.14) now takes the approximate form

$$\frac{p_{-}}{h}u_{-} + \frac{p_{-} + p}{h}u \quad \frac{p}{h}u + hr\, u = hf,$$

for $i = 1, 2, \ldots, n - 1$, with the notational convention that $u_0 = A$ and $u_n = B$, and $f_i = f(x_i)$; clearly, this is the same as the finite difference equation (13.19). Of course, had we used a different set of basis functions ϕ_i, $1 \leq i \leq n - 1$, or different numerical quadrature rules, the finite element and finite difference methods would have no longer been identical. Indeed, this example is just an illustration of the relation between the two methods; we should normally expect to compute the entries of the matrix M by using some more accurate quadrature method, such as a two-point Gauss formula.

In the next two sections we shall assess the accuracy of the finite element method. Our goal is to quantify the amount of reduction in the error $u - u^*$ as the mesh spacing h is reduced.

14.4 Error analysis of the finite element method

We begin with a fundamental result that underlies the error analysis of finite element methods.

Theorem 14.6 (Céa's Lemma) *Suppose that* u *is the function that minimises* $+(u)$ *over* $H_0^1(a, b)$ *(or, equivalently, that* u *satisfies (14.6)), and that* u^* *is its Galerkin approximation obtained by minimising* $+(\ )$ *over* S_0^* *(or, equivalently, that* u^* *satisfies (14.14)). Then,*

$$a(u - u^*, v^*) = 0 \qquad \forall\, v^* \in S_0^*, \tag{14.21}$$

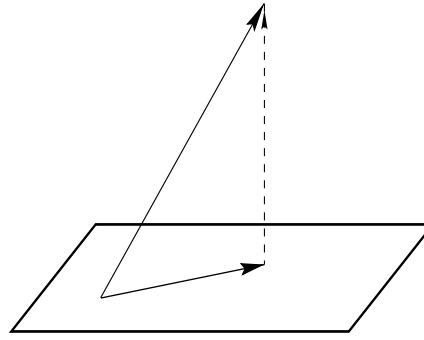
and

$$a(u - u^*, u - u^*) = \min_{v^* \in S_0^*} a(u - v^*, u - v^*). \tag{14.22}$$

The identity (14.21) is referred to as **Galerkin orthogonality**. The terminology stems from the fact that, since the bilinear functional $a(\ ,\ )$ is symmetric and $a(v, v) > 0$ for all $v \in H_0^1(a, b)$, $v \neq 0$, $a(\ ,\ )$ is an inner product in the linear space $H_0^1(a, b)$. Therefore, by virtue of Definition 9.2, (14.21) means that $u - u^*$ is orthogonal to S_0^* in $H_0^1(a, b)$. A geometrical illustration of Galerkin orthogonality is given in Figure 14.2. Given that ψ is a fixed element of $H_0^1(a, b)$, the mapping

$$R^* : u + \psi \in H_0^1(a, b) \to u^* + \psi \in S_\psi^*$$

which assigns a $u^* + \psi \in S_\psi^*$ to $u + \psi \in H_0^1(a, b)$ (where u and u^* are as in Theorem 14.6) is called the **Ritz projector**.



Motivated by the minimisation property (14.22), we define the **energy norm** $\|\cdot\|_A$ on $H_0^1(a,b)$ *via*

$$\|v\|_A = [a(v,v)]^{1/2}. \tag{14.23}$$

Under our hypotheses on p and q, it is easy to see that $\|\cdot\|_A$ satisfies all axioms of norm (see Chapter 2). The result we have just proved shows that u^* is the *best approximation* from S_0^* to the true solution $u \in H_0^1(a,b)$ of our problem, when we measure the error of the approximation in the energy norm:

$$\|u - u^*\|_A = \min_{v \in S_0^*} \|u - v\|_A. \tag{14.24}$$

A particularly relevant question is how the error $u - u^*$ depends on the spacing h of the subdivision of the computational domain $[a,b]$. We can obtain a bound on the error $u - u^*$, measured in the energy norm, by choosing a particular function $v^* \in S_0^*$ in (14.24) whose closeness to u is easy to assess. For this purpose, we introduce the **finite element interpolant** $\pi^* u \in S_0^*$ of $u \in H_0^1(a,b)$ by

$$\pi^* u(x) = \psi(x) + \sum_{j=1}^{n-1} u(x_j)\phi_j(x), \qquad x \in [a,b].$$

Clearly,

$$\pi^* u(x_j) = u(x_j), \qquad j = 0, 1, \ldots, n,$$

which justifies our use of the word *interpolant*.

We then deduce from (14.24) that

$$\|u - u^*\|_A \le \|u - \pi^* u\|_A; \tag{14.25}$$

hence, in order to quantify $\|u - u^*\|_A$, we only need to estimate the size of $\|u - \pi^* u\|_A$. This leads us to the next theorem.

Theorem 14.7 *Suppose that* $u \in H_0^1(a,b) \cap H^2(a,b)$ *and let* $\pi^* u$ *be the finite element interpolant of* u *from* S_0^* *defined above; then, the following error bounds hold:*

$$\|u - \pi^* u\|_\infty \le \frac{h}{\pi} \|u''\|_\infty,$$

$$\|u' - (\pi^* u)'\|_\infty \le \frac{h}{\pi} \|u''\|_\infty,$$

for $i = 1, 2, \ldots, n$, *where* $h = x_i - x_{i-1}$.

Proof Consider an element $[x_{i-1}, x_i]$, $1 \le i \le n$, and define $\zeta(x) = u(x) - u^*(x)$ for $x \in [x_{i-1}, x_i]$. Then, $\zeta \in H^1(x_{i-1}, x_i)$ and $\zeta(x_{i-1}) = \zeta(x_i) = 0$. Therefore ζ can be expanded into a convergent Fourier sine-series,

$$\zeta(x) = \sum_{k=1}^{\infty} a_k$$

Now, substituting the bounds from Theorem 14.7 into the definition of the norm $\|u - u^*\|_A$, we arrive at the following estimate of the interpolation error in the energy norm.

Corollary 14.1 *Suppose that* $u \in H^2(a,b) \cap H_0^1(a,b)$. *Then,*

$$\|u - u^*\|_A \leq \frac{h}{\pi} P + \frac{h}{\pi} R \|u''\|_2,$$

where $P = \max_{[a,b]} p(x)$ *and* $R = \max_{[a,b]} r(x)$.

Proof Let us observe that

$$\begin{aligned} \|v\|_A^2 &= a(v,v) \\ &= \int_a^b \left(p(x)\, v'^2(x) + r(x)\, v^2(x) \right) dx \\ &\leq \int_a^b \left(p(x)\, v'^2(x) + r(x)\, v^2(x) \right) dx \\ &\leq P \|v'\|_2^2 + R \|v\|_2^2. \end{aligned}$$

On letting $v = u - u^*$ and applying the preceding theorem on the right-hand side of the last inequality, with v' and v replaced by $u' - (u^*)'$ and $u - u^*$, respectively, the result follows. □

Inserting this estimate into (14.25) leads to the desired bound on the error between the analytical solution u and its finite element approximation u^* in the energy norm.

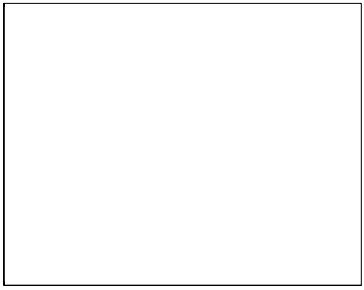
Corollary 14.2 *Suppose that* $u \in H^2(a,b) \cap H_0^1(a,b)$. *Then,*

$$\|u - u^*\|_A \leq \frac{h}{\pi} P + \frac{h}{\pi} R \|u''\|_2,$$

where $P = \max_{[a,b]} p(x)$ *and* $R = \max_{[a,b]} r(x)$. *Further,*

$$\|u - u^*\|_A \leq \frac{h}{\pi} P + \frac{h}{\pi} R \|u''\|_2, \tag{14.26}$$

where $P = \max_{[a,b]} p(x)$, $R = \max_{[a,b]} r(x)$, *and* $h = \max_{1 \leq i \leq n} h_i$.



We see from Figure 14.3 that, as the spacing h of the subdivision is reduced, the finite element solution u^* approximates the analytical solution $u(x) = \sin(\pi x)$ with increasing accuracy. Indeed, the results corresponding to $n = 2$ and $n = 4$ in Figure 14.3 indicate that as the number of intervals in the subdivision is doubled (*i.e.*, h is halved), the maximum error between $u(x)$ and $u^*(x)$ is reduced by a factor of about 4. This reduction in the error cannot be explained by Corollary 14.2 which merely implies that halving h should lead to a reduction in $\|u - u^*\|_A$ by a factor no less than 2. If you would like to learn more about the source of the observed enhancement of accuracy, consult Exercise 5 at the end of the chapter.

14.5 error analysis by duality

The bound on the error between the analytical solution u and its finite element approximation u^* formulated in Corollary 14.2 shows that, in the limit of $h \to 0$, the error $\|u - u^*\|_A$ will tend to zero as $O(h)$. This is a useful result from the theoretical point of view: it reassures us that the unknown analytical solution may be approximated arbitrarily well by making h sufficiently small. On the other hand, asymptotic error bounds of this kind are not particularly helpful for the purpose of precisely quantifying the size of the error between u and u^* for a given, *fixed*, mesh size $h > 0$: as u is unknown, it is difficult to tell just how large the right-hand side of (14.26) really is.

The aim of the present section is, therefore, to derive a computable bound on the error, and to demonstrate how such a bound may be implemented into an adaptive mesh-refinement algorithm, capable of reducing the error $\|u - u^*\|$ below a certain prescribed tolerance in an automated manner, without human intervention. The approach is based on seeking a bound on $\|u - u^*\|$ in terms of the computed solution u^* rather than in terms of norms of the unknown analytical solution u. A bound on the error in terms of u^* is referred to as an
error bound, due to the fact that it becomes *computable* only *after* the numerical solution u^* has been obtained.

In order to illuminate the key ideas while avoiding technical difficulties, we shall consider the two-point boundary value problem

$$(p(x)u')' + q(x)u' + r(x)u = f(x), \qquad a < x < b, \tag{14.28}$$

$$u(a) = A, \quad u(b) = B, \tag{14.29}$$

where $p, q \in C^1[a,b]$, $r \in C[a,b]$ and $f \in L^2(a,b)$. We shall assume, as

at the beginning of the chapter, that $p(x) \quad c > 0$, $x \quad [a,b]$; however, instead of supposing that $r(x) \quad 0$, we shall now demand that

$$r(x) \quad \frac{1}{2}q'(x) \quad c \ , \quad x \quad [a,b] \, , \tag{14.30}$$

where c is assumed to be a positive constant.

Letting

$$((w,v) = \quad [p(x)w$$

consider the auxiliary boundary value problem

$$-(p(x)z')' + q(x)z + r(x)z = (u - u^*)(x), \quad a < x < b, \tag{14.33}$$

$$z(a) = 0, \; z(b) = 0, \tag{14.34}$$

called the **dual problem** (or adjoint problem).

We begin our error analysis by noting that the definition of the dual problem and straightforward integration by parts yield (recalling that $(u - u^*)(a) = 0$, $(u - u^*)(b) = 0$)

$$\begin{aligned} \|u - u^*\|^2 &= (u - u^*, u - u^*) \\ &= (u - u^*, -(pz')' + qz) \\ &= (p(u - u^*)', z'). \end{aligned}$$

On the other hand, (14.31) and (14.32) imply the Galerkin orthogonality property

$$(p(u - u^*), z^*) = 0 \qquad z^* \in S^*.$$

In particular, by choosing

$$z^* = \pi^* z \in S^*,$$

the continuous piecewise linear interpolant of the function $z \in H_0^1(a,b)$, associated with the subdivision $a = x_0 < x_1 < \cdots < x_{N-1} < x_N = b$, we have that

$$(p(u - u^*)', (\pi^* z)') = 0.$$

Thus,

$$\begin{aligned} \|u - u^*\|^2 &= (p(u - u^*)', z' - (\pi^* z)') \\ &= (p(u, z' - (\pi^* z)') - (p(u^*)', z' - (\pi^* z)') \\ &= (f, z - \pi^* z) - (p(u^*)', z' - (\pi^* z)'), \end{aligned} \tag{14.35}$$

where the last transition follows from (14.31) with $v = z - \pi^* z$.

We observe that the right-hand side no longer involves the unknown analytical solution u. Furthermore,

$$\begin{aligned} (p(u^*)', z' - (\pi^* z)') &= \int_0^1 p(x)(u^*)'(x)(z' - (\pi^* z)')(x)\,dx \\ &+ \int_0^1 q(x)(u^*)'(x)(z - \pi^* z)(x)\,dx \\ &+ \int_0^1 r(x)u^*(x)(z - \pi^* z)(x)\,dx. \end{aligned}$$

Integrating by parts in each of the n integrals in the first sum on the right-hand side, noting that $(z_i - z^*)(x_i) = 0$, $i = 0, \ldots, n$, we deduce that

$$\begin{aligned} &(-(u^*)', z_i - z^*)' \\ &= \int_{x_{i-1}}^{x_i} (p(x)(u^*)')' + q(x)(u^*)' + r(x)u^* \,(z_i - z^*)(x)\,dx. \end{aligned}$$

Furthermore,

$$\langle f, z_i - z^* \rangle = \int_{x_{i-1}}^{x_i} f(x)\,(z_i - z^*)(x)\,dx.$$

Substituting these two identities into (14.35), we deduce that

$$a(u - u^*, z_i - z^*) = \int_{x_{i-1}}^{x_i} R(u^*)(x)\,(z_i - z^*)(x)\,dx, \tag{14.36}$$

where, for $1 \le i \le n$, and $x \in (x_{i-1}, x_i)$,

$$R(u^*)(x) = f(x) - (p(x)(u^*)')' + q(x)(u^*)' + r(x)u^*.$$

The function $R(u^*)$ is called **the finite element residual**; it measures the extent to which u^* fails to satisfy the differential equation

$$-(p(x)u')' + q(x)u' + r(x)u = f(x)$$

on the union of the intervals (x_{i-1}, x_i), $i = 1, \ldots, n$. Now, applying the Cauchy–Schwarz inequality on the right-hand side of (14.36) yields

$$a(u - u^*, z_i - z^*) \le \|R(u^*)\|_{L^2} \|z_i - z^*\|_{L^2}.$$

Recalling from Theorem 14.7 that

$$\|z_i - z^*\|_{L^2} \le \frac{h}{\pi} \|z''\|_{L^2}, \quad i = 1, 2, \ldots, n,$$

we deduce that

$$a(u - u^*, z_i - z^*) \le \frac{1}{\pi} h \|R(u^*)\|_{L^2} \|z''\|_{L^2},$$

and consequently, using the Cauchy–Schwarz inequality for finite sums,

$$\sum_{i=1}^{n} a_i b_i \le \left(\sum_{i=1}^{n} a_i \right) \max_{1 \le i \le n} b_i$$

with

a

Integrating by parts, again, in the second term on the right gives

$$(pz')' \quad (qz)' + rz, z! \qquad c \quad z' \ _{\#} \qquad \frac{1}{2} \qquad q'(x)[z \ (x)]dx$$
$$! \qquad + \qquad r(x)[z(x)] \ dx .$$

Hence, from (14.39),

$$c \quad z' \ _{\#} \qquad + \qquad r(x) \quad \frac{1}{2}q'(x) \quad [z(x)] \ dx \qquad u \quad u^*, z! ,$$

and thereby, noting (14.30) and using the Cauchy–Schwarz inequality on the right-hand side,

$$\min \ c \ , c \qquad z' \ _{\#} \qquad + \ z \ _{\#} \qquad u \quad u^*, z!$$
$$u \quad u^* \ _{\#} \qquad z \ _{\#} \qquad . \quad (14.40)$$

Therefore, also

$$\min \ c \ , c \quad z \ _{\$} \qquad u \quad u^* \ _{\#} \qquad z \ _{\$} \qquad ,$$

which means that

$$z' \ _{\#} \qquad + \ z \ _{\#} \qquad = \quad z \ _{\$}$$
$$\frac{1}{\min \ c \ , c} \ u \quad u^* \ _{\#} \qquad . \quad (14.41)$$

Now we substitute (14.41) into (14.38) to deduce that

$$z'' \ _{\#} \qquad K \ u \quad u^* \ _{\#} \qquad , \qquad (14.42)$$

where

$$K = \frac{1}{c} \quad 1 + \frac{1}{\min \ c \ , c} \quad p' + q \ _{\infty} + \ r \quad q' \ _{\infty} \qquad .$$

□

It is important to observe here that K involves only known quantities: the coefficients in the differential equation under consideration. Therefore K can be computed, or at least bounded above, without difficulties. On inserting (14.42) into (14.37), we arrive at our final result, the computable *a posteriori* error bound,

$$u \quad u^* \ _{\#} \qquad K \qquad h \quad R(u^*) \ _{\#} \qquad , \qquad (14.43)$$

where $\mathsf{K} = \mathsf{K}/\pi$.

Next we shall describe the construction of an adaptive mesh refinement algorithm based on the *a posteriori* error bound (14.43).

Suppose that 012

(4 If not, then halve those elements $[x_{-}, x]$ in $-$, with i in the set $1, 2, \ldots, n$, for which

$$h \quad R(u^*)_{\#} > \frac{1}{n} \frac{012}{K}, \tag{14.48}$$

denote by $-$ the resulting subdivision of $[a, b]$ with n elements $[x_{-}, x]$ of respective lengths

$$h = x \quad x_{-}, \quad i = 1, \ldots, n,$$

and consider the associated finite element space $S^*_{\&}$ of dimension n 1;

(! Compute the finite element approximation u^* $S^*_{\&}$, increase m by 1 and return to (,

The inequality (14.47) is called the **stopping criterion** for the mesh adaptation algorithm, and (14.48) is referred to as the **refinement criterion**. According to the *a posteriori* error bound (14.43), when the adaptive algorithm terminates, the error $u \quad u^*_{\#}$ is guaranteed not to exceed the prescribed tolerance 012.

We conclude the body of this chapter with a numerical experiment which illustrates the performance of the adaptive algorithm.

Example 14.1 *Let us consider the second-order ordinary differential equation*

$$(p(x)u')' + q(x)u' + r(x)u = f(x), \quad x \quad (0, 1), \tag{14.49}$$

subject to the boundary conditions

$$u(0) = 0, \quad u(1) = 0. \tag{14.50}$$

Suppose, for example, that

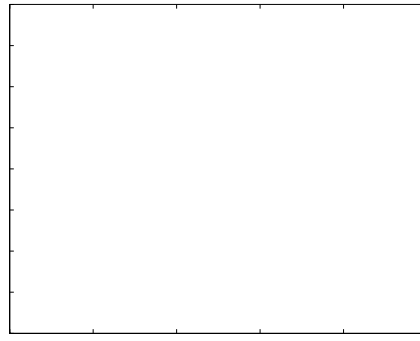
$$p(x) \quad 1, \quad q(x) \quad 20, \quad r(x) \quad 10 \quad \textit{and} \quad f(x) \quad 1.$$

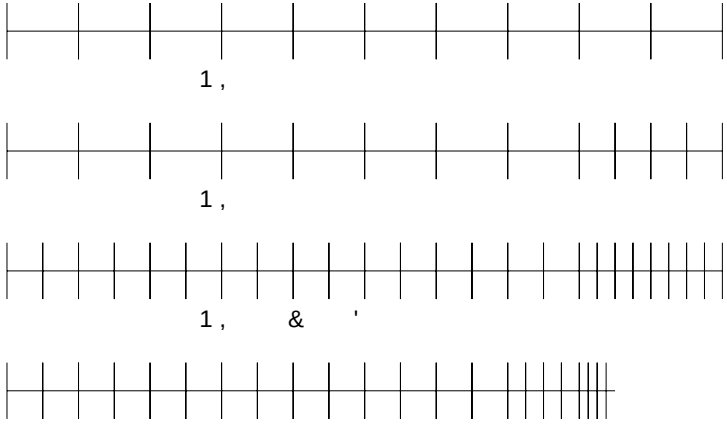
In this case, the analytical solution, u, can be expressed in closed form:

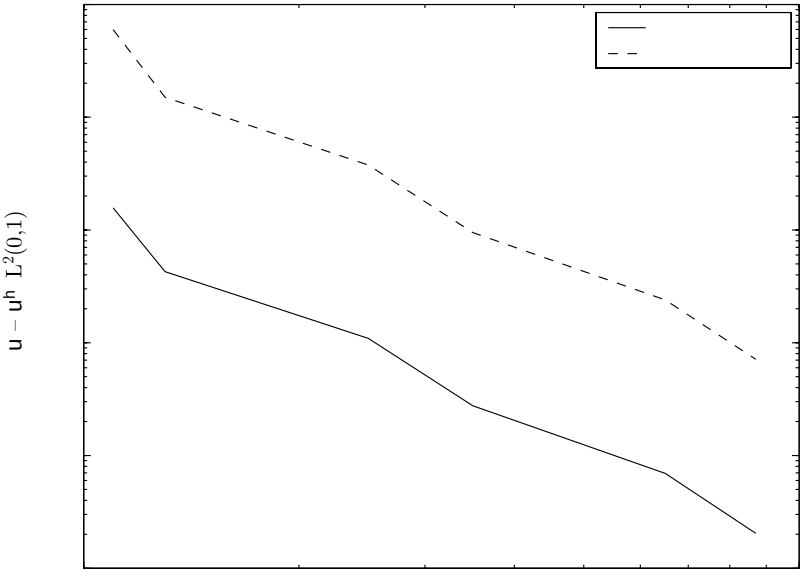
$$u(x) = C \ e^{\&} + C \ e^{\&} + \frac{1}{10},$$

where λ and λ are the two roots of the characteristic polynomial of the differential equation, $\lambda + 20\lambda + 10 = 0$, *i.e.*,

$$\lambda = 10 + \overline{110}, \quad \lambda = 10 \quad \overline{110},$$




1 ,
1 ,
1 , & '



'% 8 L (0 1) * +
,B * +
#%&# 4'

, / - " / " $ " , Introduction to adaptive methods for differential equations, in *Acta Numerica* **4** (A. Iserles, ed.), Cambridge University Press, Cambridge, 105–158, 1995.

% ! - % % , An optimal control approach to a-posteriori error estimation in finite element methods, in *Acta Numerica* **10** (A. Iserles, ed.), Cambridge University Press, Cambridge, 1–102, 2001.

+ ! / &5 , Adjoint methods for PDEs: superconvergence and adaptivity by duality, in *Acta Numerica* **11** (A. Iserles, ed.), Cambridge University Press, Cambridge, 145–236, 2002.

A detailed and general survey of the subject of *a posteriori* error estimation can be found in

+) (' , *A posteriori Error Estimation in Finite Element Analysis*, John Wiley & Sons, New York, 2000.

In this chapter we were concerned with the *a priori* error analysis of the piecewise linear finite element method in the energy norm, and its *a posteriori* error analysis in the L norm. Using similar techniques, one can establish an *a priori* error bound in the L norm and an *a posteriori* error bound in the energy norm. For extensions of the theory considered here to higher-order piecewise polynomial finite element approximations and generalisations to partial differential equations, the reader is referred to the books listed above.

Exercises

14.1 Given that (a, b) is an open interval of the real line, let

$$H_0^1(a,b) = \{ v : v \in H^1(a,b) : v(a) = 0 \}.$$

(i) By writing

$$v(x) = \int_a^x v'(\xi)\,d\xi,$$

for $v \in H_0^1(a,b)$ and $x \in [a,b]$, show the following (**Poincaré–Friedrichs**) inequality:

$$\|v\|_0 \leq \frac{1}{2}(b-a)\,\|v'\|_0 \quad \forall\, v \in H_0^1(a,b).$$

(ii) By writing

$$[v(x)]^2 = \int_a^x \frac{d}{d\xi}[v(\xi)]^2\,d\xi = 2\int_a^x v(\xi)v'(\xi)\,d\xi$$

for $v \in H_0^1(a,b)$ and $x \in [a,b]$, show the following (**Agmon's**) inequality:

$$\max_{[a,b]} |v(x)|^2 \leq 2\,\|v\|_0\,\|v'\|_0 \quad \forall\, v \in H_0^1(a,b).$$

14.2 Given that $f \in L^2(0,1)$, state the weak formulation of each of the following boundary value problems on the interval $(0,1)$:

(a) $-u'' + u = f(x)$, $u(0) = 0$, $u(1) = 0$;
(b) $-u'' + u = f(x)$, $u(0) = 0$, $u'(1) = 1$;
(c) $-u'' + u = f(x)$, $u(0) = 0$, $u(1) + u'(1) = 2$.

In each case, show that there exists at most one weak solution.

14.3 Give a proof of Theorem 14.4.

14.4 Prove Corollary 14.2.

14.5 Consider the boundary value problem

$$p\,u'' + r\,u = f(x), \quad u(0) = 0, \quad u(1) = 0,$$

on the interval $[0, 1]$, where p and r are positive constants and $f \in C[0, 1]$. Using equally spaced points

$$x_i = ih, \qquad i = 0, 1, \ldots, n, \quad \text{with } h = 1/n, \ n \geq 2,$$

and the standard piecewise linear finite element basis functions (hat functions) ϕ_i, $i = 1, 2, \ldots, n-1$, show that the finite element equations for $u_i = u^*(x_i)$ become

$$p\,(u_{i-1} - 2u_i + u_{i+1})/h + r\,(u_{i-1} + 4u_i + u_{i+1})/6 = \frac{1}{h}\,\langle f, \phi_i \rangle$$

for $i = 1, 2, \ldots, n-1$, with $u_0 = 0$ and $u_n = 0$. By expanding in Taylor series, show that

$$\frac{1}{h}\,\langle f, \phi_i \rangle = f(x_i) + \tfrac{1}{6}h^2 f''(x_i) + \cdots.$$

Interpreting this set of difference equations as a finite difference approximation to the boundary value problem, as in Chapter 13, show that the corresponding truncation error T_i satisfies

$$T_i = -\tfrac{1}{6}h^2 r\,u''(x_i) + \cdots, \qquad i = 1, \ldots, n-1,$$

and use the method of Exercise 13.2 to show that

$$\max_{0 \leq i \leq n} |u(x_i) - u^*(x_i)| \leq Mh^2,$$

where M is a positive constant.

14.6 In the notation of Exercise 5 suppose that all the integrals involved in the calculation are approximated by the trapezium rule. Show that the system of equations becomes identical to that obtained from the central difference approximation in Chapter 13, and deduce that

$$\max_{0 \leq i \leq n} |u(x_i) - u^*(x_i)| \leq Mh^2,$$

where M is a positive constant.

14.7 Consider the differential equation

$$(p(x)u')' + r(x)u = f(x), \qquad a < x < b,$$

with p, r and f as at the beginning of the chapter, subject to the boundary conditions

$$p(a)u'(a) + \alpha u(a) = A, \qquad p(b)u'(b) + \beta u(b) = B,$$

where α and β are positive real numbers, and A and B are real numbers. Show that the weak formulation of the boundary value problem is

$$\text{find } u \quad H(a,b) \text{ such that } (u,v) = (v) \text{ for all } v \quad H(a,b),$$

where

$$(u,v) = \int_a^b [p(x)u'(x)v'(x) + r(x)u(x)v(x)]dx + \alpha u(a)v(a) + \beta u(b)v(b),$$

and

$$(v) = f,v + Av(a) + Bv(b).$$

Construct a finite element approximation of the boundary value problem based on this weak formulation using piecewise linear finite element basis functions on the subdivision

$$a = x_0 < x_1 < \cdots < x_{n-1} < x_n = b$$

of the interval $[a,b]$. Show that the finite element method gives rise to a set of $n+1$ simultaneous linear equations with $n+1$ unknowns $u_i = u^*(x_i)$, $i = 0,1,\ldots,n$. Show that this linear system has a unique solution.

Comment on the structure of the matrix M $\times$ of the linear system: (a) Is M symmetric? (b) Is M positive definite? (c) Is M tridiagonal?

14.8 Given that α is a nonnegative real number, consider the differential equation

$$u'' + u = f(x) \quad \text{for } x \quad (0,1),$$

subject to the boundary conditions

$$u(0) = 0, \qquad \alpha u(1) + u'(1) = 0.$$

State the weak formulation of the problem. Using continuous piecewise linear basis functions on a uniform subdivision of $[0,1]$ into elements of size $h = 1/n$, $n \quad 2$, write down the finite element approximation to this problem and show that this has a unique solution u^*. Expand u^* in terms of the standard piecewise linear finite element basis functions (hat functions) ϕ,

$i = 1, 2, \ldots, n$, by writing

$$u^*(x) = \quad U\,\phi\,(x)$$

to obtain a system of linear equations for the vector of unknowns $(U\ , \ldots, U\ )$.

Suppose that $\alpha = 0$, $f(x) \quad 1$ and $h = 1/3$. Solve the resulting system of linear equations and compare the corresponding numerical solution $u^*(x)$ with the exact solution $u(x)$ of the boundary value problem.

14.9 Consider the differential equation

$$(p(x)u')' + r(x)u = f(x), \quad x \quad (0,1),$$

subject to the boundary conditions $u(0) = 0$, $u(1) = 0$, where $p(x) \quad c > 0$, $r(x) \quad 0$ for all x in the closed interval $[0,1]$, with $p \quad C\ [0,1]$, $r \quad C[0,1]$ and $f \quad L\ (0,1)$. Given that u^* denotes the continuous piecewise linear finite element approximation to u on a uniform subdivision of $[0,1]$ into elements of size $h = 1/n$, $n \quad 2$, show that

$$u \quad u^*\ _{\$} \qquad C\ h\ u''\ _{\#} \qquad ,$$

where C is a positive constant that you should specify. Show further that there exists a positive constant C such that

$$u \quad u^*\ _{\$} \qquad Ch\ f\ _{\#} \qquad .$$

Calculate the right-hand sides in these inequalities in the case when

$$p(x) \quad 1\,, \qquad r(x) \quad 0\,, \qquad f(x) \quad 1\,,$$

for $x \quad [0,1]$, and $h = 10^{-}$.

14.10 Consider the two-point boundary value problem

$$u'' + u = f(x)\,, \quad x \quad (0,1)\,, \qquad u(0) = 0\,,\ u(1) = 0\,,$$

with $f \quad C\ [0,1]$. State the piecewise linear finite element approximation to this problem on a nonuniform subdivision

$$0 = x\ < x\ < \qquad < x\ = 1\,, \qquad n \quad 2\,,$$

with $h\ = x \quad x\ _{-}$, assuming that, for a continuous piecewise

linear function v^*,

$$\int f(x)v^*(x)dx$$

has been approximated by applying the trapezium rule on each element $[x_{i-1}, x_i]$.

Verify that the following *a posteriori* bound holds for the error between u and its finite element approximation u^*:

$$\|u - u^*\|_{\infty} \leq K_0 h^2 \|R(u^*)\|_{\infty}$$

$$+ K_1 \max_{x_0 \leq x \leq x_n} h^2 \left(\max_{x_0 \leq x \leq x_n} |f''(x)| + 4 \max_{x_0 \leq x \leq x_n} |f'(x)| \right),$$

where $R(u^*) = f(x) - ((u^*)''(x) + u^*(x))$ for $x \in (x_{i-1}, x_i)$, $i = 1, \ldots, n$, and K_0, K_1 are constants which you should specify.

How would you use this bound to compute u to within a specified tolerance ε?

Appendix A

An overview of results from real analysis

In this Appendix we gather a number of results from real analysis which are assumed at various places in the text. Some of these will be familiar from any course on the subject, and no proofs are given; a small number may be less familiar, and we give proofs of these for completeness.

Theorem A.1 (The Intermediate Value Theorem) *Suppose that* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$ *of . Then,* f *is a bounded function on the interval* $[a, b]$ *and, if* y *is any number such that*

$$\inf_{\square} \quad f(x) \qquad y \qquad \sup_{\square} \quad f(x),$$

then there is a number $\xi \quad [a, b]$ *such that* $f(\xi) = y$. *In particular, the infimum and the supremum of* f *are achieved, and can be replaced by* $\min_{\square}$ *and* $\max_{\square}$ *, respectively.*

The next result, known as Rolle's Theorem, was published in an obscure book in 1691 by the French mathematician Michel Rolle (1652–1719) who invented the notation $\sqrt[n]{x}$ for the nth root of x.

Theorem A. 2 (Rolle's Theorem) *Suppose that* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$ *of , differentiable in the open interval* (a, b), *and such that* $f(a) = f(b)$. *Then, there exists a number* $\xi \quad (a, b)$ *such that* $f'(\xi) = 0$.

It is often important in our applications that the point $\xi \quad (a, b)$, *i.e.*, $a < \xi < b$. For instance it may happen that $f'(a) = f'(b) = 0$, as well as $f(a) = f(b)$; Theorem A.2 then states that, in addition to the endpoints

of the interval $[a, b]$, there is also an interior point $\xi \quad (a, b)$ at which the derivative vanishes.

Theorem A.3 (The Mean Value Theorem) *Suppose that* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$ *of , and* f *is differentiable in the open interval* (a, b)*. Then, there exists a number* $\xi \quad (a, b)$ *such that*

$$f(b) \quad f(a) = f'(\xi)(b \quad a).$$

Theorem A.4 (Taylor's Theorem) *Suppose that* n *is a nonnegative integer, and* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$ *of , such that the derivatives of* f *of order up to and including* n *are defined and continuous on the closed interval* $[a, b]$*. Suppose further that* f *is differentiable on the open interval* (a, b)*. Then, for each value of* x *in* $[a, b]$*, there exists a number* $\xi = \xi(x)$ *in the open interval* (a, b) *such that*

$$\begin{aligned} f(x) &= f(a) + (x \quad a)f'(a) + \quad + \frac{(x \quad a)}{n!} f \quad (a) \\ &\quad + \frac{(x \quad a)}{(n+1)!} f \quad (\xi). \end{aligned}$$

Theorem A.5 (Taylor's Theorem with integral remainder) *Let* n *be a nonnegative integer and suppose that* f *is a real-valued function, defined and continuous on the closed interval* $[a, b]$ *of , such that the derivatives of* f *of order up to and including* n *are defined and continuous on* $[a, b]$*,* f *is differentiable on the open interval* (a, b)*, and* f *is integrable on* (a, b)*. Then, for each* $x \quad [a, b]$*,*

$$\begin{aligned} f(x) &= f(a) + (x \quad a)f'(a) + \quad + \frac{(x \quad a)}{n!} f \quad (a) \\ &\quad + \quad \frac{(x \quad t)}{n!} f \quad (t)dt. \end{aligned}$$

Proof As this version of the theorem may be rather less familiar we include a proof.

The theorem is trivially true for $n = 0$. Suppose that the theorem is true for some nonnegative integer, say $n = k$. Then, provided that f is differentiable on (a, b) and f is integrable on (a, b), integration

by parts shows that

$$\int \frac{(x-t)^{k}}{(k+1)!} f^{(k+1)}(t)\,dt = \frac{(x-a)^{k+1}}{(k+1)!} f^{(k+1)}(a) + \int \frac{(x-t)^{k+1}}{k!} f^{(k+1)}(t)\,dt;$$

use of the theorem when $n = k$ now shows that it is also true for $n = k+1$. The proof by induction is then complete. □

Theorem A.6 (The Integral Mean Value Theorem) *Suppose that* f *is a real-valued function, defined and continuous on a closed interval* $[a,b]$ *of , and let* g *be a function, defined, nonnegative and integrable on* (a,b). *Then, there exists a number* $\xi \in (a,b)$ *such that*

$$\int f(x)g(x)\,dx = f(\xi) \int g(x)\,dx.$$

Proof Since f is continuous on $[a,b]$, it is bounded on $[a,b]$, say

$$m \le f(x) \le M, \quad x \in [a,b].$$

Then, as $g(x) \ge 0$ for all $x \in (a,b)$, we have that

$$mg(x) \le f(x)g(x) \le Mg(x), \quad x \in (a,b).$$

Integrating these inequalities gives

$$m \int g(x)\,dx \le \int f(x)g(x)\,dx \le M \int g(x)\,dx.$$

If $\int g(x)dx = 0$, then the result trivially follows. If, on the other hand, $\int g(x)dx > 0$, then

$$m \le \frac{\int f(x)g(x)\,dx}{\int g(x)\,dx} \le M.$$

The existence of the required value of $\xi \in (a,b)$ now follows from the Intermediate Value Theorem. □

Theorem A.6 obviously also holds provided that $g(x) \le 0$ on (a,b); it is only important that g has constant sign on (a,b). Note also that we do not require that g is continuous, only that it is integrable. For example, Theorem A.6 will hold if f is a continuous function defined on $[0,1]$ and $g(x) = x^{-\alpha}$, $x \in (0,1)$.

Theorem A.7 (Taylor's Theorem for several variables) *Suppose that* f *is a real-valued function of* n *real variables,* $n \quad 1$*, such that* f *and all of its partial derivatives up to and including order* $k+1$ *are defined, continuous and bounded in a neighbourhood of the point in . Let* A *denote an upper bound on the absolute values of all the derivatives of order* $k+1$ *in this neighbourhood. Then*

$$f(\quad + \quad) = f(\quad) + \qquad \frac{U(\quad)}{r!} + E \quad ,$$

where

$$U(\quad) = \quad \eta \frac{\partial}{\partial x} + \quad + \eta \frac{\partial}{\partial x} \quad f \quad (\quad), \qquad r = 1, \ldots, k,$$

and

$$E \qquad \text{——} An \qquad _{\infty} \quad .$$

Proof The proof involves the application of Theorem A.4, Taylor's Theorem, to the function of one variable

$$\phi(t) = f(\quad + t \quad)$$

to give a series expansion for $\phi(1)$. Then, the expressions for the derivatives of ϕ in terms of the partial derivatives of , *via* the chain rule, yield the required result; n is the number of partial derivatives of order $k+1$ for a function of n variables. □

Appendix B

WWW-resources

The book would not be complete without some mention of numerical analysis software and software repositories on the World Wide Web.

An excellent source of mathematical software is the Netlib Repository on the website

A detailed classified list of the available mathematical software libraries can be viewed by clicking on the - button on this webpage. It is also possible to search the repository for a specific piece of software.

Another useful resource is the website of the *ACM Transactions on Mathematical Software* (TOMS) at

5

The site maintains a well-organised repository, including a range of freely available packages for both numerical and symbolical computations, as well as a number of helpful links to various software vendors. The latter include the developers of Maple (a software for symbolical and numerical computations, scientific visualisation and programming), the makers of Mathematica (a software system for symbolical, numerical and graphical computations), the Numerical Algorithms Group (NAG), MathWorks, Inc., the developers of Matlab (a technical computing environment for high-performance numerical computation and visualisation), and many others. Most of the numerical experiments included in the book were performed by using either Matlab or Maple.

Concerning the history of mathematics, we refer to the Mac Tutor history of mathematics website at St Andrews University in Scotland:

A more recent site, dedicated specifically to the history of approximation theory, resides on

˜ ./0

Bibliography

C, : F 1 7 C * $) +
/ *E
/ K >+
C , L 9 / - / M L 3 * $%)+
F *C
@ / K >+
C F ? L 5 * +
*L M B N 1 / K >+
C 4 F F A < F F < < E * $$(+
! " # $ % &
8 $ *17CF @ +
C! ? * $$%+ ' (*8 , 4 B @
8 , +
2 * &+)
) * $ + ,
- ' *) * ,
($.+)+ *8 , 4 "
B @ 8 , +
2 3 > 8 5 * $% + / ! + +
&)I '
2 > < < < * + C "

I C 7 *8 , 4 B @ 8 , +
2 1 / * $ O &+ EP P Q M P
, , P) + + + / I
2AR> C * $$%+ 0 & " *17CF
@ +
2 B 5 1 < , - * $$ + 0 * *1 3 "
+
2 E * + *8 , 4 B @ 8 , +
2 1 1 3 < * +
1 - *1 / K >+
2 L 8 * $)+ # $ % &

+ 1 * 2/ 0 *L M B N 1
8 +
8 , L "3 * $$$+ * "
*1 / K >+
8 B - M * $%%+ ' - *F "9
/ K >+
8 @ * $ $+ ' 0 * #
*8 , 4 B @ 8 , +
8 < * $'&+ T , G "
, , + + + + I &
8 *)(+ ' 3 4 * *5 &
*8 : Q 8 N 8 @ , Q +
E G * $%&+ C , B ,
')I'&
E @ L < , : @ * $ '+ ' * 1 "
- *C @ ? 3+
E 2 8 * + " < - *1
/ K >+
E L "@ 1 , F * + F /
+) +) I $
E >> J T L * $ '+ 1 * 2/ %
$ % & */ "9 C +
E 1 C * $$%+ ' *C I
M B < FC+
E : @ * $$ + *8 , 4 B @ 8 "
, +
E B @ 1 * $"+ C ! G
B ! A +
+ + I $
- 9 * $ + 6) 8 F "
C *C @ 3 +
- > J - E 9 , @ L 8 * $$(+ 7
H G (I (
C 7 *8 , 4 B @ 8 , +
* $) + # * " *@ @ ?! /
K >+
A 1 L * $$ + 4; 1 C * $ I $&&+
+ + * + + + (" $
8 * $+
+ * @ 7 9 2 9 , +
8 * $+ F !
8 , %&I $% *E R %&+
M * $$%+ ? B C
C 7 *8 , 4 B @ 8 "
, +
M * $$)+ 7 ' *2 > R
2 FC+

M , 9 ? * $$$+)
* " *2 > R 2 +
F 2 1R - * + C A @E- "
B ,B B '(I &% C 7
*8 , 4 B @ 8 , +
9 * $))+ - *
) *1 / K >+
, 9 T 3 8 * $$%+ -) 5 -
*L 9 > 4 B @ 2 +
9 C * $ + E F > > C R
> + + $'I&
9 - /U 1 @ M * $$&+ (* # $ %
& '7 % " 1 - *1 2 +
9 - M * $$ + (* # $ % & "7
% $ % * " *1 2 +
9 C * $$ + 0 * *8 , 4 B @ 8 "
, +
9 @ * $% + $! # $ % &
*L M B N 1 / K >+
9 / L * $$%+ * *17CF
@ +
9 , 2 * $(%+ ' *F I9
/ K >+
9 B L 8 * $()+ 1 ! +) +
&&I '&
9 < C L 8 < * $$ + - *8 , 4 B
@ 8 , +
9 C 1 * $%'+
*2 / K >+
9 C 1 * $) + *
& *F I9 / K >+
7 C * $$%+) $ %
& *8 , 4 B @ 8 , +
L > F C 5 , L * $) + C " B"
G ' .+ + + ('(I(%%
L 8 * $$%+ " $ % &
*8 , 4 B @ 8 , +
J : < * $$%+ * 1 7 0
*2 > R 2 FC+
J 3 T * $(+ B 8
+ $I ($' - < ($ /
2 1 M E8
J 3 T C> @ * $ + 1 - "
*@ @ ?! / K >+
J 9 2 * $$ + " ! "
@ *17CF @ +
J 9 2 * $$ + " ! "

< $% , ,B 2 *E /
K >+
J L * % (+ ((+
L J
, *9 : 9 2 7V (I &&
FR $((+
J C < 4 , , 8 M * $$ +) ' *
*17CF @ +
J B T 7 * $% + -) ' * *F /
K >+
3 , L E * $$ + # $ %
*L M B N 1 8 +
F 8 *)' + 9 - + 7 , > * ,B 5 M
5 < - , +
F , 2 * $))+ 7) $ *M 9 "
1 +
F , 2 * $ &+ *M 9 /
K >+
F L M * $$)+ * $ % ! *@
4 B @ @ /L+
F < * $ %+ - *4 B
8 @ 8 +
/P @ * $)$+ # * " *C F 1 B
@ <7+
? L F < , M 8 * + ' (
& (! < $) *17CF
@ +
@ T * $$)+ 1 B G B
' 1 (+)I
@ 2 * $ + * (" *@ I9 - "
8 H /L+
@ > C * + M ! B .+ -+ +
I%%
@ F L E * $$%+ - *8 , 4 "
B @ 8 , +
@ B L E * $$&+ 20 (" *8 "
@ ?! +
< , : @ * $) +) *F I9
/ K >+
< C < , : @ * $) +)
1 - *F I9 / K >+
< 8 * $ %+ 2) *1 / K >+
< ,, 9 * $((+ C > 1 0 + +
%I $
< , M * $((+ T 7 ! + +
+ & I&%

< M * $)%+ " 5 -
*F I9 / K >+
1 > 3 3 * $ + 7 *L M B N 1
/ K >+
1 J 8 B L 3 C M "8 * $$$+) * +
) 7))
*?! 4 B @ ?! +
1 1 * $ %+ / 0
* * $ 7 $ ")
+ < - J 8 F *1 /
K >+
1 L *)& + % 7 (
: * 1 3 +
1 C F 9 C < * $$$+ $
*8 , 4 B @ 8 , +
1 L 8 * &(+ FP P P G P G
5 5 5 ((5 * 3 4 5
+ + +)&I&
1 L 8 * &%+ 1 PG HP P
.+ + " + %I %
1: W * $($+ # * " *C F 1 B
@ <7+
5 3 9 * $'$+ " 0 $ % & (
M 1 8 3 , < *8 , 4 B /
K >+
5 C F * $' + < " H ! 6 + .+ +
+ +)I&
5 3 / 2 E 777 * $$)+ 0 * *17CF
@ +
T E E * $$$+ ' + 0 0+ +.+
7 $ * 1 (*8 @ ?! +
T < 1 * $% + - ' (*@ I9 -
8 H /L+
M > L 9 * $% + - B !
.+ +) + + I&&
M > L 9 * $ + * * (" *8 @
?! 4 B @ / K >+
K J * $) + 5 - *1 2 +
K 5 * $$(+ 9 / I<
' 1 (+ (& I((

Index