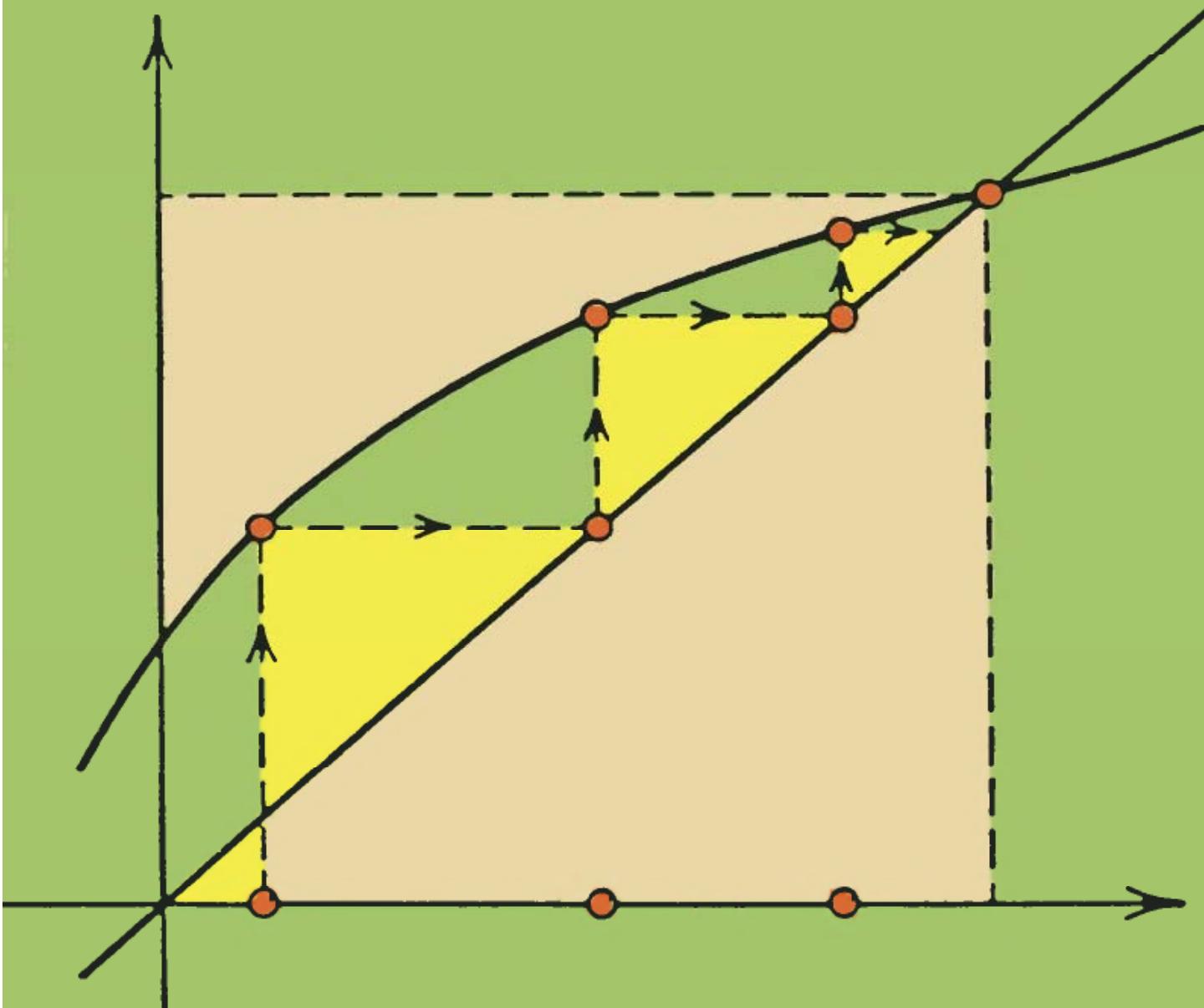


# ANALYSIS OF NUMERICAL METHODS



Eugene Isaacson | Herbert Bishop Keller

# ANALYSIS OF NUMERICAL METHODS

EUGENE ISAACSON

Professor Emeritus of Mathematics  
Courant Institute of Mathematical  
Sciences  
New York University

HERBERT BISHOP KELLER

Professor of Applied Mathematics  
Applied Mathematics  
California Institute of Technology

DOVER PUBLICATIONS, INC., NEW YORK

### *Copyright*

Copyright © 1966 by John Wiley & Sons.  
All rights reserved under Pan American and International Copyright  
Conventions.

Published in Canada by General Publishing Company, Ltd., 30 Lesmill Road,  
Don Mills, Toronto, Ontario

Published in the United Kingdom by Constable and Company, Ltd., 3 The  
Lanchesters, 162–164 Fulham Palace Road, London W6 9ER

### *Bibliographical Note*

This Dover edition, first published in 1994, is an unabridged, corrected republication of the work first published by John Wiley & Sons, New York, 1966. For this edition the authors have corrected a number of errors and provided a new Preface.

### *Library of Congress Cataloging-in-Publication Data*

Isaacson, Eugene.

Analysis of numerical methods / Eugene Isaacson, Herbert Bishop Keller.

p. cm.

Originally published: New York : Wiley, 1966. With new pref.

Includes bibliographical references and index.

ISBN 0-486-68029-0 (pbk.)

1. Numerical analysis. I Keller, Herbert Bishop. II Title

QA297.I8 1994

519.4—dc20

94-7740

CIP

Manufactured in the United States of America  
Dover Publications, Inc., 31 East 2nd Street, Mineola, N.Y. 11501

*To our understanding wives, Muriel and Loretta*



# Preface to the Dover Edition

This edition contains minor corrections to the original edition. In the 28 years that have elapsed between these two editions, there have been great changes in computing equipment and in the development of numerical methods. However, the analysis required to understand and to devise new methods has not changed, and, thus, this somewhat mature text is still relevant. To the list of important topics omitted in the original edition (namely, linear programming, rational approximation and Monte Carlo) we must now add fast transforms, finite elements, wavelets, complexity theory, multigrid methods, adaptive gridding, path following and parallel algorithms. Hopefully, some energetic young numerical analyst will incorporate all these missing topics into an updated version to aid the burgeoning field of scientific computing.

We thank the many people who have pointed out errors and misprints in the original edition. In particular, Mr. Carsten Elsner suggested an elegant improvement in our demonstration of the Runge phenomenon, which we have adopted in Problem 8 on page 280.

EUGENE ISAACSON AND HERBERT B KELLER

*New York and Pasadena  
July 1993*



# Preface to the First Edition

Digital computers, though mass produced for no more than fifteen years, have become indispensable for much current scientific research. One basic reason for this is that by implementing numerical methods, computers form a universal tool for “solving” broad classes of problems. While numerical methods have always been useful it is clear that their role in scientific research is now of fundamental importance. No modern applied mathematician, physical scientist, or engineer can be properly trained without some understanding of numerical methods.

We attempt, in this book, to supply some of the required knowledge. In presenting the material we stress techniques for the development of new methods. This requires knowing why a particular method is effective on some problems but not on others. Hence we are led to the analysis of numerical methods rather than merely their description and listing.

Certainly the solving of scientific problems should not be and is not the sole motivation for studying numerical methods. Our opinion is that the analysis of numerical methods is a broad and challenging mathematical activity whose central theme is the effective constructibility of various kinds of approximations.

Many numerical methods have been neglected in this book since we do not attempt to be exhaustive. Procedures treated are either quite good and efficient by present standards or else their study is considered instructive (while their use may not be advocated). Unfortunately the limitations of space and our own experience have resulted in the omission of many important topics that we would have liked to include (for example, linear programming, rational approximation, Monte Carlo methods).

The present work, it turns out, could be considered a mathematics text in selected areas of analysis and matrix theory. Essentially no

mathematical preparation beyond advanced calculus and elementary linear algebra (or matrix theory) is assumed. Relatively important material on norms in finite-dimensional spaces, not taught in most elementary courses, is included in Chapter 1. Some familiarity with the existence theory for differential equations would be useful, but is not necessary. A cursory knowledge of the classical partial differential equations of mathematical physics would help in Chapter 9. No significant use is made of the theory of functions of a complex variable and our book is elementary in that sense. Deeper studies of numerical methods would also rely heavily on functional analysis, which we avoid here.

The listing of algorithms to concretely describe a method is avoided. Hence some practical experience in using numerical methods is assumed or should be obtained. Examples and problems are given which extend or amplify the analysis in many cases (starred problems are more difficult). It is assumed that the instructor will supplement these with computational problems, according to the availability of computing facilities.

References have been kept minimal and are usually to one of the general texts we have found most useful and compiled into a brief bibliography. Lists of additional, more specialized references are given for the four different areas covered by Chapters 1–4, Chapters 5–7, Chapter 8, and Chapter 9. A few outstanding journal articles have been included here. Complete bibliographies can be found in several of the general texts.

Key equations (and all theorems, problems, and figures) are numbered consecutively by integers within each section. Equations, etc., in other sections are referred to by a decimal notation with explicit mention of the chapter if it is not the current one [that is, equation (3.15) of Chapter 5]. Yielding to customary usage we have not sought historical accuracy in associating names with theorems, methods, etc.

Several different one-semester and two-semester courses have been based on the material in this book. Not all of the subject matter can be covered in the usual one-year course. As examples of some plans that have worked well, we suggest:

*Two-semester courses:*

- (A) Prerequisite—Advanced Calculus and Linear Algebra, Chapters 1–9;
- (B) Prerequisite—Advanced Calculus (with Linear Algebra required only for the second semester), Chapters 3, 5–7, 8 (through Section 3), 1, 2, 4, 8, 9.

*One-semester courses:*

- (A) Chapters 3, 5–7, 8 (through Section 3);
- (B) Chapters 1–5;

(C) Chapters 8, 9 (plus some material from Chapter 2 on iterative methods).

This book benefits from our experience in trying to teach such courses at New York University for over fifteen years and from our students' reactions. Many of our former and present colleagues at the Courant Institute of Mathematical Sciences are responsible for our education in this field. We acknowledge our indebtedness to them, and to the stimulating environment of the Courant Institute. Help was given to us by our friends who have read and used preliminary versions of the text. In this connection we are happy to thank Prof. T. E. Hull, who carefully read our entire manuscript and offered much constructive criticism; Dr. William Morton, who gave valuable suggestions for Chapters 5-7; Professor Gene Golub, who helped us to improve Chapters 1, 2, and 4. We are grateful for the advice given us by Professors H. O. Kreiss, Beresford Parlett, Alan Solomon, Peter Ungar, Richard Varga, and Bernard Levinger, and Dr. Olof Widlund. Thanks are also due to Mr. Julius Rosenthal and Dr. Eva Swenson who helped in the preparation of mimeographed lecture notes for some of our courses. This book grew from two sets of these notes upon the suggestion of Mr. Earle Brach. We are most grateful to Miss Connie Engle who carefully typed our manuscript and to Mr. Richard Swenson who helped in reading galleys. Finally, we must thank Miss Sallyanne Riggione, who as copy editor made many helpful suggestions to improve the book.

*New York and Pasadena  
April, 1966*

E. ISAACSON AND H. B. KELLER



# Contents

**Chapter 1 Norms, Arithmetic, and Well-Posed Computations**

Chapter 2 Numerical Solution of Linear Systems and Matrix Inversion

0.	Introduction . . . . .	26
1.	Gaussian Elimination . . . . .	29
	1.1. Operational Counts . . . . .	34
	1.2. A Priori Error Estimates; Condition Number . . . . .	37
	1.3. A Posteriori Error Estimates . . . . .	46
2.	Variants of Gaussian Elimination . . . . .	50
3.	Direct Factorization Methods . . . . .	52
	3.1. Symmetric Matrices (Cholesky Method) . . . . .	54
	3.2. Tridiagonal or Jacobi Matrices . . . . .	55
	3.3. Block-Tridiagonal Matrices . . . . .	58
4.	Iterative Methods . . . . .	61
	4.1. Jacobi or Simultaneous Iterations . . . . .	64
	4.2. Gauss-Seidel or Successive Iterations . . . . .	66
	4.3. Method of Residual Correction . . . . .	68
	4.4. Positive Definite Systems . . . . .	70
	4.5. Block Iterations . . . . .	72
5.	The Acceleration of Iterative Methods . . . . .	73

5.1. Practical Application of Acceleration Methods . . . . .	78
5.2. Generalizations of the Acceleration Method . . . . .	80
<b>6. Matrix Inversion by Higher Order Iterations . . . . .</b>	<b>82</b>

### **Chapter 3 Iterative Solutions of Non-Linear Equations**

<b>0. Introduction . . . . .</b>	<b>85</b>
<b>1. Functional Iteration for a Single Equation . . . . .</b>	<b>86</b>
1.1. Error Propagation . . . . .	91
1.2. Second and Higher Order Iteration Methods . . . . .	94
<b>2. Some Explicit Iteration Procedures . . . . .</b>	<b>96</b>
2.1. The Simple Iteration or Chord Method (First Order) . . . . .	97
2.2. Newton's Method (Second Order) . . . . .	97
2.3. Method of False Position (Fractional Order) . . . . .	99
2.4. Aitken's $\delta^2$ -Method (Arbitrary Order) . . . . .	102
<b>3. Functional Iteration for a System of Equations . . . . .</b>	<b>109</b>
3.1. Some Explicit Iteration Schemes for Systems . . . . .	113
3.2. Convergence of Newton's Method . . . . .	115
3.3. A Special Acceleration Procedure for Non-Linear Systems	120
<b>4. Special Methods for Polynomials . . . . .</b>	<b>123</b>
4.1. Evaluation of Polynomials and Their Derivatives . . . . .	124
4.2. Sturm Sequences . . . . .	126
4.3. Bernoulli's Method . . . . .	128
4.4. Bairstow's Method . . . . .	131

### **Chapter 4 Computation of Eigenvalues and Eigenvectors**

<b>0. Introduction . . . . .</b>	<b>134</b>
<b>1. Well-Posedness, Error Estimates . . . . .</b>	<b>135</b>
1.1. A Posteriori Error Estimates . . . . .	140
<b>2. The Power Method . . . . .</b>	<b>147</b>
2.1. Acceleration of the Power Method . . . . .	151
2.2. Intermediate Eigenvalues and Eigenvectors (Orthogonalization, Deflation, Inverse Iteration) . . . . .	152
<b>3. Methods Based on Matrix Transformations . . . . .</b>	<b>159</b>

### **Chapter 5 Basic Theory of Polynomial Approximation**

<b>0. Introduction . . . . .</b>	<b>176</b>
<b>1. Weierstrass' Approximation Theorem and Bernstein Polynomials . . . . .</b>	<b>183</b>
<b>2. The Interpolation Polynomials . . . . .</b>	<b>187</b>
2.1. The Pointwise Error in Interpolation Polynomials . . . . .	189

2.2. Hermite or Osculating Interpolation . . . . .	192
<b>3. Least Squares Approximation . . . . .</b>	<b>194</b>
3.1. Construction of Orthonormal Functions . . . . .	199
3.2. Weighted Least Squares Approximation . . . . .	202
3.3. Some Properties of Orthogonal Polynomials . . . . .	203
3.4. Pointwise Convergence of Least Squares Approximations	205
3.5. Discrete Least Squares Approximation . . . . .	211
<b>4. Polynomials of “Best” Approximation . . . . .</b>	<b>221</b>
4.1. The Error in the Best Approximation . . . . .	223
4.2. Chebyshev Polynomials . . . . .	226
<b>5. Trigonometric Approximation . . . . .</b>	<b>229</b>
5.1. Trigonometric Interpolation . . . . .	230
5.2. Least Squares Trigonometric Approximation, Fourier Series . . . . .	237
5.3. “Best” Trigonometric Approximation . . . . .	240

## Chapter 6 Differences, Interpolation Polynomials, and Approximate Differentiation

<b>0. Introduction . . . . .</b>	<b>245</b>
<b>1. Newton’s Interpolation Polynomial and Divided Differences</b>	<b>246</b>
<b>2. Iterative Linear Interpolation . . . . .</b>	<b>258</b>
<b>3. Forward Differences and Equally Spaced Interpolation Points. . . . .</b>	<b>260</b>
3.1. Interpolation Polynomials and Remainder Terms for Equidistant Points . . . . .	264
3.2. Centered Interpolation Formulae . . . . .	270
3.3. Practical Observations on Interpolation . . . . .	273
3.4. Divergence of Sequences of Interpolation Polynomials	275
<b>4. Calculus of Difference Operators . . . . .</b>	<b>281</b>
<b>5. Numerical Differentiation . . . . .</b>	<b>288</b>
5.1. Differentiation Using Equidistant Points . . . . .	292
<b>6. Multivariate Interpolation . . . . .</b>	<b>294</b>

## Chapter 7 Numerical Integration

<b>0. Introduction . . . . .</b>	<b>300</b>
<b>1. Interpolatory Quadrature . . . . .</b>	<b>303</b>
1.1. Newton-Cotes Formulae . . . . .	308
1.2. Determination of the Coefficients . . . . .	315
<b>2. Roundoff Errors and Uniform Coefficient Formulae . . . . .</b>	<b>319</b>
2.1. Determination of Uniform Coefficient Formulae . . . . .	323

<b>3.</b>	<b>Gaussian Quadrature; Maximum Degree of Precision . . . . .</b>	<b>327</b>
<b>4.</b>	<b>Weighted Quadrature Formulae . . . . .</b>	<b>331</b>
<b>4.1.</b>	<b>Gauss-Chebyshev Quadrature . . . . .</b>	<b>334</b>
<b>5.</b>	<b>Composite Quadrature Formulae . . . . .</b>	<b>336</b>
<b>5.1.</b>	<b>Periodic Functions and the Trapezoidal Rule . . . . .</b>	<b>340</b>
<b>5.2.</b>	<b>Convergence for Continuous Functions . . . . .</b>	<b>341</b>
<b>6.</b>	<b>Singular Integrals; Discontinuous Integrands . . . . .</b>	<b>346</b>
<b>6.1.</b>	<b>Finite Jump Discontinuities . . . . .</b>	<b>346</b>
<b>6.2.</b>	<b>Infinite Integrand . . . . .</b>	<b>346</b>
<b>6.3.</b>	<b>Infinite Integration Limits . . . . .</b>	<b>350</b>
<b>7.</b>	<b>Multiple Integrals . . . . .</b>	<b>352</b>
<b>7.1.</b>	<b>The Use of Interpolation Polynomials . . . . .</b>	<b>354</b>
<b>7.2.</b>	<b>Undetermined Coefficients (and Nodes) . . . . .</b>	<b>356</b>
<b>7.3.</b>	<b>Separation of Variables . . . . .</b>	<b>359</b>
<b>7.4.</b>	<b>Composite Formulae for Multiple Integrals . . . . .</b>	<b>361</b>

## Chapter 8 Numerical Solution of Ordinary Differential Equations

<b>0.</b>	<b>Introduction . . . . .</b>	<b>364</b>
<b>1.</b>	<b>Methods Based on Approximating the Derivative: Euler-Cauchy Method . . . . .</b>	<b>367</b>
<b>1.1.</b>	<b>Improving the Accuracy of the Numerical Solution . . . . .</b>	<b>372</b>
<b>1.2.</b>	<b>Roundoff Errors . . . . .</b>	<b>374</b>
<b>1.3.</b>	<b>Centered Difference Method . . . . .</b>	<b>377</b>
<b>1.4.</b>	<b>A Divergent Method with Higher Order Truncation Error . . . . .</b>	<b>379</b>
<b>2.</b>	<b>Multistep Methods Based on Quadrature Formulae . . . . .</b>	<b>384</b>
<b>2.1.</b>	<b>Error Estimates in Predictor-Corrector Methods . . . . .</b>	<b>388</b>
<b>2.2.</b>	<b>Change of Net Spacing . . . . .</b>	<b>393</b>
<b>3.</b>	<b>One-Step Methods . . . . .</b>	<b>395</b>
<b>3.1.</b>	<b>Finite Taylor's Series . . . . .</b>	<b>397</b>
<b>3.2.</b>	<b>One-Step Methods Based on Quadrature Formulae . . . . .</b>	<b>400</b>
<b>4.</b>	<b>Linear Difference Equations . . . . .</b>	<b>405</b>
<b>5.</b>	<b>Consistency, Convergence, and Stability of Difference Methods</b>	<b>410</b>
<b>6.</b>	<b>Higher Order Equations and Systems . . . . .</b>	<b>418</b>
<b>7.</b>	<b>Boundary Value and Eigenvalue Problems . . . . .</b>	<b>421</b>
<b>7.1.</b>	<b>Initial Value or "Shooting" Methods . . . . .</b>	<b>424</b>
<b>7.2.</b>	<b>Finite Difference Methods . . . . .</b>	<b>427</b>
<b>7.3.</b>	<b>Eigenvalue Problems . . . . .</b>	<b>434</b>

## Chapter 9 Difference Methods for Partial Differential Equations

<b>0.</b>	<b>Introduction . . . . .</b>	<b>442</b>
-----------	-------------------------------	------------

0.1. Conventions of Notation . . . . .	444
<b>1. Laplace Equation in a Rectangle . . . . .</b>	<b>445</b>
1.1. Matrix Formulation . . . . .	452
1.2. An Eigenvalue Problem for the Laplacian Operator . . . . .	458
<b>2. Solution of Laplace Difference Equations . . . . .</b>	<b>463</b>
2.1. Line or Block Iterations . . . . .	471
2.2. Alternating Direction Iterations . . . . .	475
<b>3. Wave Equation and an Equivalent System . . . . .</b>	<b>479</b>
3.1. Difference Approximations and Domains of Dependence . . . . .	485
3.2. Convergence of Difference Solutions . . . . .	491
3.3. Difference Methods for a First Order Hyperbolic System . . . . .	495
<b>4. Heat Equation . . . . .</b>	<b>501</b>
4.1. Implicit Methods . . . . .	505
<b>5. General Theory: Consistency, Convergence, and Stability . . . . .</b>	<b>514</b>
5.1. Further Consequences of Stability . . . . .	522
5.2. The von Neumann Stability Test . . . . .	523
Bibliography . . . . .	531
Index . . . . .	535



# 1

## Norms, Arithmetic, and Well-Posed Computations

### 0. INTRODUCTION

In this chapter, we treat three topics that are generally useful for the analysis of the various numerical methods studied throughout the book. In Section 1, we give the elements of the theory of *norms* of finite dimensional vectors and matrices. This subject properly belongs to the field of *linear algebra*. In later chapters, we may occasionally employ the notion of the norm of a function. This is a straightforward extension of the notion of a vector norm to the infinite-dimensional case. On the other hand, we shall not introduce the corresponding natural generalization, i.e., the notion of the norm of a linear transformation that acts on a space of functions. Such ideas are dealt with in *functional analysis*, and might profitably be used in a more sophisticated study of numerical methods.

We study briefly, in Section 2, the practical problem of the effect of rounding errors on the basic operations of arithmetic. Except for calculations involving only exact-integer arithmetic, rounding errors are invariably present in any computation. A most important feature of the later analysis of numerical methods is the incorporation of a treatment of the effects of such rounding errors.

Finally, in Section 3, we describe the computational problems that are “reasonable” in some general sense. In effect, a numerical method which produces a solution insensitive to small changes in data or to rounding errors is said to yield a *well-posed computation*. How to determine the sensitivity of a numerical procedure is dealt with in special cases throughout the book. We indicate heuristically that any *convergent* algorithm is a well-posed computation.

## 1. NORMS OF VECTORS AND MATRICES

We assume that the reader is familiar with the basic theory of linear algebra, not necessarily in its abstract setting, but at least with specific reference to finite-dimensional linear vector spaces over the field of complex scalars. By “basic theory” we of course include: the theory of linear systems of equations, some elementary theory of determinants, and the theory of matrices or linear transformations to about the Jordan normal form. We hardly employ the Jordan form in the present study. In fact a much weaker result can frequently be used in its place (when the divisor theory or invariant subspaces are not actually involved). This result is all too frequently skipped in basic linear algebra courses, so we present it as

**THEOREM 1.** *For any square matrix  $A$  of order  $n$  there exists a non-singular matrix  $P$ , of order  $n$ , such that*

$$B = P^{-1}AP$$

*is upper triangular and has the eigenvalues of  $A$ , say  $\lambda_j \equiv \lambda_j(A)$ ,  $j = 1, 2, \dots, n$ , on the principal diagonal (i.e., any square matrix is equivalent to a triangular matrix).*

*Proof.* We sketch the proof of this result. The reader should have no difficulty in completing the proof in detail.

Let  $\lambda_1$  be an eigenvalue of  $A$  with corresponding eigenvector  $\mathbf{u}_1$ .† Then pick a basis for the  $n$ -dimensional complex vector space,  $C_n$ , with  $\mathbf{u}_1$  as the first such vector. Let the independent basis vectors be the columns of a non-singular matrix  $P_1$ , which then determines the transformation to the new basis. In this new basis the transformation determined by  $A$  is given by  $B_1 \equiv P_1^{-1}AP_1$  and since  $A\mathbf{u}_1 = \lambda_1\mathbf{u}_1$ ,

$$B_1 = P_1^{-1}AP_1 = \begin{pmatrix} \lambda_1 & \alpha_1 & \alpha_2 & \cdots & \alpha_{n-1} \\ 0 & & & & \\ \vdots & & & A_2 & \\ 0 & & & & \end{pmatrix}$$

where  $A_2$  is some matrix of order  $n - 1$ .

The characteristic polynomial of  $B_1$  is clearly

$$\det(\lambda I_n - B_1) = (\lambda - \lambda_1) \det(\lambda I_{n-1} - A_2),$$

† Unless otherwise indicated, boldface type denotes column vectors. For example, an  $n$ -dimensional vector  $\mathbf{u}_k$  has the components  $u_{ik}$ ; i.e.,

$$\mathbf{u}_k \equiv \begin{pmatrix} u_{1k} \\ u_{2k} \\ \vdots \\ u_{nk} \end{pmatrix}.$$

where  $I_n$  is the identity matrix of order  $n$ . Now pick some eigenvalue  $\lambda_2$  of  $A_2$  and corresponding  $(n - 1)$ -dimensional eigenvector,  $\mathbf{v}_2$ ; i.e.,

$$A_2 \mathbf{v}_2 = \lambda_2 \mathbf{v}_2.$$

With this vector we define the independent  $n$ -dimensional vectors

$$\hat{\mathbf{u}}_1 = \begin{pmatrix} 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}, \quad \hat{\mathbf{u}}_2 = \begin{pmatrix} 0 \\ \mathbf{v}_2 \end{pmatrix}.$$

Note that with the scalar  $\alpha = \alpha_1 v_{12} + \alpha_2 v_{22} + \cdots + \alpha_{n-1} v_{n-1,2}$

$$B_1 \hat{\mathbf{u}}_1 = \lambda_1 \hat{\mathbf{u}}_1, \quad B_1 \hat{\mathbf{u}}_2 = \lambda_2 \hat{\mathbf{u}}_2 + \alpha \hat{\mathbf{u}}_1,$$

and thus if we set  $\mathbf{u}_1 = P_1 \hat{\mathbf{u}}_1$ ,  $\mathbf{u}_2 = P_1 \hat{\mathbf{u}}_2$ , then

$$A \mathbf{u}_1 = \lambda_1 \mathbf{u}_1, \quad A \mathbf{u}_2 = \lambda_2 \mathbf{u}_2 + \alpha \mathbf{u}_1.$$

Now we introduce a new basis of  $C_n$  with the first two vectors being  $\mathbf{u}_1$  and  $\mathbf{u}_2$ . The non-singular matrix  $P_2$  which determines this change of basis has  $\mathbf{u}_1$  and  $\mathbf{u}_2$  as its first two columns; and the original linear transformation in the new basis has the representation

$$B_2 = P_2^{-1} A P_2 = \begin{pmatrix} \lambda_1 & x & x & \cdots & x \\ 0 & \lambda_2 & x & \cdots & x \\ 0 & 0 & & & \\ \vdots & \vdots & & & A_3 \\ 0 & 0 & & & \end{pmatrix},$$

where  $A_3$  is some matrix of order  $n - 2$ .

The theorem clearly follows by the above procedure; a formal inductive proof could be given. ■

It is easy to prove the related stronger result of Schur stated in Theorem 2.4 of Chapter 4 (see Problem 2.13(b) of Chapter 4). We turn now to the basic content of this section, which is concerned with the generalization of the concept of distance in  $n$ -dimensional linear vector spaces.

The “distance” between a vector and the null vector, i.e., the origin, is a measure of the “size” or “length” of the vector. This generalized notion of distance or size is called a *norm*. In particular, all such generalizations are required to have the following properties:

- (0) To each vector  $\mathbf{x}$  in the linear space,  $\mathcal{V}$ , say, a unique real number is assigned; this number, denoted by  $\|\mathbf{x}\|$  or  $N(\mathbf{x})$ , is called the norm of  $\mathbf{x}$  iff:

- (i)  $\|\mathbf{x}\| \geq 0$  for all  $\mathbf{x} \in \mathcal{V}$  and  $\|\mathbf{x}\| = 0$  iff  $\mathbf{x} = \mathbf{0}$ ;  
where  $\mathbf{0}$  denotes the *zero vector* (if  $\mathcal{V} \equiv C_n$ , then  $o_i = 0$ );
- (ii)  $\|\alpha\mathbf{x}\| = |\alpha| \cdot \|\mathbf{x}\|$  for all scalars  $\alpha$  and all  $\mathbf{x} \in \mathcal{V}$ ;
- (iii)  $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$ , the *triangle inequality*,<sup>†</sup> for all  $\mathbf{x}, \mathbf{y} \in \mathcal{V}$ .

Some examples of norms in the complex  $n$ -dimensional space  $C_n$  are

$$(1a) \quad \|\mathbf{x}\|_1 \equiv N_1(\mathbf{x}) \equiv \sum_{j=1}^n |x_j|,$$

$$(1b) \quad \|\mathbf{x}\|_2 \equiv N_2(\mathbf{x}) \equiv \left( \sum_{j=1}^n |x_j|^2 \right)^{\frac{1}{2}},$$

$$(1c) \quad \|\mathbf{x}\|_p \equiv N_p(\mathbf{x}) \equiv \left( \sum_{j=1}^n |x_j|^p \right)^{1/p}, \quad p \geq 1,$$

$$(1d) \quad \|\mathbf{x}\|_\infty \equiv N_\infty(\mathbf{x}) \equiv \max_j |x_j|.$$

It is an easy exercise for the reader to justify the use of the notation in (1d) by verifying that

$$\lim_{p \rightarrow \infty} N_p(\mathbf{x}) = N_\infty(\mathbf{x}).$$

The norm,  $\|\cdot\|_2$ , is frequently called the *Euclidean norm* as it is just the formula for distance in ordinary three-dimensional Euclidean space extended to dimension  $n$ . The norm,  $\|\cdot\|_\infty$ , is called the *maximum norm* or occasionally the *uniform norm*. In general,  $\|\cdot\|_p$ , for  $p \geq 1$  is termed the  *$p$ -norm*.

To verify that (1) actually defines norms, we observe that conditions (0), (i), and (ii) are trivially satisfied. Only the triangle inequality, (iii), offers any difficulty. However,

$$\begin{aligned} N_1(\mathbf{x} + \mathbf{y}) &= \sum_{j=1}^n |x_j + y_j| \\ &\leq \sum_{j=1}^n (|x_j| + |y_j|) = \sum_{j=1}^n |x_j| + \sum_{j=1}^n |y_j| \\ &= N_1(\mathbf{x}) + N_1(\mathbf{y}); \end{aligned}$$

<sup>†</sup> For complex numbers  $x$  and  $y$  the elementary inequality  $|x + y| \leq |x| + |y|$  expresses the fact that the length of any side of a triangle is not greater than the sum of the lengths of the other two sides.

and

$$\begin{aligned} N_\infty(\mathbf{x} + \mathbf{y}) &= \max_j |x_j + y_j| \\ &\leq \max_j (|x_j| + |y_j|) \leq \max_j |x_j| + \max_k |y_k| \\ &= N_\infty(\mathbf{x}) + N_\infty(\mathbf{y}), \end{aligned}$$

so (1a) and (1d) define norms.

The proof of (iii) for (1b), the Euclidean norm, is based on the well-known *Cauchy-Schwarz inequality* which states that

$$(2) \quad \left| \sum_{j=1}^n x_j y_j \right| \leq \left( \sum_{j=1}^n |x_j|^2 \right)^{\frac{1}{2}} \left( \sum_{j=1}^n |y_j|^2 \right)^{\frac{1}{2}} = N_2(\mathbf{x}) N_2(\mathbf{y}).$$

To prove this basic result, let  $|\mathbf{x}|$  and  $|\mathbf{y}|$  be the  $n$ -dimensional vectors with components  $|x_j|$  and  $|y_j|$ ,  $j = 1, 2, \dots, n$ , respectively. Then for any real scalar,  $\xi$ ,

$$0 \leq N_2^2(\xi|\mathbf{x}| + |\mathbf{y}|) = \xi^2 \sum_{j=1}^n |x_j|^2 + 2\xi \sum_{j=1}^n |x_j| |y_j| + \sum_{j=1}^n |y_j|^2.$$

But since the real quadratic polynomial in  $\xi$  above does not change sign its discriminant must be non-positive; i.e.,

$$\left( \sum_{j=1}^n |x_j| \cdot |y_j| \right)^2 \leq \left( \sum_{j=1}^n |x_j|^2 \right) \left( \sum_{j=1}^n |y_j|^2 \right).$$

However, we note that

$$\left| \sum_{j=1}^n x_j y_j \right|^2 \leq \left( \sum_{j=1}^n |x_j| |y_j| \right)^2,$$

and (2) follows from the above pair of inequalities.

Now we form

$$\begin{aligned} N_2(\mathbf{x} + \mathbf{y}) &= \left( \sum_{j=1}^n |x_j + y_j|^2 \right)^{\frac{1}{2}} = \left( \sum_{j=1}^n (x_j + y_j)(\bar{x}_j + \bar{y}_j) \right)^{\frac{1}{2}} \\ &= \left( \sum_{j=1}^n |x_j|^2 + \sum_{j=1}^n (x_j \bar{y}_j + \bar{x}_j y_j) + \sum_{j=1}^n |y_j|^2 \right)^{\frac{1}{2}} \\ &\leq \left( N_2^2(\mathbf{x}) + 2 \sum_{j=1}^n |x_j| \cdot |y_j| + N_2^2(\mathbf{y}) \right)^{\frac{1}{2}}. \end{aligned}$$

An application of the Cauchy-Schwarz inequality yields finally

$$N_2(\mathbf{x} + \mathbf{y}) \leq N_2(\mathbf{x}) + N_2(\mathbf{y})$$

and so the Euclidean norm also satisfies the triangle inequality.

The statement that

$$(3) \quad N_p(\mathbf{x} + \mathbf{y}) \leq N_p(\mathbf{x}) + N_p(\mathbf{y}), \quad p \geq 1,$$

is known as *Minkowski's inequality*. We do not derive it here as general  $p$ -norms will not be employed further. (A proof of (3) can be found in most advanced calculus texts.)

We can show quite generally that all vector norms are continuous functions in  $C_n$ . That is,

**LEMMA 1.** *Every vector norm,  $N(\mathbf{x})$ , is a continuous function of  $x_1, x_2, \dots, x_n$ , the components of  $\mathbf{x}$ .*

*Proof.* For any vectors  $\mathbf{x}$  and  $\boldsymbol{\delta}$  we have by (iii)

$$N(\mathbf{x} + \boldsymbol{\delta}) \leq N(\mathbf{x}) + N(\boldsymbol{\delta}),$$

so that

$$N(\mathbf{x} + \boldsymbol{\delta}) - N(\mathbf{x}) \leq N(\boldsymbol{\delta}).$$

On the other hand, by (ii) and (iii),

$$\begin{aligned} N(\mathbf{x}) &= N(\mathbf{x} + \boldsymbol{\delta} - \boldsymbol{\delta}) \\ &\leq N(\mathbf{x} + \boldsymbol{\delta}) + N(\boldsymbol{\delta}), \end{aligned}$$

so that

$$-N(\boldsymbol{\delta}) \leq N(\mathbf{x} + \boldsymbol{\delta}) - N(\mathbf{x}).$$

Thus, in general

$$|N(\mathbf{x} + \boldsymbol{\delta}) - N(\mathbf{x})| \leq N(\boldsymbol{\delta}).$$

With the unit vectors†  $\{\mathbf{e}_k\}$ , any  $\boldsymbol{\delta}$  has the representation

$$\boldsymbol{\delta} = \sum_{k=1}^n \delta_k \mathbf{e}_k.$$

Using (ii) and (iii) repeatedly implies

$$\begin{aligned} (4a) \quad N(\boldsymbol{\delta}) &\leq \sum_{k=1}^n N(\delta_k \mathbf{e}_k) \\ &\leq \sum_{k=1}^n |\delta_k| N(\mathbf{e}_k) \\ &\leq \max_k |\delta_k| \sum_{j=1}^n N(\mathbf{e}_j) \\ &= MN_\infty(\boldsymbol{\delta}), \end{aligned}$$

†  $\mathbf{e}_k$  has the components  $e_{ik}$ , where  $e_{ik} = 0$ ,  $i \neq k$ ;  $e_{kk} = 1$ .

where

$$(4b) \quad M \equiv \sum_{j=1}^n N(\mathbf{e}_j).$$

Using this result in the previous inequality yields, for any  $\epsilon > 0$  and all  $\delta$  with  $N_\infty(\delta) \leq \epsilon/M$ ,

$$|N(\mathbf{x} + \delta) - N(\mathbf{x})| \leq \epsilon.$$

This is essentially the definition of continuity for a function of the  $n$  variables  $x_1, x_2, \dots, x_n$ . ■

See Problem 6 for a mild generalization.

Now we can show that all vector norms are equivalent in the sense of

**THEOREM 2.** *For each pair of vector norms, say  $N(\mathbf{x})$  and  $N'(\mathbf{x})$ , there exist positive constants  $m$  and  $M$  such that for all  $\mathbf{x} \in C_n$ :*

$$mN'(\mathbf{x}) \leq N(\mathbf{x}) \leq MN'(\mathbf{x}).$$

*Proof.* The proof need only be given when one of the norms is  $N_\infty$ , since  $N$  and  $N'$  are equivalent if they are each equivalent to  $N_\infty$ . Let  $S \subset C_n$  be defined by

$$S \equiv \{\mathbf{x} \mid N_\infty(\mathbf{x}) = 1, \mathbf{x} \in C_n\}$$

(this is frequently called the surface of the unit ball in  $C_n$ ).  $S$  is a closed bounded set of points. Then since  $N(\mathbf{x})$  is a continuous function (see Lemma 1), we conclude by a theorem of Weierstrass that  $N(\mathbf{x})$  attains its minimum and its maximum on  $S$  at some points of  $S$ . That is, for some  $\mathbf{x}^0 \in S$  and  $\mathbf{x}^1 \in S$

$$N(\mathbf{x}^0) = \min_{\mathbf{x} \in S} N(\mathbf{x}), \quad N(\mathbf{x}^1) = \max_{\mathbf{x} \in S} N(\mathbf{x})$$

or

$$0 < N(\mathbf{x}^0) \leq N(\mathbf{x}) \leq N(\mathbf{x}^1) < \infty$$

for all  $\mathbf{x} \in S$ .

For any  $\mathbf{y} \neq \mathbf{0}$  we see that  $\mathbf{y}/N_\infty(\mathbf{y})$  is in  $S$  and so

$$N(\mathbf{x}^0) \leq N\left(\frac{\mathbf{y}}{N_\infty(\mathbf{y})}\right) \leq N(\mathbf{x}^1)$$

or

$$N(\mathbf{x}^0)N_\infty(\mathbf{y}) \leq N(\mathbf{y}) \leq N(\mathbf{x}^1)N_\infty(\mathbf{y}).$$

The last two inequalities yield

$$mN_\infty(\mathbf{y}) \leq N(\mathbf{y}) \leq MN_\infty(\mathbf{y}),$$

where  $m \equiv N(\mathbf{x}^0)$  and  $M \equiv N(\mathbf{x}^1)$ . ■

A matrix of order  $n$  could be treated as a vector in a space of dimension  $n^2$  (with some fixed convention as to the manner of listing its elements). Then matrix norms satisfying the conditions (0)–(iii) could be defined as in (1). However, since the product of two matrices of order  $n$  is also such a matrix, we impose an additional condition on matrix norms, namely that

$$(iv) \quad \|AB\| \leq \|A\| \cdot \|B\|.$$

With this requirement the vector norms (1) do not all become matrix norms (see Problem 2). However, there is a more natural, geometric, way in which the norm of a matrix can be defined. Thus, if  $\mathbf{x} \in C_n$  and  $\|\cdot\|$  is some vector norm on  $C_n$ , then  $\|\mathbf{x}\|$  is the “length” of  $\mathbf{x}$ ,  $\|A\mathbf{x}\|$  is the “length” of  $A\mathbf{x}$ , and we define a norm of  $A$ , written as  $\|A\|$  or  $N(A)$ , by the maximum relative “stretching,”

$$(5) \quad \|A\| \equiv \sup_{\mathbf{x} \neq \mathbf{0}} \frac{\|A\mathbf{x}\|}{\|\mathbf{x}\|}.$$

Note that we use the same notation,  $\|\cdot\|$ , to denote vector and matrix norms; the context will always clarify which is implied. We call (5) a *natural norm* or the *matrix norm induced by the vector norm*,  $\|\cdot\|$ . This is also known as the *operator norm* in functional analysis. Since for any  $\mathbf{x} \neq \mathbf{0}$  we can define  $\mathbf{u} = \mathbf{x}/\|\mathbf{x}\|$  so that  $\|\mathbf{u}\| = 1$ , the definition (5) is equivalent to

$$(6) \quad \|A\| = \max_{\|\mathbf{u}\|=1} \|A\mathbf{u}\| = \|A\mathbf{y}\|, \quad \|\mathbf{y}\| = 1.$$

That is, by Problems 6 and 7,  $\|A\mathbf{u}\|$  is a continuous function of  $\mathbf{u}$  and hence the maximum is attained for some  $\mathbf{y}$ , with  $\|\mathbf{y}\| = 1$ .

Before verifying the fact that (5) or (6) defines a matrix norm, we note that they imply, for any vector  $\mathbf{x}$ , that

$$(7) \quad \|A\mathbf{x}\| \leq \|A\| \cdot \|\mathbf{x}\|.$$

There are many other ways in which matrix norms may be defined. But if (7) holds for some such norm then it is said to be *compatible* with the vector norm employed in (7). The natural norm (5) is essentially the “smallest” matrix norm compatible with a given vector norm.

To see that (5) yields a norm, we first note that conditions (i) and (ii) are trivially verified. For checking the triangle inequality, let  $\mathbf{y}$  be such that  $\|\mathbf{y}\| = 1$  and from (6),

$$\|(A + B)\mathbf{y}\| = \|(A + B)\mathbf{y}\|.$$

But then, upon recalling (7),

$$\begin{aligned} \|A + B\| &\leq \|A\mathbf{y}\| + \|B\mathbf{y}\| \\ &\leq \|A\| + \|B\|. \end{aligned}$$

Finally, to verify (iv), let  $\mathbf{y}$  with  $\|\mathbf{y}\| = 1$  now be such that

$$\|(AB)\| = \|(AB)\mathbf{y}\|.$$

Again by (7), we have

$$\begin{aligned} \|AB\| &\leq \|A\| \cdot \|B\mathbf{y}\| \\ &\leq \|A\| \cdot \|B\|, \end{aligned}$$

so that (5) and (6) do define a matrix norm.

We shall now determine the natural matrix norms induced by some of the vector  $p$ -norms ( $p = 1, 2, \infty$ ) defined in (1). Let the  $n$ th order matrix  $A$  have elements  $a_{jk}$ ,  $j, k = 1, 2, \dots, n$ .

(A) The matrix norm induced by the *maximum norm* (1d) is

$$(8) \quad \|A\|_\infty = \max_j \sum_{k=1}^n |a_{jk}|,$$

i.e., the *maximum absolute row sum*. To prove (8), let  $\mathbf{y}$  be such that  $\|\mathbf{y}\|_\infty = 1$  and

$$\|A\|_\infty = \|A\mathbf{y}\|_\infty.$$

Then,

$$\begin{aligned} \|A\|_\infty &= \max_j \left| \sum_{k=1}^n a_{jk} y_k \right| \leq \max_j \left( \sum_{k=1}^n |a_{jk}| |y_k| \right) \\ &\leq \max_k |y_k| \cdot \max_j \sum_{k=1}^n |a_{jk}| = \|\mathbf{y}\|_\infty \cdot \max_j \sum_{k=1}^n |a_{jk}| \\ &= \max_j \sum_{k=1}^n |a_{jk}|, \end{aligned}$$

so the right-hand side of (8) is an upper bound of  $\|A\|_\infty$ . Now if the maximum row sum occurs for, say,  $j = J$  then let  $\mathbf{x}$  have the components

$$x_k = \begin{cases} \bar{a}_{Jk}/|a_{Jk}|, & a_{Jk} \neq 0 \\ 0, & a_{Jk} = 0, \end{cases} \quad k = 1, 2, \dots, n.$$

Clearly  $\|\mathbf{x}\|_\infty = 1$ , if  $A$  is non-trivial, and

$$\|A\mathbf{x}\|_\infty = \sum_{k=1}^n |a_{Jk}| \leq \|A\|_\infty,$$

so (8) holds. [If  $A \equiv O$ , property (ii) implies  $\|A\| = 0$  for any natural norm.] ■

(B) Next, we claim that

$$(9) \quad \|A\|_1 = \max_k \sum_{j=1}^n |a_{jk}|,$$

i.e., the *maximum absolute column sum*. Now let  $\|\mathbf{y}\|_1 = 1$  and be such that

$$\|A\|_1 = \|A\mathbf{y}\|_1.$$

Then,

$$\begin{aligned} \|A\|_1 &= \sum_{j=1}^n \left| \sum_{k=1}^n a_{jk} y_k \right| \leq \sum_{j=1}^n \sum_{k=1}^n |a_{jk}| \cdot |y_k| \\ &= \sum_{k=1}^n \left( |y_k| \sum_{j=1}^n |a_{jk}| \right) \leq \sum_{k=1}^n |y_k| \left( \max_m \sum_{j=1}^n |a_{jm}| \right) \\ &= \|\mathbf{y}\|_1 \max_m \sum_{j=1}^n |a_{jm}| = \max_m \sum_{j=1}^n |a_{jm}|, \end{aligned}$$

and the right-hand side of (9) is an upper bound of  $\|A\|_1$ . If the maximum is attained for  $m = K$ , then this bound is actually attained for  $\mathbf{x} = \mathbf{e}_K$ , the  $K$ th unit vector, since  $\|\mathbf{e}_K\|_1 = 1$  and

$$\begin{aligned} \|A\mathbf{e}_K\|_1 &= \sum_{j=1}^n \left| \sum_{k=1}^n a_{jk} \delta_{kK} \right| \\ &= \sum_{j=1}^n |a_{jK}|. \end{aligned}$$

Thus (9) is established. ■

(C) Finally, we consider the Euclidean norm, for which case we recall the notation for the *Hermitian transpose* or *conjugate transpose* of any rectangular matrix  $A \equiv (a_{ij})$ ,

$$A^* \equiv \bar{A}^T,$$

i.e., if  $A^* \equiv (b_{ij})$ , then  $b_{ij} = \bar{a}_{ji}$ . Further, the *spectral radius* of any square matrix  $A$  is defined by

$$(10) \quad \rho(A) \equiv \max_s |\lambda_s(A)|,$$

where  $\lambda_s(A)$  denotes the  $s$ th eigenvalue of  $A$ . Now we can state that

$$(11) \quad \|A\|_2 = \sqrt{\rho(A^* A)}.$$

To prove (11), we again pick  $\mathbf{y}$  such that  $\|\mathbf{y}\|_2 = 1$  and

$$\|A\|_2 = \|A\mathbf{y}\|_2.$$

From (1b) it is clear that  $\|\mathbf{x}\|_2^2 = \mathbf{x}^* \mathbf{x}$ , since  $\mathbf{x}^* \equiv (\bar{x}_1, \bar{x}_2, \dots, \bar{x}_n)$ . Therefore, from the identity  $(A\mathbf{y})^* = \mathbf{y}^* A^*$ , we find

$$\begin{aligned} (12) \quad \|A\|_2^2 &= \|A\mathbf{y}\|_2^2 = (A\mathbf{y})^*(A\mathbf{y}) \\ &= \mathbf{y}^* A^* A \mathbf{y}. \end{aligned}$$

But since  $A^*A$  is Hermitian it has a complete set of  $n$  orthonormal eigenvectors, say  $\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n$ , such that

$$(13a) \quad \mathbf{u}_j^* \mathbf{u}_k = \delta_{jk},$$

$$(13b) \quad A^* A \mathbf{u}_s = \lambda_s \mathbf{u}_s.$$

The multiplication of (13b) by  $\mathbf{u}_s^*$  on the left yields further

$$\lambda_s = \mathbf{u}_s^* A^* A \mathbf{u}_s \geq 0.$$

Every vector has a unique expansion in the basis  $\{\mathbf{u}_s\}$ . Say in particular that

$$\mathbf{y} = \sum_{s=1}^n \alpha_s \mathbf{u}_s,$$

and then (12) becomes, upon recalling (13),

$$\begin{aligned} \|A\|_2^2 &= \sum_{t=1}^n \bar{\alpha}_t \mathbf{u}_t^* A^* A \sum_{s=1}^n \alpha_s \mathbf{u}_s \\ &= \sum_{t=1}^n \bar{\alpha}_t \mathbf{u}_t^* \sum_{s=1}^n \alpha_s \lambda_s \mathbf{u}_s = \sum_{s=1}^n \lambda_s |\alpha_s|^2 \\ &\leq \max_s \lambda_s \sum_{t=1}^n |\alpha_t|^2 = \max_s \lambda_s = \rho(A^* A). \end{aligned}$$

Thus  $\rho^{1/2}(A^* A)$  is an upper bound of  $\|A\|_2$ . However, using  $\mathbf{y} = \mathbf{u}_s$ , where  $\lambda_s = \rho(A^* A)$ , we get

$$\begin{aligned} \|A\mathbf{u}_s\|_2 &= (\mathbf{u}_s^* A^* A \mathbf{u}_s)^{1/2} \\ &= \rho^{1/2}(A^* A), \end{aligned}$$

and so (11) follows. ■

We have observed that a matrix of order  $n$  can be considered as a vector of dimension  $n^2$ . But since every matrix norm satisfies the conditions (0)–(iii) of a vector norm the results of Lemma 1 and Theorem 2 also apply to matrix norms. Thus we have

**LEMMA 1'.** *Every matrix norm,  $\|A\|$ , is a continuous function of the  $n^2$  elements  $a_{ij}$  of  $A$ .* ■

**THEOREM 2'.** *For each pair of matrix norms, say  $\|A\|$  and  $\|A\|'$ , there exist positive constants  $m$  and  $M$  such that for all  $n$ th order matrices  $A$*

$$m\|A\|' \leq \|A\| \leq M\|A\|'. \quad \blacksquare$$

The proofs of these results follow exactly the corresponding proofs for vector norms so we leave their detailed exposition to the reader.

There is frequently confusion between the spectral radius (10) of a matrix and the Euclidean norm (11) of a matrix. (To add to this confusion,  $\|A\|_2$  is sometimes called the *spectral norm* of  $A$ .) It should be observed that if  $A$  is Hermitian, i.e.,  $A^* = A$ , then  $\lambda_s(A^*A) = \lambda_s^2(A)$  and so *the spectral radius is equal to the Euclidean norm for Hermitian matrices*. However, in general this is not true, but we have

**LEMMA 2.** *For any natural norm,  $\|\cdot\|$ , and square matrix,  $A$ ,*

$$\rho(A) \leq \|A\|.$$

*Proof.* For each eigenvalue  $\lambda_s(A)$  there is a corresponding eigenvector, say  $\mathbf{u}_s$ , which can be chosen to be normalized for any particular vector norm,  $\|\mathbf{u}_s\| = 1$ . But then for the corresponding natural matrix norm

$$\|A\| = \max_{\|\mathbf{y}\|=1} \|A\mathbf{y}\| \geq \|A\mathbf{u}_s\| = \|\lambda_s\mathbf{u}_s\| = |\lambda_s|.$$

As this holds for all  $s = 1, 2, \dots, n$ , the result follows. ■

On the other hand, for each matrix some natural norm is arbitrarily close to the spectral radius. More precisely we have

**THEOREM 3.** *For each  $n$ th order matrix  $A$  and each arbitrary  $\epsilon > 0$  a natural norm,  $\|A\|$ , can be found such that*

$$\rho(A) \leq \|A\| \leq \rho(A) + \epsilon.$$

*Proof.* The left-hand inequality has been verified above. We shall show how to construct a norm satisfying the right-hand inequality. By Theorem 1 we can find a non-singular matrix  $P$  such that

$$PAP^{-1} \equiv B \equiv \Lambda + U$$

where  $\Lambda = (\lambda_j(A)\delta_{ij})$  and  $U \equiv (u_{ij})$  has zeros on and below the diagonal. With  $\delta > 0$ , a “sufficiently small” positive number, we form the diagonal matrix of order  $n$

$$D \equiv (\delta^{1-j}\delta_{ij}) = \begin{pmatrix} 1 & & & & 0 \\ & \delta^{-1} & & & \\ & & \ddots & & \\ 0 & & & & \delta^{1-n} \end{pmatrix}.$$

Now consider

$$C = DBD^{-1} = \Lambda + E,$$

where  $E \equiv (e_{ij}) = DUD^{-1}$  has elements

$$e_{ij} = \begin{cases} 0, & j \leq i \\ u_{ij}\delta^{j-i}, & j > i, \quad i = 1, 2, \dots, n. \end{cases}$$

Note that the elements  $e_{ij}$  can be made arbitrarily small in magnitude by choosing  $\delta$  appropriately. Also we have that

$$A = P^{-1}D^{-1}CDP.$$

Since  $DP$  is non-singular, a vector norm can be defined by

$$\|\mathbf{x}\| \equiv N_2(DP\mathbf{x}) = (\mathbf{x}^*P^*D^*DP\mathbf{x})^{1/2}.$$

The proof of this fact is left to the reader in Problem 5. The natural matrix norm induced by this vector norm is of course

$$\|A\| \equiv \max_{\|\mathbf{y}\|=1} \|A\mathbf{y}\|.$$

However, from the above form for  $A$ , we have, for any  $\mathbf{y}$ ,

$$\|A\mathbf{y}\| = N_2(DPA\mathbf{y}) = N_2(CDP\mathbf{y}).$$

If we let  $\mathbf{z} \equiv DP\mathbf{y}$ , this becomes

$$\|A\mathbf{y}\| = N_2(C\mathbf{z}) = (\mathbf{z}^*C^*C\mathbf{z})^{1/2}.$$

Now observe that

$$\begin{aligned} C^*C &= (\Lambda^* + E^*)(\Lambda + E) \\ &= \Lambda^*\Lambda + \mathcal{M}(\delta). \end{aligned}$$

Here the term  $\mathcal{M}(\delta)$  represents an  $n$ th order matrix each of whose terms is  $\mathcal{O}(\delta)$ .† Thus, we can conclude that

$$\begin{aligned} \mathbf{z}^*C^*C\mathbf{z} &\leq \max_s |\lambda_s^2(A)| \mathbf{z}^*\mathbf{z} + |\mathbf{z}^*\mathcal{M}(\delta)\mathbf{z}| \\ &\leq [\rho^2(A) + \mathcal{O}(\delta)]\mathbf{z}^*\mathbf{z}, \end{aligned}$$

since

$$|\mathbf{z}^*\mathcal{M}(\delta)\mathbf{z}| \leq n^2 \mathbf{z}^*\mathbf{z} \mathcal{O}(\delta) = \mathbf{z}^*\mathbf{z} \mathcal{O}(\delta).$$

Recalling  $\|\mathbf{y}\| = N_2(\mathbf{z})$ , we find from  $\|\mathbf{y}\| = 1$  that  $\mathbf{z}^*\mathbf{z} = 1$ . Then it follows that

$$\begin{aligned} \|A\| &\leq [\rho^2(A) + \mathcal{O}(\delta)]^{1/2} \\ &= \rho(A) + \mathcal{O}(\delta). \end{aligned}$$

For  $\delta$  sufficiently small  $\mathcal{O}(\delta) < \epsilon$ . ■

† A quantity, say  $f$ , is said to be  $\mathcal{O}(\delta)$ , or briefly  $f = \mathcal{O}(\delta)$  iff for some constants  $K \geq 0$  and  $\delta_0 > 0$ ,

$$|f| \leq K |\delta|, \quad \text{for } |\delta| \leq \delta_0.$$

It should be observed that the natural norm employed in Theorem 3 depends upon the matrix  $A$  as well as the arbitrary small parameter  $\epsilon$ . However, this result leads to an interesting characterization of the spectral radius of any matrix; namely,

**COROLLARY.** *For any square matrix  $A$*

$$\rho(A) = \inf_{\{N(\cdot)\}} \left( \max_{N(\mathbf{x})=1} N(A\mathbf{x}) \right)$$

*where the inf is taken over all vector norms,  $N(\cdot)$ ; or equivalently*

$$\rho(A) = \inf_{\{\|\cdot\|\}} \|A\|$$

*where the inf is taken over all natural norms,  $\|\cdot\|$ .*

*Proof.* By using Lemma 2 and Theorem 3, since  $\epsilon > 0$  is arbitrary and the natural norm there depends upon  $\epsilon$ , the result follows from the definition of inf. ■

### 1.1. Convergent Matrices

To study the convergence of various iteration procedures as well as for many other purposes, we investigate matrices  $A$  for which

$$(14) \quad \lim_{m \rightarrow \infty} A^m = O,$$

where  $O$  denotes the *zero matrix* all of whose entries are 0. Any square matrix satisfying condition (14) is said to be *convergent*. Equivalent conditions are contained in

**THEOREM 4.** *The following three statements are equivalent:*

- (a)  $A$  is convergent;
- (b)  $\lim_{m \rightarrow \infty} \|A^m\| = 0$ , for some matrix norm;
- (c)  $\rho(A) < 1$ .

*Proof.* We first show that (a) and (b) are equivalent. Since  $\|\cdot\|$  is continuous, by Lemma 1', and  $\|O\| = 0$ , then (a) implies (b). But if (b) holds for some norm, then Theorem 2' implies there exists an  $M$  such that

$$\|A^m\|_\infty \leq M \|A^m\| \rightarrow 0.$$

Hence, (a) holds.

Next we show that (b) and (c) are equivalent. Note that by Theorem 2' there is no loss in generality if we assume the norm to be a natural norm. But then, by Lemma 2 and the fact that  $\lambda(A^m) = \lambda^m(A)$ , we have

$$\|A^m\| \geq \rho(A^m) = \rho^m(A),$$

so that (b) implies (c). On the other hand, if (c) holds, then by Theorem 3 we can find an  $\epsilon > 0$  and a natural norm, say  $N(\cdot)$ , such that

$$N(A) \leq \rho(A) + \epsilon \equiv \theta < 1.$$

Now use the property (iv) of matrix norms to get

$$N(A^m) \leq [N(A)]^m \leq \theta^m$$

so that  $\lim_{m \rightarrow \infty} N(A^m) = 0$  and hence (b) holds.  $\blacksquare$

A test for convergent matrices which is frequently easy to apply is the content of the

**COROLLARY.** *A is convergent if for some matrix norm*

$$\|A\| < 1.$$

*Proof.* Again by (iv) we have

$$\|A^m\| \leq \|A\|^m$$

so that condition (b) of Theorem 4 holds.  $\blacksquare$

Another important characterization and property of convergent matrices is contained in

**THEOREM 5.** (a) *The geometric series*

$$I + A + A^2 + A^3 + \dots,$$

*converges iff A is convergent.*

(b) *If A is convergent, then I - A is non-singular and*

$$(I - A)^{-1} = I + A + A^2 + A^3 + \dots.$$

*Proof.* A necessary condition for the series in part (a) to converge is that  $\lim_{m \rightarrow \infty} A^m = O$ , i.e., that A be convergent. The sufficiency will follow from part (b).

Let A be convergent, whence by Theorem 4 we know that  $\rho(A) < 1$ . Since the eigenvalues of  $I - A$  are  $1 - \lambda(A)$ , it follows that  $\det(I - A) \neq 0$  and hence this matrix is non-singular. Now consider the identity

$$(I - A)(I + A + A^2 + \dots + A^m) = I - A^{m+1}$$

which is valid for all integers  $m$ . Since A is convergent, the limit as  $m \rightarrow \infty$  of the right-hand side exists. The limit, after multiplying both sides on the left by  $(I - A)^{-1}$ , yields

$$(I + A + A^2 + \dots) = (I - A)^{-1}$$

and part (b) follows.  $\blacksquare$

A useful corollary to this theorem is

**COROLLARY.** *If in some natural norm,  $\|A\| < 1$ , then  $I - A$  is non-singular and*

$$\frac{1}{1 + \|A\|} \leq \|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

*Proof.* By the corollary to Theorem 4 and part (b) of Theorem 5 it follows that  $I - A$  is non-singular. For a natural norm we note that  $\|I\| = 1$  and so taking the norm of the identity

$$I = (I - A)(I - A)^{-1}$$

yields

$$\begin{aligned} 1 &\leq \|(I - A)\| \cdot \|(I - A)^{-1}\| \\ &\leq (1 + \|A\|) \|(I - A)^{-1}\|. \end{aligned}$$

Thus the left-hand inequality is established.

Now write the identity as

$$(I - A)^{-1} = I + A(I - A)^{-1}$$

and take the norm to get

$$\|(I - A)^{-1}\| \leq 1 + \|A\| \cdot \|(I - A)^{-1}\|.$$

Since  $\|A\| < 1$  this yields

$$\|(I - A)^{-1}\| \leq \frac{1}{1 - \|A\|}. \quad \blacksquare$$

It should be observed that if  $A$  is convergent, so is  $(-A)$ , and  $\|A\| = \|-A\|$ . Thus Theorem 5 and its corollary are immediately applicable to matrices of the form  $I + A$ . That is, if in some natural norm,  $\|A\| < 1$ , then

$$\frac{1}{1 + \|A\|} \leq \|(I + A)^{-1}\| \leq \frac{1}{1 - \|A\|}.$$

## PROBLEMS, SECTION 1

1. (a) Verify that (1b) defines a norm in the linear space of square matrices of order  $n$ ; i.e., check properties (i)–(iv), for  $\|A\|_E^2 = \sum_{ij} |a_{ij}|^2$ .

(b) Similarly, verify that (1a) defines a matrix norm, i.e.,  $\|A\| = \sum_{ij} |a_{ij}|$ .

2. Show by example that the maximum vector norm,  $\eta(A) = \max_{i,j} |a_{ij}|$ , when applied to a matrix, does not satisfy condition (iv) that we impose on a matrix norm.

3. Show that if  $A$  is non-singular, then  $B \equiv A^*A$  is Hermitian and positive definite. That is,  $\mathbf{x}^*B\mathbf{x} > 0$  if  $\mathbf{x} \neq \mathbf{0}$ . Hence the eigenvalues of  $B$  are all positive.

4. Show for any non-singular matrix  $A$  and any matrix norm that

$$\|I\| \geq 1 \quad \text{and} \quad \|A^{-1}\| \geq \frac{1}{\|A\|}.$$

[Hint:  $\|I\| = \|II\| \leq \|I\|^2$ ;  $\|A^{-1}A\| \leq \|A^{-1}\| \cdot \|A\|$ .]

5. Show that if  $\eta(\mathbf{x})$  is a norm and  $A$  is any non-singular matrix, then  $N(\mathbf{x})$  defined by

$$N(\mathbf{x}) \equiv \eta(A\mathbf{x}),$$

is a (vector) norm.

6. We call  $\eta(\mathbf{x})$  a *semi-norm* iff  $\eta(\mathbf{x})$  satisfies all of the conditions, (0)–(iii), for a norm with condition (i) replaced by the weaker condition

$$(i'): \eta(\mathbf{x}) \geq 0 \text{ for all } \mathbf{x} \in \mathcal{V}.$$

We say that  $\eta(\mathbf{x})$  is *non-trivial* iff  $\eta(\mathbf{x}) > 0$  for some  $\mathbf{x} \in \mathcal{V}$ . Prove the following generalization of Lemma 1:

**LEMMA 1".** Every non-trivial semi-norm,  $\eta(\mathbf{x})$ , is a continuous function of  $x_1, x_2, \dots, x_n$ , the components of  $\mathbf{x}$ . Hence every semi-norm is continuous.

7. Show that if  $\eta(\mathbf{x})$  is a semi-norm and  $A$  any square matrix, then  $N(\mathbf{x}) \equiv \eta(A\mathbf{x})$  defines a semi-norm.

## 2. FLOATING-POINT ARITHMETIC AND ROUNDING ERRORS

In the following chapters we will have to refer, on occasion, to the errors due to “rounding” in the basic arithmetic operations. Such errors are inherent in all computations in which only a fixed number of digits are retained. This is, of course, the case with all modern digital computers and we consider here an example of one way in which many of them do or can do arithmetic; so-called *floating-point arithmetic*. Although most electronic computers operate with numbers in some kind of binary representation, most humans still think in terms of a decimal representation and so we shall employ the latter here.

Suppose the number  $a \neq 0$  has the exact decimal representation

$$(1) \quad a = \pm 10^q(.d_1d_2\dots)$$

where  $q$  is an integer and the  $d_1, d_2, \dots$  are digits with  $d_1 \neq 0$ . Then the “ $t$ -digit floating-decimal representation of  $a$ ,” or for brevity the “floating  $a$ ” used in the machine, is of the form

$$(2) \quad \text{fl}(a) \equiv \pm 10^q(. \delta_1 \delta_2 \dots \delta_t)$$

where  $\delta_1 \neq 0$  and  $\delta_1, \delta_2, \dots, \delta_t$  are digits. The number  $(. \delta_1 \delta_2 \dots \delta_t)$  is

called the *mantissa* and  $q$  is called the *exponent* of  $\text{fl}(a)$ . There is usually a restriction on the exponent, of the form

$$(3) \quad -N \leq q \leq M,$$

for some large positive integers  $N, M$ . If a number  $a \neq 0$  has an exponent outside of this range it cannot be represented in the form (2), (3). If, during the course of a calculation, some computed quantity has an exponent  $q > M$  (called *overflow*) or  $q < -N$  (called *underflow*), meaningless results usually follow. However, special precautions can be taken on most computers to at least detect the occurrence of such over- or underflows. We do not consider these practical difficulties further; rather, we shall assume that they do not occur or are somehow taken into account.

There are two popular ways in which the floating digits  $\delta_j$  are obtained from the exact digits,  $d_j$ . The obvious *chopping* representation takes

$$(4) \quad \delta_j = d_j, \quad j = 1, 2, \dots, t.$$

Thus the exact mantissa is chopped off after the  $t$ th decimal digit to get the floating mantissa. The other and preferable procedure is to *round*, in which case†

$$(5) \quad \delta_1 \delta_2 \cdots \delta_t = [d_1 d_2 \cdots d_t . d_{t+1} + 0.5]$$

and the brackets on the right-hand side indicate the integral part. The error in either of these procedures can be bounded as in

**LEMMA 1.** *The error in  $t$ -digit floating-decimal representation of a number  $a \neq 0$  is bounded by*

$$|a - \text{fl}(a)| \leq 5|a|10^{-t}p \quad \begin{cases} p = 1, & \text{rounded,} \\ p = 2, & \text{chopped.} \end{cases}$$

*Proof.* From (1), (2), and (4) we have

$$\begin{aligned} |a - \text{fl}(a)| &= 10^{q-t}(.d_{t+1}d_{t+2}\cdots) \\ &= 10^{q-t}(.d_{t+1}d_{t+2}\cdots) \frac{|a|}{|a|} \\ &= 10^{-t} \frac{(.d_{t+1}d_{t+2}\cdots)}{(.d_1d_2\cdots)} |a|. \end{aligned}$$

† For simplicity we are neglecting the special case that occurs when  $d_1 = d_2 = \cdots = d_t = 9$  and  $d_{t+1} \geq 5$ . Here we would increase the exponent  $q$  in (2) by unity and set  $\delta_1 = 1, \delta_j = 0, j > 1$ . Note that when  $d_{t+1} = 5$ , if we were to round up iff  $d_t$  is odd, then an *unbiased rounding* procedure would result. Some electronic computers employ an unbiased rounding procedure (in a binary system).

But since  $1 \leq d_1 \leq 9$  and  $0.d_{t+1}d_{t+2}\dots \leq 1$  this implies

$$|a - \text{fl}(a)| \leq 10^{1-t}|a|,$$

which is the bound for the chopped representation. For the case of rounding we have, similarly,

$$|a - \text{fl}(a)| \leq \frac{1}{2} 10^{q-t} = \frac{1}{2} 10^{q-t} \frac{|a|}{|a|} \leq 5|a|10^{-t}. \quad \blacksquare$$

We shall assume that our idealized computer performs each basic arithmetic operation correctly to  $2t$  digits and then either rounds or chops the result to a  $t$ -digit floating number. With such operations it clearly follows from Lemma 1 that

$$\left. \begin{array}{l} (6a) \quad \text{fl}(a \pm b) = (a \pm b)(1 + \phi 10^{-t}) \\ (6b) \quad \text{fl}(ab) = a \cdot b(1 + \phi 10^{-t}) \\ (6c) \quad \text{fl}\left(\frac{a}{b}\right) = \frac{a}{b}(1 + \phi 10^{-t}) \end{array} \right\} \begin{array}{l} 0 \leq |\phi| \leq 5 \quad \text{rounding,} \\ 0 \leq |\phi| \leq 10 \quad \text{chopping.} \end{array}$$

In many calculations, particularly those concerned with linear systems, the accumulation of products is required (e.g., the inner product of two vectors). We assume that rounding (or chopping) is done after each multiplication and after each successive addition. That is,

$$(7a) \quad \text{fl}(a_1b_1 + a_2b_2) = [a_1b_1(1 + \phi_1 10^{-t}) + a_2b_2(1 + \phi_2 10^{-t})](1 + \theta 10^{-t})$$

and in general

$$(7b) \quad \text{fl}\left(\sum_{i=1}^n a_i b_i\right) = \text{fl}\left[\text{fl}\left(\sum_{i=1}^{n-1} a_i b_i\right) + \text{fl}(a_n b_n)\right].$$

The result of such computations can be represented as an exact inner product with, say, the  $a_i$  slightly altered. We state this as

**LEMMA 2.** *Let the floating-point inner product (7) be computed with rounding. Then if  $n$  and  $t$  satisfy*

$$(8) \quad n10^{1-t} \leq 1$$

*it follows that*

$$(9a) \quad \text{fl}\left(\sum_{i=1}^n a_i b_i\right) = \sum_{i=1}^n (a_i + \delta a_i) b_i$$

*where*

$$(9b) \quad |\delta a_1| \leq n|a_1|10^{1-t}, \quad |\delta a_i| \leq (n-i+2)|a_i|10^{1-t}, \quad i = 2, 3, \dots, n.$$

*Proof.* By (6b) we can write

$$\text{fl}(a_k b_k) = a_k b_k (1 + \phi_k 10^{-t}), \quad |\phi_k| \leq 5,$$

since rounding is assumed. Similarly from (6a) and (7b) with  $n = k$  we have

$$\text{fl}\left(\sum_{i=1}^k a_i b_i\right) = \left[\text{fl}\left(\sum_{i=1}^{k-1} a_i b_i\right) + (a_k b_k)(1 + \phi_k 10^{-t})\right](1 + \theta_k 10^{-t})$$

where

$$\theta_1 = 0; \quad |\theta_k| \leq 5, \quad k = 2, 3, \dots$$

Now a simple recursive application of the above yields

$$\begin{aligned} \text{fl}\left(\sum_{i=1}^n a_i b_i\right) &= \sum_{k=1}^n \left[ a_k b_k (1 + \phi_k 10^{-t}) \prod_{j=k}^n (1 + \theta_j 10^{-t}) \right] \\ &\equiv \sum_{k=1}^n a_k b_k (1 + E_k), \end{aligned}$$

where we have introduced  $E_k$  by

$$1 + E_k \equiv (1 + \phi_k 10^{-t}) \prod_{j=k}^n (1 + \theta_j 10^{-t}).$$

A formal verification of this result is easily obtained by induction.

Since  $\theta_1 = 0$ , it follows that

$$(1 - 5 \cdot 10^{-t})^{n-k+2} \leq 1 + E_k \leq (1 + 5 \cdot 10^{-t})^{n-k+2}, \quad k = 2, 3, \dots, n,$$

and

$$(1 - 5 \cdot 10^{-t})^n \leq 1 + E_1 \leq (1 + 5 \cdot 10^{-t})^n.$$

Hence, with  $\epsilon = 5 \cdot 10^{-t}$ ,

$$|E_1| \leq (1 + \epsilon)^n - 1,$$

$$|E_k| \leq (1 + \epsilon)^{n-k+2} - 1, \quad k = 2, 3, \dots, n.$$

But, for  $p \leq n$ , (8) implies that  $p\epsilon \leq \frac{1}{2}$ , so that

$$\begin{aligned} (1 + \epsilon)^p - 1 &\equiv p\epsilon \left(1 + \frac{p-1}{2}\epsilon + \frac{p-1}{2}\frac{p-2}{3}\epsilon^2 + \dots\right) \\ &\leq p\epsilon(1 + \frac{1}{2} + (\frac{1}{2})^2 + \dots) \\ &\leq 2p\epsilon = p \cdot 10^{1-t}. \end{aligned}$$

Therefore,

$$|E_k| \leq (n - k + 2)10^{1-t}, \quad k = 2, 3, \dots, n.$$

Clearly for  $k = 1$  we find, as above with  $k = 2$ , that

$$|E_1| \leq n \cdot 10^{1-t}.$$

The result now follows upon setting

$$\delta a_k = a_k E_k.$$

(Note that we could just as well have set  $\delta b_k = b_k E_k$ .)  $\blacksquare$

Obviously a similar result can be obtained for the error due to chopping if condition (8) is strengthened slightly; see Problem 1.

## PROBLEMS, SECTION 2

1. Determine the result analogous to Lemma 2, when “chopping” replaces “rounding” in the statement.

[Hint: The factor  $10^{1-t}$  need only be replaced by  $2 \cdot 10^{1-t}$ , throughout.]

2. (a) Find a representation for  $\text{fl}\left(\sum_{i=1}^n c_i\right)$ .

(b) If  $c_1 > c_2 > \dots > c_n > 0$ , in what order should  $\text{fl}\left(\sum_{i=1}^n c_i\right)$  be calculated to minimize the effect of rounding?

3. What are the analogues of equations (6a, b, c) in the binary representation:

$$\text{fl}(a) = \pm 2^q(. \delta_1 \delta_2 \dots \delta_t)$$

where  $\delta_1 = 1$  and  $\delta_j = 0$  or  $1$ ?

## 3. WELL-POSED COMPUTATIONS

Hadamard introduced the notion of well-posed or properly posed problems in the theory of partial differential equations (see Section 0 of Chapter 9). However, it seems that a related concept is quite useful in discussing computational problems of almost all kinds. We refer to this as the notion of a well-posed computing problem.

First, we must clarify what is meant by a “computing problem” in general. Here we shall take it to mean an *algorithm* or equivalently: *a set of rules specifying the order and kind of arithmetic operations (i.e., rounding rules) to be used on specified data*. Such a computing problem may have as its object, for example, the determination of the roots of a quadratic equation or of an approximation to the solution of a nonlinear partial differential equation. How any such rules are determined for a particular purpose need not concern us at present (this is, in fact, what much of the rest of this book is about).

Suppose the specified data for some particular computing problem are the quantities  $a_1, a_2, \dots, a_m$ , which we denote as the  $m$ -dimensional vector  $\mathbf{a}$ . Then if the quantities to be computed are  $x_1, x_2, \dots, x_n$ , we can write

$$(1) \quad \mathbf{x} = \mathbf{f}(\mathbf{a}),$$

where of course the  $n$ -dimensional function  $\mathbf{f}(\cdot)$  is determined by the rules.

Now we will define a computing problem to be *well-posed* iff the algorithm meets **three requirements**. The *first requirement* is that a “solution,”  $\mathbf{x}$ , should *exist* for the given data,  $\mathbf{a}$ . This is implied by the notation (1). However, if we recall that (1) represents the evaluation of some algorithm it would seem that a solution (i.e., a result of using the algorithm) must always exist. But this is not true, a trivial example being given by data that lead to a division by zero in the algorithm. (The algorithm in this case is not properly specified since it should have provided for such a possibility. If it did not, then the corresponding computing problem is *not well-posed* for data that lead to this difficulty.) There are other, more subtle situations that result in algorithms which cannot be evaluated and it is by no means easy, *a priori*, to determine that  $\mathbf{x}$  is indeed defined by (1).

The *second requirement* is that the computation be *unique*. That is, when performed several times (with the same data) identical results are obtained. This is quite invariably true of algorithms which can be evaluated. If in actual practice it seems to be violated, the trouble usually lies with faulty calculations (i.e., machine errors). The functions  $\mathbf{f}(\mathbf{a})$  must be single valued to insure uniqueness.

The *third requirement* is that the result of the computation should depend *Lipschitz continuously on the data with a constant that is not too large*. That is, “small” changes in the data,  $\mathbf{a}$ , should result in only “small” changes in the computed  $\mathbf{x}$ . For example, let the computation represented by (1) satisfy the first two requirements for all data  $\mathbf{a}$  in some set, say  $\mathbf{a} \in D$ . If we change the data  $\mathbf{a}$  by a small amount  $\delta\mathbf{a}$  so that  $(\mathbf{a} + \delta\mathbf{a}) \in D$ , then we can write the result of the computation with the altered data as

$$(2) \quad \mathbf{x} + \delta\mathbf{x} = \mathbf{f}(\mathbf{a} + \delta\mathbf{a}).$$

Now if there exists a constant  $M$  such that for any  $\delta\mathbf{a}$ ,

$$(3) \quad \|\delta\mathbf{x}\| \leq M \|\delta\mathbf{a}\|,$$

we say that the computation depends Lipschitz continuously on the data. Finally, we say (1) is *well-posed* iff the three requirements are satisfied and (3) holds with a not too large constant,  $M = M(\mathbf{a}, \eta)$ , for some not too small  $\eta > 0$  and all  $\delta\mathbf{a}$  such that  $\|\delta\mathbf{a}\| \leq \eta$ . Since the Lipschitz constant

$M$  depends on  $(\mathbf{a}, \eta)$  we see that a computing problem or algorithm may be well-posed for some data,  $\mathbf{a}$ , but not for all data.

Let  $\mathcal{P}(\mathbf{a})$  denote the original problem which the algorithm (1) was devised to “solve.” This problem is also said to be *well-posed* if it has a unique solution, say

$$\mathbf{y} = \mathbf{g}(\mathbf{a}),$$

which depends Lipschitz continuously on the data. That is,  $\mathcal{P}(\mathbf{a})$  is well-posed if for all  $\delta\mathbf{a}$  satisfying  $\|\delta\mathbf{a}\| \leq \zeta$ , there is a constant  $N = N(\mathbf{a}, \zeta)$  such that

$$(4) \quad \|\mathbf{g}(\mathbf{a} + \delta\mathbf{a}) - \mathbf{g}(\mathbf{a})\| \leq N\|\delta\mathbf{a}\|.$$

We call the algorithm (1) *convergent* iff  $\mathbf{f}$  depends on a parameter, say  $\epsilon$  (e.g.,  $\epsilon$  may determine the size of the rounding errors), so that for any small  $\epsilon > 0$ ,

$$(5) \quad \|\mathbf{f}(\mathbf{a} + \delta\mathbf{a}) - \mathbf{g}(\mathbf{a} + \delta\mathbf{a})\| \leq \epsilon,$$

for all  $\delta\mathbf{a}$  such that  $\|\delta\mathbf{a}\| \leq \delta$ . Now, if  $\mathcal{P}(\mathbf{a})$  is well-posed and (1) is convergent, then (4) and (5) yield

$$(6) \quad \begin{aligned} \|\mathbf{f}(\mathbf{a}) - \mathbf{f}(\mathbf{a} + \delta\mathbf{a})\| &\leq \|\mathbf{f}(\mathbf{a}) - \mathbf{g}(\mathbf{a})\| + \|\mathbf{g}(\mathbf{a}) - \mathbf{g}(\mathbf{a} + \delta\mathbf{a})\| \\ &\quad + \|\mathbf{g}(\mathbf{a} + \delta\mathbf{a}) - \mathbf{f}(\mathbf{a} + \delta\mathbf{a})\| \\ &\leq \epsilon + N\|\delta\mathbf{a}\| + \epsilon. \end{aligned}$$

Thus, recalling (3), we are led to the heuristic

**OBSERVATION 1.** *If  $\mathcal{P}(\mathbf{a})$  is a well-posed problem, then a necessary condition that (1) be a convergent algorithm is that (1) be a well-posed computation.*

Therefore we are interested in determining whether a given algorithm (1) is a well-posed computation simply because *only such an algorithm is sure to be convergent* for all problems of the form  $\mathcal{P}(\mathbf{a} + \delta\mathbf{a})$ , when  $\mathcal{P}(\mathbf{a})$  is well-posed and  $\|\delta\mathbf{a}\| \leq \delta$ .

Similarly, by interchanging  $\mathbf{f}$  and  $\mathbf{g}$  in (6), we may justify

**OBSERVATION 2.** *If  $\mathcal{P}$  is a not well-posed problem, then a necessary condition that (1) be an accurate algorithm is that (1) be a not well-posed computation.*

In fact, for certain problems of linear algebra (see Subsection 1.2 of Chapter 2), it has been possible to prove that the commonly used algorithms, (1), produce approximations,  $\mathbf{x}$ , which are exact solutions of slightly perturbed original mathematical problems. In these algebraic cases, the accuracy of the solution  $\mathbf{x}$ , as measured in (5), is seen to depend on the well-posedness of the original mathematical problem. In algorithms,

(1), that arise from differential equation problems, other techniques are developed to estimate the accuracy of the approximation. For differential equation problems the well-posedness of the resulting algorithms (1) is referred to as the *stability* of the finite difference schemes (see Chapters 8 and 9).

We now consider two elementary examples to illustrate some of the previous notions.

The most overworked example of how a simple change in the algorithm can affect the accuracy of a single precision calculation is the case of determining the smallest root of a quadratic equation. If in

$$x^2 + 2bx + c,$$

$b < 0$  and  $c$  are given to  $t$  digits, but  $|c|/b^2 < 10^{-t}$ , then the smallest root,  $x_2$ , should be found from  $x_2 = c/x_1$ , after finding  $x_1 = -b + \sqrt{b^2 - c}$  in single precision arithmetic. Using

$$x_2 = -b - \sqrt{b^2 - c}$$

in single precision arithmetic would be disastrous!

A more sophisticated well-posedness discussion, without reference to the type of arithmetic, is afforded by the problem of determining the zeros of a polynomial

$$P_n(z) \equiv z^n + a_{n-1}z^{n-1} + \cdots + a_1z + a_0.$$

If  $Q_n(z) \equiv z^n + b_{n-1}z^{n-1} + \cdots + b_1z + b_0$ , then the zeros of  $P_n(z; \epsilon) \equiv P_n(z) + \epsilon Q_n(z)$  are “close” to the zeros of  $P_n(z)$ . That is, in the theory of functions of a complex variable it is shown that

LEMMA. *If  $z = z_1$  is a simple zero of  $P_n(z)$ , then for  $|\epsilon|$  sufficiently small  $P_n(z; \epsilon)$  has a zero  $z_1(\epsilon)$ , such that*

$$\left| z_1(\epsilon) - z_1 + \epsilon \frac{Q_n(z_1)}{P_n'(z_1)} \right| = \mathcal{O}(\epsilon^2).$$

*If  $z_1$  is a zero of multiplicity  $r$  of  $P_n(z)$ , there are  $r$  neighboring zeros of  $P_n(z; \epsilon)$  with*

$$\left| z_1(\epsilon) - z_1 - \left[ -\frac{r! Q_n(z_1)}{P_n^{(r)}(z_1)} \right]^{1/r} \epsilon^{1/r} \right| = \mathcal{O}(\epsilon^{2/r}).$$

Now it is clear that in the case of a simple zero,  $z_1$ , the computing problem, to determine the zero, might be well-posed if  $P_n'(z_1)$  were not too small and  $Q_n(z_1)$  not too large, since then  $|z_1(\epsilon) - z_1|/|\epsilon|$  would not be large for small  $\epsilon$ . On the other hand, the determination of the multiple root would most likely lead to a not well-posed computing problem.

The latter example illustrates Observation (2), that is, a computing problem is not well-posed if the original mathematical problem is not well-posed. On the other hand, the example of the quadratic equation indicates how an ill-chosen formulation of an algorithm may be well-posed but yet *inaccurate* in single precision.

Given an  $\epsilon > 0$  and a problem  $\mathcal{P}(\mathbf{a})$  we do not, in general, know how to determine an algorithm, (1), that requires the least amount of work to find  $\mathbf{x}$  so that  $\|\mathbf{x} - \mathbf{y}\| \leq \epsilon$ . This is an important aspect of algorithms for which there is no general mathematical theory. For most of the algorithms that are described in later chapters, we estimate the number of arithmetic operations required to find  $\mathbf{x}$ .

### PROBLEM, SECTION 3

#### 1. For the quadratic equation

$$x^2 + 2bx + c = 0,$$

find the small root by using single precision arithmetic in the iterative schemes

$$(a) \quad x_{n+1} = -\frac{c}{2b} - \frac{x_n^2}{2b},$$

and

$$(b) \quad x_{n+1} = x_n - \frac{x_n^2 + 2bx_n + c}{2x_n + 2b}.$$

If your computer has a mantissa with approximately  $t = 2p$  digits, use

$$c = 1, \quad b = -10^p$$

for the two initial values

$$(i) \quad x_0 = 0; \quad (ii) \quad x_0 = -\frac{c}{b}.$$

Which scheme gives the smaller root to approximately  $t$  digits with the smaller number of iterations? Which scheme requires less work?

# 2

## Numerical Solution of Linear Systems and Matrix Inversion

### 0. INTRODUCTION

Finding the solution of a linear algebraic equation system of “large” order and calculating the inverse of a matrix of “large” order can be difficult numerical tasks. While in principle there are standard methods for solving such problems, the difficulties are practical and stem from

- (a) the labor required in a lengthy sequence of calculations, and
- (b) the possible loss of accuracy in such lengthy calculations performed with a fixed number of decimal places.

The first difficulty renders manual computation impractical and the second limits the applicability of high speed digital computers with fixed “word” length. Thus to determine the feasibility of solving a particular problem with given equipment, several questions should be answered:

- (i) How many arithmetic operations are required to apply a proposed method?
- (ii) What will be the accuracy of a solution to be found by the proposed method (*a priori estimate*)?
- (iii) How can the accuracy of the computed answer be checked (*a posteriori estimate*)?

The first question can frequently† be answered in a straightforward manner and this is done, by means of an “operational count,” for most

† For “direct” methods, the operational count is easily made; while for “indirect” or iterative methods, the operational count is made by multiplying the estimated number of iterations by the operational count per iteration.

of the methods in this chapter. The third question can be easily answered if we have a bound for the norm of the inverse matrix. We therefore indicate, in Subsection 1.3, how such a bound may be obtained if we have an *approximate* inverse. However, the second question has only been recently answered for some methods. After discussing the notions of “well-posed problem” and “condition number” of a matrix, we give an account of Wilkinson’s a priori estimate for the Gaussian elimination method in Subsection 1.2. This treatment, in Section 1, of the Gaussian elimination method is followed, in Section 2, by a discussion of some modifications of the procedure. Direct factorization methods, which include Gaussian elimination as a special case, are described in Section 3. Iterative methods and techniques for accelerating them are studied in the remaining three sections.

The matrix inversion problem may be formulated as follows: *Given a square matrix of order n,*

$$(1) \quad A \equiv (a_{ij}) \equiv \begin{pmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{pmatrix}$$

*find its inverse, i.e., a square matrix of order n, say  $A^{-1}$ , such that*

$$(2) \quad A^{-1}A = AA^{-1} = I \equiv (\delta_{ij}).$$

Here  $I$  is the  $n$ th order identity matrix whose elements are given by the Kronecker delta:

$$(3) \quad \delta_{ij} \equiv \begin{cases} 0, & \text{if } i \neq j; \\ 1, & \text{if } i = j. \end{cases}$$

It is well known that this problem has one and only one solution iff the determinant of  $A$  is non-zero ( $\det A \neq 0$ ), i.e., iff  $A$  is non-singular.

The problem of solving a general linear system is formulated as follows: Given a square matrix  $A$  and an arbitrary  $n$ -component column vector  $\mathbf{f}$ , find a vector  $\mathbf{x}$  which satisfies

$$(4a) \quad A\mathbf{x} = \mathbf{f},$$

or, in component form,

$$(4b) \quad \begin{aligned} a_{11}x_1 + a_{12}x_2 + \cdots + a_{1n}x_n &= f_1, \\ a_{21}x_1 + a_{22}x_2 + \cdots + a_{2n}x_n &= f_2, \\ &\vdots \\ a_{n1}x_1 + a_{n2}x_2 + \cdots + a_{nn}x_n &= f_n. \end{aligned}$$

Again it is known that this problem has a solution which is unique for *every* inhomogeneous term  $\mathbf{f}$ , iff  $A$  is non-singular. [If  $A$  is singular the system (4) will have a solution only for *special* vectors  $\mathbf{f}$  and such a solution is not unique. The numerical solution of such “singular” problems is briefly touched on in Section 1 and in Problems 1.3, 1.4 of Chapter 4.]

It is easy to see that the problem of matrix inversion is equivalent to that of solving linear systems. For, let the inverse of  $A$ , assumed non-singular, be known and have elements  $c_{ij}$ , that is,

$$A^{-1} \equiv (c_{ij}).$$

Then multiplication of (4a) on the left by  $A^{-1}$  yields, since  $I\mathbf{x} = \mathbf{x}$ ,

$$(5a) \quad \mathbf{x} = A^{-1}\mathbf{f},$$

or componentwise,

$$(5b) \quad x_i = \sum_{j=1}^n c_{ij}f_j, \quad i = 1, 2, \dots, n.$$

Thus when the  $c_{ij}$  are known it requires, at most,  $n$  multiplications and  $(n - 1)$  additions to evaluate each component of the solution, or, for the complete solution, a total of  $n^2$  multiplications and  $n(n - 1)$  additions.

On the other hand, assume a procedure is known for solving the non-singular system (4) with an *arbitrary* inhomogeneous term  $\mathbf{f}$ . We then consider the  $n$  special systems

$$(6) \quad A\mathbf{x} = \mathbf{e}^{(j)}, \quad j = 1, 2, \dots, n,$$

where  $\mathbf{e}^{(j)}$  is the  $j$ th column of the identity matrix; that is, the elements of  $\mathbf{e}^{(j)}$  are  $e_i^{(j)} = \delta_{ij}$ ,  $i = 1, 2, \dots, n$ . The solutions of these systems are  $n$  vectors which we call  $\mathbf{x}^{(j)}$ ,  $j = 1, 2, \dots, n$ ; the components of  $\mathbf{x}^{(j)}$  are denoted by  $x_i^{(j)}$ . With these vectors we form the square matrix

$$(7) \quad X \equiv (x_i^{(j)}),$$

in which the  $j$ th column is the solution  $\mathbf{x}^{(j)}$  of the  $j$ th system in (6). Then it follows from the row by column rule for matrix multiplication that

$$(8) \quad AX = (\delta_{ij}) = I.$$

Since  $A$  was assumed to be non-singular, we find upon multiplying both sides of (8) on the left by  $A^{-1}$  that

$$X = A^{-1}.$$

Thus, by solving the  $n$  special systems (6) the inverse may be computed; this is the procedure generally used in practice. The number of operations required is, *at most*,  $n$  times that required to solve a single system. However,

this number can be reduced by efficiently organizing the computations and by taking account of the special form of the inhomogeneous terms,  $\mathbf{e}^{(j)}$ , as we shall explain later in Subsection 1.1.

## PROBLEMS, SECTION 0

1. If the columns of  $A$  form a set of vectors such that at most  $c$  of the columns are linearly independent, then we say that the *column rank* of  $A$  is  $c$ . (Similarly define the *row rank* of  $A$  to be  $r$  by replacing “columns” by “rows” and “ $c$ ” by “ $r$ .”) Prove that the row rank of  $A$  equals the column rank of  $A$ .

[Hint: Let row rank  $(A) \equiv r$ ; and use a set of  $r$  rows of  $A$  to define a submatrix  $B$  with row rank  $(B) = r$ . Then show that  $c \equiv \text{column rank } (A) = \text{column rank } (B)$ . Hence, since  $B$  has  $r$  rows,  $c \leq r$ . Similarly show that  $r \leq c$ .] Hence define *rank* of  $A$  by  $\text{rank } (A) \equiv r = c$ .

2. (*Alternative Principle*) If  $A$  is of order  $n$ , then either

$$Ax = \mathbf{0} \quad \text{iff } x = \mathbf{0},$$

or else

$$r \equiv \text{rank } (A) < n$$

and there exist a finite number,  $p$ , of linearly independent solutions  $\{\mathbf{x}^{(j)}\}$  that span the null space of  $A$ , i.e.,

$$Ax^{(j)} = \mathbf{0}, \quad j = 1, 2, \dots, p,$$

and if  $Ax = \mathbf{0}$  there exist constants  $\{\alpha_j\}$  such that  $x = \sum_{j=1}^p \alpha_j \mathbf{x}^{(j)}$ . Show that  $p = n - r$ .

3. Observe that

$$(9) \quad Ax = \mathbf{f}$$

has a solution iff  $\mathbf{f}$  is a linear combination of the columns of  $A$ . Hence show that: (9) has a solution  $x$  iff  $\text{rank } (A) = \text{rank } (A, \mathbf{f})$ ; (9) has a solution  $x$  iff  $y^T \mathbf{f} = 0$  for all vectors  $y \neq \mathbf{0}$  such that  $y^T A = \mathbf{0}$ . (In this problem,  $A$  may be rectangular;  $(A, \mathbf{f})$  is the *augmented matrix*).

## 1. GAUSSIAN ELIMINATION

The best known and most widely used method for solving linear systems of algebraic equations and for inverting matrices is attributed to Gauss. It is, basically, the elementary procedure in which the “first” equation is used to eliminate the “first” variable from the last  $n - 1$  equations, then the new “second” equation is used to eliminate the “second” variable from the last  $n - 2$  equations, etc. If  $n - 1$  such eliminations can be performed, then the resulting linear system which is equivalent† to the

† Two linear systems are *equivalent* iff every solution of one is a solution of the other.

original system is triangular and is easily solved. Of course, the ordering of the equations in the system and of the unknowns in the equations is arbitrary and so there is no unique order in which the procedure must be employed. As we shall see, the ordering is important since some orderings may not permit  $n - 1$  eliminations, while among the permissible orderings, some are to be preferred since they yield more accurate results.

In order to describe the specific sequence of arithmetic operations used in Gaussian elimination, we will first use the natural order in which the system is given, say

$$(1a) \quad A\mathbf{x} = \mathbf{f},$$

where

$$(1b) \quad A \equiv (a_{ij}), \quad \mathbf{x} \equiv \begin{pmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{pmatrix}, \quad \mathbf{f} \equiv \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_n \end{pmatrix}.$$

Before the variable  $x_k$  is eliminated, we denote the equivalent system (i.e., the reduced system), from which  $x_1, x_2, \dots, x_{k-1}$  have been eliminated, by

$$(2a) \quad A^{(k)}\mathbf{x} = \mathbf{f}^{(k)}, \quad k = 1, 2, \dots, n,$$

where

$$(2b) \quad A^{(k)} \equiv (a_{ij}^{(k)}), \quad \mathbf{f}^{(k)} \equiv \begin{pmatrix} f_1^{(k)} \\ f_2^{(k)} \\ \vdots \\ f_n^{(k)} \end{pmatrix}.$$

For  $k = 1$  we have  $A^{(1)} \equiv A$ ,  $\mathbf{f}^{(1)} \equiv \mathbf{f}$ , and the elements in (2b) for  $k = 2, 3, \dots, n$ , are computed recursively by

$$(3a) \quad a_{ij}^{(k)} = \begin{cases} a_{ij}^{(k-1)} & \text{for } i \leq k-1, \\ 0 & \text{for } i \geq k, j \leq k-1, \\ a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)} & \text{for } i \geq k, j \geq k; \end{cases}$$

$$(3b) \quad f_i^{(k)} = \begin{cases} f_i^{(k-1)} & \text{for } i \leq k-1, \\ f_i^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} f_{k-1}^{(k-1)} & \text{for } i \geq k. \end{cases}$$

These formulae represent the result of multiplying the  $(k-1)$ st equation in  $A^{(k-1)}\mathbf{x} = \mathbf{f}^{(k-1)}$  by the ratio  $(a_{i,k-1}^{(k-1)}/a_{k-1,k-1}^{(k-1)})$  and subtracting the

result from the  $i$ th equation for all  $i \geq k$ . In this way, the variable  $x_{k-1}$  is eliminated from the last  $n - k + 1$  equations. The resulting coefficient matrix and inhomogeneous term have the forms

$$(4) \quad A^{(k)} \equiv \begin{pmatrix} a_{11}^{(1)} & a_{12}^{(1)} & a_{13}^{(1)} & \cdots & a_{1,k-1}^{(1)} & a_{1k}^{(1)} & \cdots & a_{1n}^{(1)} \\ 0 & a_{22}^{(2)} & a_{23}^{(2)} & \cdots & a_{2,k-1}^{(2)} & a_{2k}^{(2)} & \cdots & a_{2n}^{(2)} \\ 0 & 0 & a_{33}^{(3)} & \cdots & \cdot & \cdot & \cdots & \cdot \\ & & & & \cdot & \cdot & \cdots & \cdot \\ 0 & & \cdot & & \cdot & \cdot & \cdots & \cdot \\ & & & a_{k-1,k-1}^{(k-1)} & a_{k-1,k}^{(k-1)} & \cdots & a_{k-1,n}^{(k-1)} \\ 0 & & & a_{kk}^{(k)} & \cdots & a_{kn}^{(k)} \\ 0 & & & a_{k+1,k}^{(k)} & \cdots & a_{k+1,n}^{(k)} \\ & & & \vdots & & \vdots \\ 0 & 0 & 0 & \cdots & 0 & a_{nk}^{(k)} & \cdots & a_{nn}^{(k)} \end{pmatrix}; \quad \mathbf{f}^{(k)} \equiv \begin{pmatrix} f_1^{(1)} \\ f_2^{(2)} \\ \vdots \\ f_{k-1}^{(k-1)} \\ f_k^{(k)} \\ f_{k+1}^{(k)} \\ \vdots \\ f_n^{(k)} \end{pmatrix}.$$

It has been assumed above that the elements  $a_{kk}^{(k)} \neq 0$  for  $k = 1, 2, \dots, n$ . When this is the case we have

**THEOREM 1.** *Let the matrix  $A$  be such that the Gaussian elimination procedure defined in (2)–(3) (i.e., in the natural order) yields non-zero diagonal elements  $a_{kk}^{(k)}$ ,  $k = 1, 2, \dots, n$ . Then  $A$  is non-singular and in fact,*

$$(5a) \quad \det A = a_{11}^{(1)} a_{22}^{(2)} \cdots a_{nn}^{(n)}.$$

*The final matrix  $A^{(n)} \equiv U$  is upper triangular and  $A$  has the factorization*

$$(5b) \quad LU = A,$$

*where  $L \equiv (m_{ik})$  is lower triangular with the elements*

$$(5c) \quad m_{ik} = \begin{cases} 0 & \text{for } i < k, \\ 1 & \text{for } i = k, \\ a_{ik}^{(k)} / a_{kk}^{(k)} & \text{for } i > k. \end{cases}$$

*The final vector  $\mathbf{f}^{(n)} \equiv \mathbf{g}$  is*

$$(5d) \quad \mathbf{g} = L^{-1}\mathbf{f}.$$

*Proof.* Once (5b) is established, we have that  $\det A = (\det L)(\det U) = \det U$  and so (5a) follows. To verify (5b), let us set  $LU = (c_{ij})$ . Then, since  $L$  and  $U$  are triangular and (4) is satisfied for  $k = n$ ,

$$\begin{aligned} c_{ij} &= \sum_{k=1}^n m_{ik} a_{kj}^{(n)}, \\ &= \sum_{k=1}^{\min(i,j)} m_{ik} a_{kj}^{(k)}. \end{aligned}$$

We recall that  $a_{ij}^{(1)} \equiv a_{ij}$  and note from (3a) and (5c) that

$$m_{i,k-1} a_{k-1,j}^{(k-1)} = a_{ij}^{(k-1)} - a_{ij}^{(k)} \quad \text{for } 2 \leq k \leq i, k \leq j.$$

Thus, if  $i \leq j$  we get from the above

$$\begin{aligned} c_{ij} &= \sum_{k=1}^{i-1} m_{ik} a_{kj}^{(k)} + a_{ij}^{(i)} \\ &= \sum_{k=1}^{i-1} (a_{ij}^{(k)} - a_{ij}^{(k+1)}) + a_{ij}^{(i)} \\ &= a_{ij}. \end{aligned}$$

This holds also for  $i > j$  since  $a_{ij}^{(j+1)} = 0$  and so (5b) is verified.

Define  $\mathbf{h} \equiv L\mathbf{g}$  so that

$$h_i = \sum_{k=1}^i m_{ik} g_k = \sum_{k=1}^i m_{ik} f_k^{(k)}.$$

Now from (3b) and (5c)

$$m_{ik} f_k^{(k)} = f_i^{(k)} - f_i^{(k+1)} \quad \text{for } k < i,$$

and  $f_i^{(1)} = f_i$ . Thus, we find  $h_i = f_i$ , and since  $L$  is non-singular (5d) follows. ■

Under the conditions of this theorem, the system (1) can be written as

$$LU\mathbf{x} = L\mathbf{g}.$$

Multiplication on the left by  $L^{-1}$  yields the equivalent upper triangular system

$$(6a) \quad U\mathbf{x} = \mathbf{g}.$$

If we write  $U \equiv (u_{ij})$ , then (5) is easily solved in the order  $x_n, x_{n-1}, \dots, x_1$  to get

$$(6b) \quad \begin{aligned} x_n &= \frac{1}{u_{nn}} g_n, \\ x_i &= \frac{1}{u_{ii}} \left( g_i - \sum_{j=i+1}^n u_{ij} x_j \right), \quad i = n-1, n-2, \dots, 1. \end{aligned}$$

We recall that the elements of  $U \equiv A^{(n)}$  and  $\mathbf{g} \equiv \mathbf{f}^{(n)}$  are computed by the Gaussian elimination procedure (3), without the explicit evaluation of  $L^{-1}$ .

We now consider the generalization in which the order of elimination is arbitrary. Again we set  $A^{(1)} \equiv A$  and  $\mathbf{f}^{(1)} \equiv \mathbf{f}$ . Then we select an arbitrary non-zero element  $a_{i_1 j_1}^{(1)}$  called the 1st pivot element. (If this cannot be done then  $A \equiv O$  and the system is degenerate, but also trivially in triangular

form.) Since  $a_{i_1 j_1}^{(1)} \neq 0$  is the coefficient of the  $j_1$ st variable,  $x_{j_1}$ , in the  $i_1$ st equation we can eliminate this variable from all of the other equations. To do this, we subtract an appropriate unique multiple of the  $i_1$ st equation from each of the other equations; i.e., to eliminate  $x_{j_1}$  from the  $k$ th equation the multiplier must be  $m_{k j_1} = (a_{k j_1}^{(1)} / a_{i_1 j_1}^{(1)})$ .

The reduced system is written as  $A^{(2)}\mathbf{x} = \mathbf{f}^{(2)}$  and it is such that omitting the  $i_1$ st equation yields  $n - 1$  equations in the  $n - 1$  unknowns  $x_k$ ,  $k \neq j_1$ . We now proceed with this reduced system and eliminate a second unknown, say  $x_{j_2}$ . To do this we must find some element  $a_{i_2 j_2}^{(2)} \neq 0$  with  $i_2 \neq i_1$  and  $j_2 \neq j_1$ , called the 2nd pivot element. If  $a_{rs}^{(2)} = 0$  for all  $r \neq i_1$  and  $s \neq j_1$  the process is terminated as the remaining equations are degenerate. After this second elimination the resulting system, say,  $A^{(3)}\mathbf{x} = \mathbf{f}^{(3)}$ , is composed of the  $i_1$ st equation of  $A^{(1)}\mathbf{x} = \mathbf{f}^{(1)}$ , the  $i_2$ nd equation of  $A^{(2)}\mathbf{x} = \mathbf{f}^{(2)}$  and  $n - 2$  remaining equations in only  $n - 2$  variables,  $x_k$  with  $k \neq j_1, j_2$ . The general process is now clear and can be used to prove

**THEOREM 2.** *Let the matrix  $A$  have rank  $r$ . Then we can find a sequence of distinct row and column indices  $(i_1, j_1), (i_2, j_2), \dots, (i_r, j_r)$  such that the corresponding pivot elements in  $A^{(1)}, A^{(2)}, \dots, A^{(r)}$  are non-zero and  $a_{ij}^{(r)} = 0$  if  $i \neq i_1, i_2, \dots, i_r$ . Let us define the permutation matrices, whose columns are unit vectors,*

$$P \equiv (\mathbf{e}^{(i_1)}, \mathbf{e}^{(i_2)}, \dots, \mathbf{e}^{(i_r)}, \dots, \mathbf{e}^{(i_n)}),$$

$$Q \equiv (\mathbf{e}^{(j_1)}, \mathbf{e}^{(j_2)}, \dots, \mathbf{e}^{(j_r)}, \dots, \mathbf{e}^{(j_n)}),$$

where  $i_k, j_k$ , for  $1 \leq k \leq r$ , are the above pivotal indices and the sets  $\{i_k\}$  and  $\{j_k\}$  are permutations of  $1, 2, \dots, n$ .

Then the system

$$By = g,$$

where

$$B \equiv P^T A Q, \quad \mathbf{y} \equiv Q^T \mathbf{x}, \quad \mathbf{g} \equiv P^T \mathbf{f},$$

is equivalent to the system (1) and can be reduced to triangular form by using Gaussian elimination with the natural order of pivots  $(1, 1), (2, 2), \dots, (r, r)$ .

*Proof.* The generalized elimination alters the matrix  $A \equiv A^{(1)}$  by forming successive linear combinations of the rows. Thus, whenever no non-zero pivot elements can be found the null rows must have been linearly dependent upon the rows containing non-zero pivots. The permutations by  $P$  and  $Q$  simply arrange the order of the equations and unknowns, respectively, so that  $b_{vv} = a_{i_v j_v}$ ,  $v = 1, 2, \dots, n$ . By the first part of the theorem, the reduced matrix  $B^{(r)}$  is triangular since all rows after the  $r$ th one vanish. ■

If the matrix  $A$  is non-singular, then  $r = n$  and Theorem 2 implies that, after the indicated rearrangement of the data, Theorem 1 becomes applicable. This is only useful for purposes of analysis. In actual computations on digital computers it is a simple matter to record the order of the pivot indices  $(i_v, j_v)$  for  $v = 1, 2, \dots, n$ , and to do the arithmetic accordingly. Of course, the important problem is to determine some order for the pivots so that the elimination can be completed.

One way to pick the pivots is to require that  $(i_k, j_k)$  be the indices of a maximal coefficient in the system of  $n - k + 1$  equations that remain at the  $k$ th step. This method of selecting *maximal pivots* is recommended as being likely to introduce the least loss of accuracy in the arithmetical operations that are based on working with a finite number of digits. We shall return to this feature in Subsection 1.2. Another commonly used pivotal selection method eliminates the variables  $x_1, x_2, \dots, x_{n-1}$  in succession by requiring that  $(i_k, k)$  be the indices of the maximal coefficient of  $x_k$  in the remaining system of  $n - k + 1$  equations. (This method of *maximal column pivots* is particularly convenient for use on an electronic computer if the large matrix of coefficients is stored by columns since the search for a maximal column element is then quicker than the maximal matrix element search.)

### 1.1. Operational Counts

If the  $n$ th order matrix is non-singular, Gaussian elimination might be employed to solve the  $n$  special linear systems (0.6) and thus to obtain  $A^{-1}$ . Then to solve any number, say  $m$ , of systems with the same coefficient matrix  $A$ , we need only perform  $m$  multiplications of vectors by  $A^{-1}$ . However, we shall show here that *for any value of  $m$  this procedure is less efficient than an appropriate application of Gaussian elimination to the  $m$  systems in question*. In order to show this, we must count the number of arithmetic operations required in the procedures to be compared. The current convention is to count only multiplications and divisions. This custom arose because the first modern digital computers performed additions and subtractions much faster than they did multiplications and divisions which were done in comparable lengths of time. This variation in the execution time of the arithmetic operations is at present being reduced, but it should be noted that additions and subtractions are about as numerous as multiplications for most methods of this chapter. On the other hand, for some computers, as in the case of a desk calculator, it is possible to accumulate a sequence of multiplications (scalar product of two vectors) in the same time that it takes to perform the multiplications. Hence one is justified in neglecting to count these additions since they do not contribute to the total time of performing the calculation.

Let us consider first the  $m$  systems, with arbitrary  $\mathbf{f}(j)$ ,

$$(7) \quad A\mathbf{x} = \mathbf{f}(j), \quad j = 1, 2, \dots, m.$$

We assume, for the operational count, that the elimination proceeds in the natural order. The most efficient application then performs the operations in (3a) only once and those in (3b) once for each  $j$ ; that is,  $m$  times. On digital computers, not all of the vectors  $\mathbf{f}(j)$  may be available at the same time, and thus the calculations in (3b) may be done later than those in (3a). However, since the final reduced matrix  $A^{(n)}$  is upper triangular, we may store the “multipliers”

$$m_{i, k-1} \equiv \frac{a_{i, k-1}^{(k-1)}}{a_{k-1, k-1}^{(k-1)}}, \quad 2 \leq k \leq i,$$

in the lower triangular part of the original matrix,  $A$ . (That is,  $m_{i, k-1}$  is put in the location of  $a_{i, k-1}^{(k-1)}$ ). Thus, no operations in (3a) ever need to be repeated.

From (3) and (4) we see that in eliminating  $x_{k-1}$ , a square submatrix of order  $n - k + 1$  is determined and the last  $n - k + 1$  components of each right-hand side are modified. Each element of the new submatrix and subvectors is obtained by performing a multiplication (and an addition which we ignore), but the quotients which appear as factors in (3) are computed only once. Thus, we find that it requires

$$\begin{aligned} (n - k + 1)^2 + (n - k + 1) &\quad \text{ops. for (3a),} \\ (n - k + 1) &\quad \text{ops. for (3b).} \end{aligned}$$

These operations must be done for  $k = 2, 3, \dots, n$  and hence with the aid of the formulae,

$$\sum_{\nu=1}^n \nu = \frac{n(n+1)}{2}, \quad \sum_{\nu=1}^n \nu^2 = \frac{n(n+1)(2n+1)}{6},$$

the total number of operations is found to be:

$$(8a) \quad \frac{n(n^2 - 1)}{3} \quad \text{ops. to triangularize } A,$$

$$(8b) \quad \frac{n(n-1)}{2} \quad \text{ops. to modify one inhomogeneous vector, } \mathbf{f}(j).$$

To solve the resulting triangular system we use (6). Thus, to compute  $x_i$  requires  $(n - i)$  multiplications and one division. By summing this over  $i = 1, 2, \dots, n$ , we get

$$(8c) \quad \frac{n(n+1)}{2} \quad \text{ops. to solve one triangular system.}$$

Finally, to solve the  $m$  systems in (7) we must do the operations in (8b) and (8c)  $m$  times while those in (8a) are done only once. Thus, we have

**LEMMA 1.**

$$(9) \quad \frac{n^3}{3} + mn^2 - \frac{n}{3} \quad \text{ops.}$$

are required to solve the  $m$  systems (7) by Gaussian elimination. ■

To compute  $A^{-1}$ , we could solve the  $n$  systems (0.6) so the above count would yield, upon setting  $m = n$ ,

$$\frac{4n^3}{3} - \frac{n}{3} \quad \text{ops. to compute } A^{-1}.$$

However, the  $n$  inhomogeneous vectors  $\mathbf{e}^{(j)}$  are quite special, each having only one non-zero component which is unity. If we take account of this fact, the above operational count can be reduced. That is, for any fixed  $j$ ,  $1 \leq j \leq n$ , the calculations to be counted in (3b) when  $\mathbf{f} \equiv \mathbf{e}^{(j)}$  start for  $k = j + 2$ . This follows, since  $f_v^{(v)} = 0$  for  $v = 1, 2, \dots, j - 1$  and  $f_j^{(j)} = 1$ . Thus, if  $j = n - 1$  or  $j = n$ , no multiplications are involved and in place of (8b), we have

$$\begin{aligned} \sum_{k=j+2}^n (n - k + 1) &= \sum_{v=1}^{n-j-1} v \\ &= \frac{1}{2}[j^2 - (2n - 1)j + n^2 - n] \quad \text{ops.} \end{aligned}$$

to modify the inhomogeneous vector  $\mathbf{e}^{(j)}$  for  $j = 1, 2, \dots, n - 2$ .

By summing this over the indicated values of  $j$ , we find

$$\frac{1}{6}(n^3 - 3n^2 + 2n) \quad \text{ops. to modify all } \mathbf{e}^{(j)}, \quad j = 1, 2, \dots, n.$$

The count in (8c) is unchanged and thus to solve the  $n$  resulting triangular systems takes

$$\frac{1}{2}(n^3 + n^2) \quad \text{ops.}$$

Upon combining the above with (8a) we find

**LEMMA 2.** *It need only require*

$$(10) \quad n^3 \quad \text{ops. to compute } A^{-1}. \quad ■$$

Now let us find the operational count for solving the  $m$  systems in (7) by employing the inverse. Since  $A^{-1}$  and  $\mathbf{f}(j)$  need not have any zero or unit elements, it requires in general

$$n^2 \quad \text{ops. to compute } A^{-1}\mathbf{f}(j).$$

Thus, to solve  $m$  systems requires  $mn^2$  ops. and if we include the  $n^3$  operations to compute  $A^{-1}$ , we get the result:

$$(11) \quad n^3 + mn^2 \quad \text{ops.}$$

*are required to solve the  $m$  systems (7), when using the inverse matrix.*

Upon comparing (9) and (11) it follows that *for any value of  $m$  the use of the inverse matrix is less efficient than using direct elimination.*

## 1.2. A Priori Error Estimates; Condition Number

In the course of actually carrying out the arithmetic required to solve

$$(12) \quad Ax = \mathbf{f}$$

by any procedure, roundoff errors will in general be introduced. But if the numerical procedure is “stable,” or if the problem is “well-posed” in the sense of Section 3 of Chapter 1, these errors can be kept within reasonable bounds. We shall investigate these matters for the Gaussian elimination method of solving (12).

We recall that a computation is said to be well-posed if “small” changes in the data cause “small” changes in the solution. For the linear system (12) the data are  $A$  and  $\mathbf{f}$  while  $\mathbf{x}$  is the solution. The matrix  $A$  is said to be “well-conditioned” or “ill-conditioned” if the computation is or is not, respectively, well-posed. We shall make these notions somewhat more precise here and introduce a *condition number* for  $A$  which serves as a measure of ill-conditioning. Then we will show that the Gaussian elimination procedure yields accurate answers, even for very large order systems, if  $A$  is well-conditioned, and single precision arithmetic is used.

Suppose first that the data  $A$  and  $\mathbf{f}$  in (12) are perturbed by the quantities  $\delta A$  and  $\delta \mathbf{f}$ . Then if the perturbation in the solution  $\mathbf{x}$  of (12) is  $\delta \mathbf{x}$  we have

$$(13) \quad (A + \delta A)(\mathbf{x} + \delta \mathbf{x}) = \mathbf{f} + \delta \mathbf{f}.$$

Now an estimate of the relative change in the solution can be given in terms of the relative changes in  $A$  and  $\mathbf{f}$  by means of

**THEOREM 3.** *Let  $A$  be non-singular and the perturbation  $\delta A$  be so small that*

$$(14) \quad \|\delta A\| < 1/\|A^{-1}\|.$$

*Then if  $\mathbf{x}$  and  $\delta \mathbf{x}$  satisfy (12) and (13) we have*

$$(15) \quad \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\mu}{1 - \mu} \frac{\|\delta \mathbf{f}\|}{\|\delta A\|/\|A\|} \left( \frac{\|\delta \mathbf{f}\|}{\|\mathbf{f}\|} + \frac{\|\delta A\|}{\|A\|} \right),$$

*where the condition number  $\mu$  is defined as*

$$(16) \quad \mu = \mu(A) \equiv \|A^{-1}\| \cdot \|A\|.$$

*Proof.* Since  $\|A^{-1}\delta A\| \leq \|A^{-1}\| \cdot \|\delta A\| < 1$  by (14) it follows from the Corollary to Theorem 1.5 of Chapter 1 that the matrix  $I + A^{-1}\delta A$  is non-singular, and further, that

$$\|(I + A^{-1}\delta A)^{-1}\| \leq \frac{1}{1 - \|A^{-1}\delta A\|} \leq \frac{1}{1 - \|A^{-1}\| \cdot \|\delta A\|}.$$

If we multiply (13) by  $A^{-1}$  on the left, recall (12) and solve for  $\delta x$ , we get

$$\delta x = (I + A^{-1}\delta A)^{-1}A^{-1}(\delta f - \delta Ax).$$

Now take norms of both sides, use the above bound on the inverse, and divide by  $\|x\|$  to obtain

$$\frac{\|\delta x\|}{\|x\|} \leq \frac{\|A^{-1}\|}{1 - \|A^{-1}\| \cdot \|\delta A\|} \left( \frac{\|\delta f\|}{\|x\|} + \|\delta A\| \right).$$

But from (12) it is clear that we may replace  $\|x\|$  on the right, since

$$\|x\| \geq \|f\|/\|A\|,$$

and (15) now easily follows by using the definition (16). ■

The estimate (15) shows that small relative changes in  $f$  and  $A$  cause small relative changes in the solution if the factor

$$\frac{\mu}{1 - \mu \|\delta A\|/\|A\|}$$

is not too large. Of course the condition (14) is equivalent to

$$\mu \frac{\|\delta A\|}{\|A\|} < 1.$$

Thus, it is clear that when the condition number  $\mu(A)$  is not too large, the system (12) is well-conditioned. Note that we cannot expect  $\mu(A)$  to be small compared to unity since

$$\|I\| = \|A^{-1}A\| \leq \mu(A).$$

We can apply Theorem 3 to estimate the effects of roundoff errors committed in solving linear systems by Gaussian elimination and other direct methods. Given any non-singular matrix  $A$ , the condition number  $\mu(A)$  is determined independently of the numerical procedure. But it is possible to view the computed solution as the exact solution, say  $x + \delta x$ , of a perturbed system of the form (13). The basic problem now is to determine the magnitude of the perturbations,  $\delta A$  and  $\delta f$ . This type of approach is called a *backward error analysis*. It is rather clear that there are many perturbations  $\delta A$  and  $\delta f$  which yield the same solution,  $x + \delta x$ , in (13).

Our analysis for the Gaussian elimination method will define  $\delta A$  and  $\delta \mathbf{f}$  so that

$$\delta \mathbf{f} \equiv 0.$$

Then the error estimate (15) becomes simply

$$(17) \quad \frac{\|\delta \mathbf{x}\|}{\|\mathbf{x}\|} \leq \frac{\mu \|\delta A\| / \|A\|}{1 - \mu \|\delta A\| / \|A\|},$$

and it is clear that only  $\|\delta A\|$  in this error bound depends upon the round-off errors and method of computation.

In the case of Gaussian elimination we have seen in Theorem 1, that exact calculations yield the factorization (5b),

$$LU = A.$$

Here  $L$  and  $U$  are, respectively, lower and upper triangular matrices determined by (5c) and (3a). However, with finite precision arithmetic in these evaluations, we do not obtain  $L$  and  $U$  exactly, but say some triangular matrices  $\mathcal{L}$  and  $\mathcal{U}$ . We define the perturbation  $E$  due to these inexact calculations by

$$(18) \quad \mathcal{L}\mathcal{U} \equiv A + E.$$

There are additional rounding errors committed in computing  $\mathbf{g}$  defined by (3b) or (5d), and in the final back substitution (6b) in attempting to compute the solution  $\mathbf{x}$ . With exact calculations, these vectors are defined from (5d) and (6a) as the solutions of

$$L\mathbf{g} = \mathbf{f}, \quad U\mathbf{x} = \mathbf{g}.$$

The vectors actually obtained can be written as  $\mathbf{g} + \delta \mathbf{g}$  and  $\mathbf{x} + \delta \mathbf{x}$  which are the exact solutions of, say,

$$(19a) \quad (\mathcal{L} + \delta \mathcal{L})(\mathbf{g} + \delta \mathbf{g}) = \mathbf{f},$$

$$(19b) \quad (\mathcal{U} + \delta \mathcal{U})(\mathbf{x} + \delta \mathbf{x}) = (\mathbf{g} + \delta \mathbf{g}).$$

Here  $\mathcal{L}$  and  $\mathcal{U}$  account for the fact that the matrices  $L$  and  $U$  are not determined exactly, as in (18). The perturbations  $\delta \mathcal{L}$  and  $\delta \mathcal{U}$  arise from the finite precision arithmetic performed in solving the triangular systems with the coefficients  $\mathcal{L}$  and  $\mathcal{U}$ . Upon multiplying (19b) by  $\mathcal{L} + \delta \mathcal{L}$  and using (19a) we have, from (13) with  $\delta \mathbf{f} = 0$ ,

$$(A + \delta A) = (\mathcal{L} + \delta \mathcal{L})(\mathcal{U} + \delta \mathcal{U}).$$

From (18), it follows then that

$$(20) \quad \delta A = E + \mathcal{L}(\delta \mathcal{U}) + (\delta \mathcal{L})\mathcal{U} + (\delta \mathcal{L})(\delta \mathcal{U}).$$

Thus, to apply our error bound (17) we must estimate  $\|E\|$ ,  $\|\delta \mathcal{U}\|$ , and  $\|\delta \mathcal{L}\|$ . Since  $\mathcal{L}$  and  $\mathcal{U}$  are explicitly determined by the computations, their norms can also, in principle, be obtained.

We shall assume that floating-point arithmetic operations are performed with a  $t$ -digit decimal mantissa (see Section 2 of Chapter 1) and that the system has been ordered so that the natural order of pivots is used [i.e., as in eq. (3)]. In place of the matrices  $A^{(k)} \equiv (a_{ij}^{(k)})$  defined in (2) and (3a), we shall consider the computed matrices  $B^{(k)} \equiv (b_{ij}^{(k)})$  with the final such matrix  $B^{(n)} \equiv \mathcal{U} = (u_{ij})$ , the upper triangular matrix introduced above. Similarly, the lower triangular matrix of computed multipliers  $\mathcal{L} \equiv (s_{ij})$  will replace the matrix  $L \equiv (m_{ij})$  of (5c). For simplicity, we assume that the given matrix elements  $A = (a_{ij})$  can be represented exactly with our floating-decimal numbers.

Now in place of (3a) and (5c), the floating-point calculations yield  $b_{ij}^{(k)}$  and  $s_{ij}$  which by (2.6) of Chapter 1 can be written as:

for  $k = 1$ ,

$$(21a) \quad b_{ij}^{(1)} = a_{ij}, \quad i, j = 1, 2, \dots, n;$$

for  $k = 1, 2, \dots, n - 1$ ,

$$(21b) \quad b_{ij}^{(k+1)} = \begin{cases} b_{ij}^{(k)}, & i \leq k, \\ 0, & i \geq k + 1, j \leq k, \\ [b_{ij}^{(k)} - s_{ik} b_{kj}^{(k)} (1 + \theta_{ij}^{(k)} 10^{-t})] (1 + \phi_{ij}^{(k)} 10^{-t}), & i \geq k + 1, j \geq k + 1; \end{cases}$$

and finally

$$(21c) \quad s_{ij} = \begin{cases} 0, & i < j; \\ 1, & i = j; \\ \frac{b_{ij}^{(j)}}{b_{jj}^{(j)}} (1 + \psi_{ij} 10^{-t}), & i > j. \end{cases}$$

Here the quantities,  $\theta$ ,  $\phi$ ,  $\psi$  satisfy

$$|\theta_{ij}^{(k)}| \leq 5, \quad |\phi_{ij}^{(k)}| \leq 5, \quad |\psi_{ij}| \leq 5,$$

and they account for the rounding procedures in floating-point arithmetic. Of course, the above calculations can be carried out iff the  $b_{jj}^{(j)} \neq 0$  for  $j \leq n - 1$ . However, this can be assured from

**LEMMA 1.** *If  $A$  is non-singular and  $t$  sufficiently large, then the Gaussian elimination method, with maximal pivots and floating-point arithmetic (with  $t$ -digit mantissas), yields multipliers  $s_{ij}$  with  $|s_{ij}| \leq 1$  and pivots  $b_{jj}^{(j)} \neq 0$ .*

*Proof.* See Problem 8. ■

It turns out that we require a bound on the growth of the pivot elements for our error estimates. That is, we seek a quantity  $G = G(n)$ , independent of the  $a_{ij}$ , such that

$$(22a) \quad |b_{jj}^{(j)}| \leq G(n)a, \quad j = 1, 2, \dots, n;$$

where

$$(22b) \quad a \equiv \max_{i,j} |a_{ij}|.$$

Under the conditions of Lemma 1, it is not difficult to see, by induction in (21b), that

$$(22c) \quad G(n) \leq [1 + (1 + 5 \times 10^{-t})^2]^{n-1} = 2^{n-1} + \mathcal{O}((n-1)10^{1-t}).$$

This establishes the existence of a bound of the form (22), but it is a tremendous overestimate for large  $n$ . In fact for exact elimination (i.e., no roundoff errors) using maximal pivots it can be shown that

$$(23a) \quad |a_{jj}^{(j)}| < g(j)a,$$

where

$$(23b) \quad g(j) \leq 3j^{\frac{1}{2} + \frac{1}{4} \ln j}.$$

The quantity  $g(n)$  would be a reasonable estimate for  $G(n)$  if the maximal pivots in the sequence  $\{B^{(k)}\}$  were located in the same positions as the maximal pivots in  $\{A^{(k)}\}$ . We know that if  $A$  is non-singular and  $t$  is sufficiently large, then the indices of the maximal pivotal elements used to find  $\{B^{(k)}\}$  are also indices of maximal pivots in an exact Gaussian elimination procedure for  $A$ . For two special classes of matrices it is established in Problems 6 and 7 that  $g(n) \leq 1$  and  $g(n) \leq n$ . The best (i.e., lowest) bound for  $G(n)$  [or for  $g(n)$ ] is not known at present.

We now turn to estimates of the terms in  $\delta A$ . Our first result is a bound on the elements of  $E$  which we state as

**THEOREM 4.** *Under the hypothesis of Lemma 1 the Gaussian elimination calculations (21) are such that*

$$\mathcal{L}\mathcal{U} = A + E$$

where  $E \equiv (e_{ij})$  satisfies

$$(24) \quad |e_{ij}| \leq \begin{cases} (i-1)2aG(n)10^{1-t}, & \text{for } i \leq j; \\ j2aG(n)10^{1-t}, & \text{for } i > j. \end{cases}$$

Here  $G(n)$  is any bound satisfying (22).

*Proof.* We write the last line of (21b) as

$$(25a) \quad b_{ij}^{(k+1)} = b_{ij}^{(k)} - s_{ik}b_{kj}^{(k)} + \epsilon_{ij}^{(k+1)}, \quad i \geq k+1, j \geq k+1,$$

where from Lemma 1 and (22) it follows that

$$(25b) \quad |\epsilon_{ij}^{(k+1)}| \leq 2aG(n)10^{1-t}, \quad i \geq k+1, j \geq k+1.$$

Similarly, multiplying the last line of (21c) by  $b_{jj}^{(j)}$  and dividing by  $1 + \psi_{ij}10^{-t}$ , we can write the result as

$$(26a) \quad 0 = b_{ij}^{(j)} - s_{ij}b_{jj}^{(j)} + \epsilon_{ij}^{(j+1)}, \quad i > j;$$

where again we find that

$$(26b) \quad |\epsilon_{ij}^{(j+1)}| \leq 2aG(n)10^{1-t}, \quad i > j.$$

Upon recalling (21) we have

$$\mathcal{L} \equiv (s_{ij}), \quad \mathcal{U} \equiv (b_{ij}^{(i)})$$

so that

$$\begin{aligned} (\mathcal{L}\mathcal{U})_{ij} &= \sum_{k=1}^n s_{ik}b_{kj}^{(k)} \\ &= \sum_{k=1}^{\min(i,j)} s_{ik}b_{kj}^{(k)}. \end{aligned}$$

Now let  $i > j$  and sum (25a) over  $k = 1, 2, \dots, j-1$  and then add (26a) to get, with the aid of (21),

$$0 = b_{ij}^{(1)} - \sum_{k=1}^j s_{ik}b_{kj}^{(k)} + \sum_{k=1}^j \epsilon_{ij}^{(k+1)}, \quad i > j.$$

From the last two equations above and the fact that

$$b_{ij}^{(1)} = a_{ij}$$

we see that the elements  $e_{ij}$  with  $i > j$  of  $E = \mathcal{L}\mathcal{U} - A$  are

$$(27a) \quad e_{ij} = \sum_{k=1}^j \epsilon_{ij}^{(k+1)}, \quad i > j.$$

For the elements with  $i \leq j$ , we just sum (25a) over  $k = 1, 2, \dots, i-1$ , recalling that  $s_{ii} = 1$ , and obtain as before

$$(27b) \quad e_{ij} = \sum_{k=1}^{i-1} \epsilon_{ij}^{(k+1)}, \quad i \leq j.$$

But now (24) follows from (27) by using the bounds (25b) and (26b). ■

As a simple corollary of this theorem, we note that since  $|e_{ij}| \leq 2(n-1)aG(n)10^{1-t}$ , it follows that

$$(28) \quad \|E\|_\infty \leq 2an(n-1)G(n)10^{1-t}.$$

The elements in  $\delta\mathcal{L}$  and  $\delta\mathcal{U}$  can be estimated from a single analysis of the error in solving any triangular system (with the same arithmetic and rounding). Thus we consider, say,

$$(29a) \quad T\mathbf{u} = \mathbf{h}$$

where  $T \equiv (t_{ij})$  is lower triangular and non-singular, i.e.,

$$(29b) \quad t_{ii} \neq 0; \quad t_{ij} = 0, \quad j > i; \quad i = 1, 2, \dots, n.$$

The exact solution of (29) is easily obtained by recursion and is

$$(30a) \quad u_1 = t_{11}^{-1}h_1$$

$$(30b) \quad u_i = t_{ii}^{-1} \left( h_i - \sum_{k=1}^{i-1} t_{ik}u_k \right), \quad i = 2, 3, \dots, n.$$

For numerical solutions, we have

**THEOREM 5.** *Let the “solution” of (29) be computed by using  $t$ -digit floating-point arithmetic to evaluate (30). Then the computed solution, say  $\mathbf{v}$ , satisfies*

$$(31a) \quad (T + \delta T)\mathbf{v} = \mathbf{h}$$

where the perturbations are bounded by

$$(31b) \quad |\delta t_{ij}| \leq \max [2, |i - j + 1|] |t_{ij}| 10^{1-t} \leq n |t_{ij}| 10^{1-t}.$$

Here  $t$  is required to be so large that  $n10^{1-t} \leq 1$ .

*Proof.* In the notation of Section 2 of Chapter 1 the floating-decimal evaluations,  $v_i$ , of the formulas (30) are

$$\begin{aligned} v_1 &= \text{fl}\left(\frac{h_1}{t_{11}}\right); \\ v_i &= \text{fl}\left[\frac{\left(h_i - \sum_{k=1}^{i-1} t_{ik}v_k\right)}{t_{ii}}\right] \\ &= \text{fl}\left[\frac{\text{fl}\left(h_i - \sum_{k=1}^{i-1} t_{ik}v_k\right)}{t_{ii}}\right], \quad i = 2, 3, \dots, n. \end{aligned}$$

Then by using (2.6c) of Chapter 1, we must have from the above

$$(32a) \quad v_1 = \frac{h_1}{t_{11}} (1 + \phi_1 10^{-t}), \quad |\phi_1| \leq 5;$$

$$(32b) \quad v_i = \frac{\text{fl}\left(h_i - \sum_{k=1}^{i-1} t_{ik}v_k\right)}{t_{ii}} (1 + \phi_i 10^{-t}), \quad |\phi_i| \leq 5, i = 2, 3, \dots, n.$$

If, in the floating-point evaluation of the numerator in (32b) the sum is first accumulated and then subtracted from  $h_i$ , we can write, with the use of (2.6a) and Lemma 2.2 of Chapter 1,

$$\text{fl}\left(h_i - \sum_{k=1}^{i-1} t_{ik}v_k\right) = \left[h_i - \sum_{k=1}^{i-1} (t_{ik} + \delta t_{ik})v_k\right](1 + \theta_i 10^{-t}), \\ i = 2, 3, \dots, n.$$

Here  $|\theta_i| \leq 5$  and

$$|\delta t_{ik}| \leq \begin{cases} (i-k+1)|t_{ik}|10^{1-t}, & 2 \leq k \leq i-1, \\ 2|t_{i,i-1}|10^{1-t}, & k = 1, i = 2, 3, \dots, n. \end{cases}$$

From the above and (32) we now obtain, solving for the  $h_j$ ,

$$t_{11}v_1(1 + \phi_1 10^{-t})^{-1} = h_1, \\ t_{ii}v_i(1 + \phi_i 10^{-t})^{-1}(1 + \theta_i 10^{-t})^{-1} + \sum_{k=1}^{i-1} (t_{ik} + \delta t_{ik})v_k = h_i, \\ i = 2, 3, \dots, n.$$

However, if we write

$$t_{11} + \delta t_{11} = t_{11}(1 + \phi_1 10^{-t})^{-1} \\ t_{ii} + \delta t_{ii} = t_{ii}(1 + \phi_i 10^{-t})^{-1}(1 + \theta_i 10^{-t})^{-1}$$

then it follows from  $|\phi_i| \leq 5$ ,  $|\theta_i| \leq 5$ , and  $n10^{1-t} \leq 1$  that

$$|\delta t_{11}| \leq |t_{11}|10^{1-t}, \\ |\delta t_{ii}| \leq 2|t_{ii}|10^{1-t}, \quad i > 1. \quad \blacksquare$$

We are now able to obtain estimates of the elements in  $\delta\mathcal{L}$  and  $\delta\mathcal{U}$  or more importantly those in  $\delta A$ . These results are contained in the basic

**THEOREM 6 (WILKINSON).** *Let the  $n$ th order matrix  $A$  be non-singular and employ Gaussian elimination with maximal pivots and  $t$ -digit floating-point arithmetic to solve (12). Let  $t$  be so large that Lemma 1 applies and that*

$$n10^{1-t} \leq 1.$$

*Then the computed solution, say  $\mathbf{x} + \delta\mathbf{x}$ , satisfies*

$$(A + \delta A)(\mathbf{x} + \delta\mathbf{x}) = \mathbf{f}$$

where

$$(33) \quad |\delta a_{ij}| \leq (2n + 3n^2)G(n)a10^{1-t}$$

and  $G(n)$  is any bound satisfying (22).

*Proof.* We have already shown that  $\delta A$  is given by (20) where the elements of  $E$  are estimated in (24). Since  $\delta \mathcal{L}$  is the perturbation (19a) in solving the lower triangular system  $\mathcal{L}\mathbf{g} = \mathbf{f}$ , we can apply Theorem 5. We note that the elements of  $\mathcal{L} \equiv (s_{ij})$  are given by (21c). Since maximal pivots are used  $|b_{jj}^{(j)}| \geq |b_{ij}^{(j)}|$  so that  $|s_{ij}| \leq 1$  and we easily get from (31b) in this case

$$|(\delta \mathcal{L})_{ij}| \equiv |\delta s_{ij}| \leq n10^{1-t}.$$

The elements of  $\delta \mathcal{U}$  are the perturbations in solving a system of the form  $\mathcal{U}\mathbf{y} = \mathbf{z}$  with  $\mathcal{U} \equiv (u_{ij}) \equiv (b_{ij}^{(i)})$  where the  $b_{ij}^{(i)}$  are defined by (21b). This system is, of course, upper triangular but the estimates of Theorem 5 still apply. Since maximal pivots are used we have, recalling (22),

$$|u_{ij}| \equiv |b_{ij}^{(i)}| \leq |b_{ii}^{(i)}| \leq G(n)a, \quad i \leq j, \quad i = 1, 2, \dots, n;$$

and now (31b) yields

$$|(\delta \mathcal{U})_{ij}| \equiv |\delta u_{ij}| \leq nG(n)a10^{1-t}.$$

From (20) we have

$$\delta a_{ij} = e_{ij} + \sum_{k=1}^{\min(i,j)} (s_{ik} \delta u_{kj} + \delta s_{ik} u_{kj} + \delta s_{ik} \delta u_{kj}).$$

By taking absolute values and using the above bounds on  $|\delta u_{ij}|$ ,  $|\delta s_{ij}|$ ,  $|u_{ij}|$ ,  $|s_{ij}|$ , as well as (24), we easily obtain

$$|\delta a_{ij}| \leq (2n + 2n^2 + n^3 10^{1-t}) G(n)a10^{1-t}.$$

However, since it was required that  $n10^{1-t} < 1$ , the result in (33) follows. ■

From this theorem, it follows that the computed solution is the exact solution of a system only slightly perturbed from the original if enough figures are used, i.e.,  $t$  sufficiently large. Appropriate values for  $t$  depend upon  $n$  and the bound  $G(n)$ . If, as indicated in (22c),  $G(n)$  were of the order  $2^{n-1}$ , only relatively small order systems could be treated effectively. On the other hand, if  $G(n) \cong g(n) \leq 3n^{1/2 + 1/4 \ln n}$ , as in (23), then quite large order systems can be treated with the number of digits used on modern digital computers (say  $t \geq 8$ ). It is generally believed, however, that even this latter estimate for  $G(n)$  is a generous overestimate, when using maximal pivots. It should be observed that essentially all of the previous analysis is valid if only partial pivoting, say maximal column pivoting, is employed since then  $|s_{ij}| \leq 1$  is maintained. However, the growth factor  $G(n)$  for this procedure cannot be estimated well in general. In fact, it is possible that the upper bound (22c) which still applies may be

attained. In spite of this, partial pivoting is found to be effective in practice but the absence of *any* type of maximal pivoting strategy frequently leads to catastrophic growth of rounding errors.†

From (33) we easily find that

$$(34) \quad \|\delta A\|_\infty \leq (2n^2 + 3n^3)G(n)a10^{1-t},$$

and this can be employed in (17) to obtain maximum norm bounds on the relative error. It is clear that this relative error in  $\mathbf{x}$  may not be small even though the relative perturbation,  $\|\delta A\|/\|A\|$ , is small. In such a case,  $A$  would be ill-conditioned. By (17), the relative error  $\|\delta \mathbf{x}\|/\|\mathbf{x}\|$  is small if  $\|A^{-1}\| \|\delta A\|$  is small. For a given  $G(n)$  and  $\mu(A)$  equation (34) may be used to find the value of  $t$  that assures a solution with a prescribed accuracy.

Finally, we recall that a computed inverse of  $A$  can be obtained by solving the  $n$  systems (0.6). If we denote the matrix obtained as  $A^{-1} + F$ , then as above we can show that each column vector of  $A^{-1} + F$ , i.e.,  $(A^{-1} + F)_j$ , satisfies an equation of the form, for some perturbation matrix  $\delta A^{(j)}$ ,

$$(35a) \quad (A + \delta A^{(j)})(A^{-1} + F)_j = \mathbf{e}^{(j)}, \quad j = 1, 2, \dots, n.$$

Under the assumptions of Theorem 6, the estimates (33) and (34) also apply to the current perturbations,  $\delta A^{(j)}$ . Then, if  $\|\delta A^{(j)}\| \leq 1/\|A^{-1}\|$  we have, almost as in the proof of Theorem 3,

$$(35b) \quad \frac{\|F_j\|}{\|A_j^{-1}\|} \leq \frac{\mu(A)\|\delta A^{(j)}\|/\|A\|}{1 - \mu(A)\|\delta A^{(j)}\|/\|A\|}.$$

Thus, as was to be expected, the columns of the inverse matrix are obtained to within the same relative error [i.e., compare (17)] as is the solution of any particular system.

### 1.3. A Posteriori Error Estimates

Although we do not advocate inverting a matrix to solve linear systems, it is of interest to consider error estimates related to computed inverses.

† Experience indicates that we usually achieve greater accuracy in the single precision solution, if we first *scale* the matrix  $A$ . That is, if with  $B = D_1AD_2$ , we solve

$$By = D_1f$$

for  $\mathbf{y}$ ; and then determine  $\mathbf{x}$  from  $D_2\mathbf{y} = \mathbf{x}$ . Here  $D_1$  and  $D_2$  are some diagonal matrices chosen so that the  $n$  columns and the  $n$  rows of the matrix  $B$  have approximately equal norms. A complete mathematical explanation of this phenomenon is not available.

Let  $A$  be the matrix to be inverted and let  $C$  be the computed or alleged inverse. The error in the inverse is defined by

$$(36a) \quad F = C - A^{-1};$$

we also use another measure of error called the *residual matrix*:

$$(36b) \quad R = AC - I.$$

We have first

**THEOREM 7.** *If  $\|R\| < 1$  then:*

$$(37a) \quad A \text{ and } C \text{ are non-singular};$$

$$(37b) \quad \|A^{-1}\| \leq \|C\|(1 - \|R\|);$$

$$(37c) \quad \|F\| \leq \|C\| \cdot \|R\|(1 - \|R\|).$$

*Proof.* We write (36b) as

$$AC = I + R,$$

and use the corollary to Theorem 1.5 of Chapter 1 and  $\|R\| < 1$  to deduce that  $AC$  is non-singular. Part (37a) then follows. Take the inverse of both sides in the above equation and multiply on the left by  $C$  to find

$$A^{-1} = C(I + R)^{-1}.$$

Now (37b) follows by taking norms and by again using the corollary to Theorem 1.5 of Chapter 1. From (36) we see that  $F = A^{-1}R$  and so,  $\|F\| \leq \|A^{-1}\| \cdot \|R\|$ , and (c) follows by an application of (b). ■

Note that we may just as well consider  $A$  to be an approximation to the inverse of  $C$ . Thus we obtain the

**COROLLARY.** *Under the hypothesis of Theorem 7,*

$$(37d) \quad \|C^{-1}\| \leq \|A\|(1 - \|R\|),$$

$$(37e) \quad \|A - C^{-1}\| \leq \|A\| \cdot \|R\|(1 - \|R\|). \quad \blacksquare$$

Since  $A$  and  $C$  are presumed known, we could actually compute  $\|C\|$ ,  $\|A\|$ , and  $\|R\|$  in the estimates (37). This, of course, is what is meant by a posteriori estimates. In general,  $n^3$  multiplications are required to form  $AC$  and this computation, as well as that of the norms, is subject to simply estimated roundoff errors. In contrast, the quantity  $\delta A$  entering the a priori estimates (17) and (34) cannot be computed. It is hardly necessary to point out that  $G(n)$  is determined easily after the elimination process has been completed.

It is of interest to note that, under the hypothesis of Theorem 7, with  $C$  an approximate inverse of  $A$  we can find the perturbation  $\delta A$ , so that  $C$  is the exact inverse of  $A + \delta A$ . That is, set

$$(A + \delta A)C = I.$$

Hence,

$$\begin{aligned} -\delta A &= (AC - I)C^{-1} \\ &= RC^{-1}. \end{aligned}$$

Upon taking norms and using (37d), we then have

$$(38) \quad \|\delta A\| \leq \frac{\|R\| \cdot \|A\|}{1 - \|R\|}.$$

Finally, we observe that the computed inverse can also be used to estimate the error in solving a linear system. We state this result as

**THEOREM 8.** *Let an approximate solution  $\mathbf{y}$  of*

$$A\mathbf{x} = \mathbf{f}$$

*have the residual vector*

$$(39) \quad \mathbf{r} = A\mathbf{y} - \mathbf{f}.$$

*Then, if an approximate inverse  $C$  of  $A$  satisfies  $\|R\| \equiv \|AC - I\| < 1$ , we have*

$$(40) \quad \|\mathbf{y} - \mathbf{x}\| \leq \frac{\|\mathbf{r}\| \cdot \|C\|}{1 - \|R\|}.$$

*Proof.* From Theorem 7 it follows that  $A$  is non-singular and so from (39)

$$\mathbf{y} = A^{-1}(\mathbf{r} + \mathbf{f}).$$

Subtract  $\mathbf{x} = A^{-1}\mathbf{f}$  from this to find, after taking norms,

$$\|\mathbf{y} - \mathbf{x}\| \leq \|A^{-1}\| \cdot \|\mathbf{r}\|.$$

The result (40) then follows from (37b). ■

The determination of the *residual vector*  $\mathbf{r}$  is the first step in an iterative procedure to improve upon the accuracy of the solution (see Subsection 4.3).

It should be noted that the result in this theorem is independent of the manner in which the approximate solution,  $\mathbf{y}$ , is obtained. Thus, it was not assumed that  $\mathbf{y} = Cf$ . This suggests, in fact, that the sole purpose for computing  $C$  and  $R$  might be to use them in error estimates of the form (40). That is, once the constant  $M \equiv \|C\|/(1 - \|R\|)$  is known, it requires

only  $n^2$  multiplications to compute  $\mathbf{r}$  for each approximate solution  $\mathbf{y}$  of a system with coefficient matrix  $A$ . If one wished to use (40), after finding  $\mathbf{y}$  by Gaussian elimination, then an approximate inverse  $C$  could be obtained by using the approximate factorization of  $A$ . This, by (8a) and (10), would require twice as much labor as was already expended to find  $\mathbf{y}$ .

## PROBLEMS, SECTION 1

1. Show that  $A^{(k)}$  is non-singular iff  $A$  is non-singular, for the Gaussian elimination method.
2. Describe how the maximal pivot scheme permits the completion of the elimination method, when  $A$  is singular.
3. Prove the following corollary to Theorem 2: If interchanges of rows and of columns are made and  $r = n$ , then

$$\det A = (-1)^j a_{i_1 j_1}^{(1)} a_{i_2 j_2}^{(2)} \dots a_{i_n j_n}^{(n)},$$

where  $a_{i_k j_k}^{(k)}$  are the successive pivotal elements in the Gaussian elimination scheme and  $(-1)^j = \det P \det Q$ .

4. If  $A$  is symmetric and positive definite (that is,  $\mathbf{x}^* A \mathbf{x} \geq 0$  and  $\sum_{i,j=1}^n a_{ij} x_i x_j = 0$  only if  $x_i = 0$  for all  $i$ ;  $a_{ij}$  and  $x_i$  real), show that
  - (a)  $a_{ii} > 0$
  - (b)  $\max_i a_{ii} = \max_{i,j} |a_{ij}|$ . [Hint: For (b), if  $|a_{rs}| = \max_{i,j} |a_{ij}|$ , then with  $x_i = 0$  for  $i \neq r, s$

$$\sum_{i,j=1}^n a_{ij} x_i x_j \equiv a_{rr} x_r^2 + 2a_{rs} x_r x_s + a_{ss} x_s^2 = 0$$

for non-trivial  $x_r, x_s$  if  $a_{rr} a_{ss} < a_{rs}^2$ .]

5. If  $A$  is symmetric, positive definite, then the submatrices  $(a_{ij}^{(k)})$  for  $k \leq i, j \leq n$  are symmetric, positive definite. [Hint: Use mathematical induction on  $k$ . Symmetry from

$$a_{ij}^{(2)} = a_{ii}^{(1)} - \frac{a_{i1}^{(1)}}{a_{11}^{(1)}} a_{1j}^{(1)}.$$

Positive definiteness from Problem (4a) and

$$\sum_{i,j=2}^n a_{ij}^{(2)} x_i x_j \equiv \sum_{i,j=1}^n a_{ij}^{(1)} x_i x_j - a_{11}^{(1)} \left[ x_1 + \sum_{i=2}^n \frac{a_{j1}^{(1)}}{a_{11}^{(1)}} x_i \right]^2.$$

That is, if  $(a_{ij}^{(2)})$  is not positive definite, then  $(a_{ij}^{(1)})$  is not.]

6. (Von Neumann-Goldstine.) If  $A$  is symmetric, positive definite, then  $a_{ii}^{(k)} \leq a_{ii}^{(k-1)}$  for  $k \leq i \leq n$ ,  $k = 2, 3, \dots, n$ . (Hence by Problem (4b),  $\max_{i,j} |a_{ij}^{(n)}| \leq \max_{i,j} |a_{ij}|$ .)

7. (Wilkinson.) If  $A$  is a *Hessenberg matrix* (i.e.,  $a_{ij} = 0$  for  $i \geq j + 2$ ),

$$\max_{i,j} |a_{ij}^{(n)}| \leq n \max_{i,j} |a_{ij}|,$$

if maximal column pivots are used. [Hint: Only one row is changed in passing from  $A^{(k-1)}$  to  $A^{(k)}$ .]

**8. Prove Lemma 1.**

**9.** If  $A$  is symmetric, use the first part of Problem 5 to show: the number of operations to solve (1) by Gaussian elimination, with diagonal pivots, is  $n^3/6 + \mathcal{O}(n^2)$ .

**10.** If in (1),  $A$  and  $\mathbf{f}$  are complex, show that (1) may be converted to the solution of a real system of order  $2n$ .

## 2. VARIANTS OF GAUSSIAN ELIMINATION

There are many methods for solving linear systems that are slight variations of the Gaussian elimination method. None of these methods has succeeded in reducing the number of operations required, but some have eliminated much of the intermediate storage or recording requirements. Caution should be taken in applying any variation that does not allow for the selection of some sort of maximal pivots, which is generally necessary to prevent the growth of rounding errors.

The modification due to Jordan circumvents the final back substitution. This is accomplished by additional computations which serve to eliminate the variable  $x_k$  from the first  $k - 1$  equations as well as from the last  $n - k$  equations at the  $k$ th stage of the reduction. In other words, the coefficients above the diagonal are also reduced to zero and the final coefficient matrix which results is a diagonal matrix. The obvious modifications which are required for this Gauss-Jordan elimination are contained in

$$(1a) \quad a_{ij}^{(1)} = a_{ij}$$

$$(1b) \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} a_{k-1,j}^{(k-1)} \quad \text{for } i \neq k-1, j \geq k-1$$

$$(1c) \quad a_{k-1,j}^{(k)} = a_{k-1,j}^{(k-1)} \quad \text{for } j \geq k-1$$

$$(1d) \quad a_{ij}^{(k)} = a_{ij}^{(k-1)} \quad \text{for } j < k-1.$$

$$(2a) \quad f_i^{(1)} = f_i$$

$$(2b) \quad f_i^{(k)} = f_i^{(k-1)} - \frac{a_{i,k-1}^{(k-1)}}{a_{k-1,k-1}^{(k-1)}} f_{k-1}^{(k-1)} \quad \text{for } i \neq k-1$$

$$(2c) \quad f_{k-1}^{(k)} = f_{k-1}^{(k-1)} \quad \text{for } k = 2, \dots, n.$$

The solution is then

$$x_i = \frac{f_i^{(n)}}{a_{ii}^{(n)}} = \frac{f_i^{(n)}}{a_{ii}^{(1)}}.$$

It is clear that pivoting on the maximal element in the remaining square submatrix may be retained in this procedure. Hence, multipliers for

$i < k - 1$  may exceed unity. Furthermore, the number of operations is somewhat greater than in the ordinary Gaussian elimination with back substitution; it is now

$$(3) \quad \frac{n^3}{2} + n^2 - \frac{n}{2} \text{ ops.}$$

Thus, there does not seem to be any great advantage in using the Gauss-Jordan elimination in actual calculations with automatic computing equipment.

Another variation is the so-called *Crout reduction*. This method is applicable if the rows and columns are so arranged that *no column interchanges are required* in the Gaussian elimination (as in the case of symmetric, positive definite matrices; see Theorem 3.3). Thus, in general, the pivots will not be the maximal elements. Hence, errors may grow very rapidly in the Crout method and it is not recommended unless the system is of relatively small order or if it can be determined that the error growth will not be catastrophic. (In practice one may apply the method and test the accuracy of the solution *a posteriori*.) On the other hand, the Crout method is specifically designed to reduce the number of intermediate quantities which must be retained. Thus, for hand computations and digital computers with small storage capacities it may be of great value. The Crout method may be modified to use maximal column pivots, by incorporating row interchanges as described in Theorem 3.1 (or see Theorem 1.2).

This “compact” elimination procedure is based on the fact that only those elements  $a_{ij}^{(k)}$ , in the Gaussian elimination, for which  $j \geq i$  and  $i \leq k$ , are required for the final back substitution.

Thus, we seek a recursive method of defining the columns of  $L$  (lower triangular matrix of multipliers) and rows of  $U$  (upper triangular matrix). From Theorem 1.1, we know that

$$LU = A.$$

Hence, let us write the formula for  $a_{kj}$  from the rule for matrix multiplication, after a simple algebraic transformation, in the form

$$(4) \quad u_{kj} = a_{kj} - \sum_{p=1}^{k-1} m_{kp} u_{pj}, \quad \text{if } k \leq j;$$

and the formula for  $a_{ik}$  in the form

$$(5) \quad m_{ik} = \frac{1}{u_{kk}} \left[ a_{ik} - \sum_{p=1}^{k-1} m_{ip} u_{pk} \right], \quad \text{if } i > k.$$

We now may use (4) and (5) for  $k \geq 2$  to find first the elements of the  $k$ th row of  $U$  and then the elements of the  $k$ th column of  $L$ , provided that we know the previous rows and columns, respectively, of  $U$  and  $L$ . Hence, we need only define the first row

$$(4)_1 \quad u_{1j} = a_{1j} \quad \text{for } 1 \leq j \leq n$$

and first column

$$(5)_1 \quad m_{i1} = \frac{a_{i1}}{a_{11}} \quad \text{for } 2 \leq i \leq n.$$

If we define  $u_{1,n+1} = f_1$ ;  $a_{i,n+1} = f_i$  and use (4) for  $j = n + 1$ , we find a column  $(u_{i,n+1})$  which is the vector  $\mathbf{g}$  of Theorem 1.1.

We then use the back substitution to solve  $U\mathbf{x} = \mathbf{g}$  as before [where  $U$  represents the first  $n$  columns of  $(u_{ij})$ ]. The operational count, for producing  $L$ ,  $U$ , and  $\mathbf{g}$ , is easily found to be  $(2n^3 + 3n^2 - 5n)/6$ . It is not surprising that this is the same as the number of operations required by the conventional Gaussian elimination scheme to produce  $L$ ,  $U$ , and  $\mathbf{g}$  (since we merely avoid writing down the intermediate elements but have ultimately to do the same multiplications and divisions).

We could show now, if the inner products in (4) and (5) are accumulated in *double precision* before the sum is rounded, that the effect of rounding errors is appreciably diminished. In fact, the estimate in place of (1.34) becomes

$$\|\delta A\|_\infty = \mathcal{O}(n^2 G(n) \alpha 10^{-t}).$$

## PROBLEM, SECTION 2

- Verify the operational count for the Crout method:  $(2n^3 + 3n^2 - 5n)/6$ .

## 3. DIRECT FACTORIZATION METHODS

The final forms, (2.4) and (2.5), that are used in the Crout method, suggest a more general study of the direct triangular decomposition

$$(1) \quad LU = A,$$

in which the diagonal elements of  $L$  are not necessarily unity. In fact, if we consider  $L \equiv (l_{ij})$  then (1) implies

$$(2a) \quad l_{kk}u_{kk} = a_{kk} - \sum_{p=1}^{k-1} l_{kp}u_{pk}, \quad \text{for } k \geq 2$$

$$(2b) \quad u_{kj} = \frac{1}{l_{kk}} \left( a_{kj} - \sum_{p=1}^{k-1} l_{kp} u_{pj} \right), \quad \text{for } j > k \geq 2$$

$$(2c) \quad l_{ik} = \frac{1}{u_{kk}} \left( a_{ik} - \sum_{p=1}^{k-1} l_{ip} u_{pk} \right), \quad \text{for } i > k \geq 2.$$

[Equations (2a, b, c) hold for  $k = 1$ , if we remove the  $\sum$  term.] Equation (2a) determines the product  $l_{kk}u_{kk}$  in terms of data in previous rows of  $U$  and columns of  $L$ . Once  $l_{kk}$  and  $u_{kk}$  are chosen to satisfy (2a), we then use (2b) and (2c) to determine the remaining elements in the  $k$ th row and column.

If  $l_{kk}u_{kk} = 0$ , the factorization is not possible, unless all of the brackets vanish in (2b), for  $j > k$ , or all of the brackets vanish in (2c), for  $i > k$ . If  $A$  is non-singular, then the use of maximal column pivots results in the sequence  $(i_1, 1), (i_2, 2), \dots, (i_n, n)$  as pivotal elements. Hence, the Gaussian elimination process shows that the triangular decomposition

$$LU = P^T A,$$

is possible, where  $P$  is defined in Theorem 1.2. In fact, if  $A$  is non-singular, one of the bracketed expressions in (2c) does not vanish, for some  $i \geq k$ . Therefore, one of the bracketed expressions in (2c) is of maximum absolute value for  $i \geq k$ , say for  $i = i_k \geq k$ . We may then move the elements of the row  $i_k$  in both  $A$  and in the part of  $L$  that has already been found up to row  $k$ . (The rows  $k, k+1, \dots, i_k - 1$  are moved down in both  $L$  and  $A$  to fill the gap.) Hence, if  $A$  is non-singular we may, with row interchanges, employ (2a, b, and c) to achieve a triangular factorization. We summarize these facts in

**THEOREM 1.** *If  $A$  is non-singular, a triangular decomposition,  $LU = A$ , may not be possible. But a permutation of the rows of  $A$  can be found such that  $B \equiv P^T A = LU$ , where  $P = (p_{rs})$  and*

$$p_{rs} = \begin{cases} 0, & r \neq i_s \\ 1, & r = i_s. \end{cases}$$

*In fact, the  $P$  may be found so that*

$$|l_{kk}| \geq |l_{ik}| \quad \text{for } i > k; \quad k = 1, 2, \dots, n-1. \quad \blacksquare$$

Note that in this result, in contrast to that in Theorem 1.2, we have only employed row interchanges.

A symmetric choice  $l_{kk} = u_{kk}$  may lead to imaginary numbers, if the right-hand side of (2a) is negative; a less symmetric choice  $l_{kk} = |u_{kk}|$  keeps the arithmetic real if  $A$  is real (see Problem 1).

As in the Crout method, we may consider  $\mathbf{f}$  as an additional column of  $A$  (i.e.,  $a_{i,n+1} \equiv f_i$ ) and use (2b) for  $j = n + 1$  to define the elements  $g_i \equiv u_{i,n+1}$  such that

$$U\mathbf{x} = L^{-1}\mathbf{f} = \mathbf{g}.$$

In Subsections 3.1, 3.2, and 3.3, we consider special applications of this procedure.

### 3.1. Symmetric Matrices (Cholesky Method)

We begin with

**THEOREM 2.** *Let  $A$  be symmetric. If the factorization  $LU = A$  is possible, then the choice  $l_{kk} = u_{kk}$  implies  $l_{ik} = u_{ki}$ , that is,  $LL^T = A$ .*

*Proof.* Use (2) and induction on  $k$ . ■

A simple, non-singular, symmetric matrix for which the factorization is not possible is

$$\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

On the other hand, if the symmetric matrix  $A$  is positive definite (i.e.,  $\mathbf{x}^*A\mathbf{x} > 0$  if  $\mathbf{x}^*\mathbf{x} > 0$ ), then the factorization is possible. We have

**THEOREM 3.** *Let  $A$  be symmetric, positive definite. Then,  $A$  can be factored in the form*

$$LL^T = A.$$

*Proof.* Problems 4 and 5 of Section 1 show that the Gaussian elimination method can be carried out, without any interchanges, to give the factorization  $(m_{ij})(b_{ij}) = A$ , where  $b_{ii} > 0$ . But if we define

$$l_{kk} = u_{kk} = \sqrt{b_{kk}},$$

then by Problem 1, we will obtain from (2b, c) the elements in the factorization

$$LU = A,$$

where

$$l_{ik} = u_{ki}. \quad \blacksquare$$

A count of the arithmetic operations can be made if we remark that only the elements  $l_{ik}$  defined by (2a, c) are involved. If we count the square root operations separately, we have

$$\frac{n^3}{6} + n^2 - \frac{n}{6} \text{ ops.} + n \text{ square roots} = \text{no. of ops. to find } L \text{ and } g.$$

In addition, to find  $\mathbf{x}$  we must solve a triangular system which requires  $(n^2 + n)/2$  operations. Thus, *to solve one system using the Cholesky method requires  $n^3/6 + 3n^2/2 + n/3$  operation plus  $n$  square roots.*

To apply our previous error analysis, we deduce from

$$a_{ii} = \sum_{k=1}^i l_{ik} u_{ki} = \sum_{k=1}^i |l_{ik}|^2,$$

the bounds

$$|l_{ik}|^2 \leq a_{ii} \leq a = \max_{i,j} |a_{ij}|.$$

For single precision square roots and  $G(n) = 1$ , we could prove (as we do in Theorem 1.6)

**THEOREM 4.** *If  $A$  is symmetric and positive definite, then the approximate solution of  $A\mathbf{x} = \mathbf{f}$  obtained by factorization and floating-point arithmetic with  $t$  digits satisfies*

$$(A + \delta A)\mathbf{y} = \mathbf{f},$$

where for  $t$  sufficiently large

$$\|\delta A\|_\infty \leq a 10^{1-t} (2n^2 + 3n^3). \quad \blacksquare$$

**COROLLARY.** *Under the hypothesis of Theorem 4, if inner products are accumulated exactly, prior to a final rounding, then for  $t$  sufficiently large*

$$\|\delta A\|_\infty = \mathcal{O}(n^2 a 10^{-t}). \quad \blacksquare$$

### 3.2. Tridiagonal or Jacobi Matrices

A coefficient matrix which frequently occurs is the tridiagonal or Jacobi form, in which  $a_{ij} = 0$  if  $|i - j| > 1$ . That is,

$$(3) \quad A = \begin{bmatrix} a_1 & c_1 & & & & & \\ b_2 & a_2 & c_2 & & & & \\ . & . & . & . & & & \\ & . & . & . & . & & \\ & & & & . & & \\ & & & & & b_{n-1} & a_{n-1} & c_{n-1} \\ & & & & & b_n & a_n & & \end{bmatrix}.$$

Assume this matrix can be factored in the bidiagonal form

$$A = LU \equiv \begin{bmatrix} \alpha_1 & & & \\ b_2 & \alpha_2 & & \\ b_3 & \alpha_3 & & \\ \vdots & \ddots & & \\ & \ddots & \ddots & \\ & & b_n & \alpha_n \end{bmatrix} \begin{bmatrix} 1 & \gamma_1 & & & \\ & 1 & \gamma_2 & & \\ & & 1 & \cdot & \\ & & & \ddots & \cdot \\ & & & & \gamma_{n-1} \\ & & & & 1 \end{bmatrix}.$$

Then we find,

$$(4a) \quad \alpha_1 = a_1, \quad \gamma_1 = c_1/\alpha_1;$$

$$(4b) \quad \alpha_i = a_i - b_i \gamma_{i-1}, \quad i = 2, 3, \dots, n;$$

$$(4c) \quad \gamma_i = c_i/\alpha_i, \quad i = 2, 3, \dots, n-1.$$

Thus, if none of the  $\alpha_i$  vanish, the factorization is accomplished by evaluating the recursions in (4). The “intermediate” solution  $\mathbf{g}$  of  $L\mathbf{g} = \mathbf{f}$  becomes

$$(5a) \quad g_1 = f_1/\alpha_1;$$

$$(5b) \quad g_i = (f_i - b_i g_{i-1})/\alpha_i, \quad i = 2, \dots, n;$$

and the final solution  $\mathbf{x}$  of  $U\mathbf{x} = \mathbf{g}$  is given by

$$(6a) \quad x_n = g_n,$$

$$(6b) \quad x_j = g_j - \gamma_j x_{j+1}, \quad j = n-1, n-2, \dots, 1.$$

In many of the applications of this procedure, the elements (3) of  $A$  satisfy

$$(7a) \quad |a_1| > |c_1| > 0;$$

$$(7b) \quad |a_i| \geq |b_i| + |c_i|, \quad b_i c_i \neq 0, \quad i = 2, 3, \dots, n-1;$$

$$(7c) \quad |a_n| > |b_n| > 0.$$

In such cases, the quantities  $\alpha_i$  and  $\gamma_i$  can be shown to be nicely bounded and in fact  $A$  is non-singular. We state this as

**THEOREM 5.** *If the elements of  $A$  in (3) satisfy (7) then  $\det A \neq 0$  and the quantities in (4) are bounded by*

$$(a) \quad |\gamma_i| < 1; \quad (b) \quad |a_i| - |b_i| < |\alpha_i| < |a_i| + |b_i|.$$

*Proof.* From (4a) and (7a), it follows that  $|\gamma_1| < 1$ . Assume  $|\gamma_i| < 1$  for  $i = 1, 2, \dots, j - 1$ . Then by (4b, c)

$$\gamma_j = \frac{c_j}{a_j - b_j \gamma_{j-1}},$$

and thus

$$|\gamma_j| \leq \frac{|c_j|}{||a_j| - |b_j| |\gamma_{j-1}|} < \frac{|c_j|}{|a_j| - |b_j|},$$

by the inductive assumption. Finally, by using (7b, c) in the above, it follows that  $|\gamma_j| < 1$  and hence (a) is proved. Using this result and (4b) it follows that

$$|a_i| + |b_i| > |\alpha_i| > |a_i| - |b_i| (\geq |c_i|),$$

which concludes the proof of the inequalities (a) and (b). But then

$$\det A = (\det L) \cdot (\det U) = \prod_{j=1}^n \alpha_j \neq 0. \quad \blacksquare$$

It should be noted that, when the conditions (7) hold, the procedure defined in (4) must be valid. Further if  $b_i c_i = 0$  for some  $i \neq 1, n$ , then the system can be reduced to two systems which are essentially uncoupled. Similarly, if  $c_1 = 0$  or  $b_n = 0$  then  $x_1$  or  $x_n$ , respectively, can be eliminated to get a reduced system.

The operational count for this procedure is somewhat striking:

- (4) requires  $2(n - 1)$  ops.
- (5) requires  $1 + 2(n - 1)$  ops.
- (6) requires  $n - 1$  ops.

or a total of

$$(8) \quad 5n - 4 \text{ ops.}$$

*to solve a single system.* If there are  $m$  such systems to be solved, the quantities  $\alpha_i, \gamma_i$  in (4) need be computed only once and (5) and (6) are then each done  $m$  times for a total of

$$(9) \quad (3n - 2)m + 2n - 2 \text{ ops.}$$

*to solve  $m$  systems.* Consequently, the inverse can be obtained, although it should never be used in such circumstances to solve the system, in not more than

$$3n^2 - 2 \text{ ops.}$$

The low operational counts in (8) and (9) are due to the fact that the zero elements of  $A$  have been accounted for in performing the calculations.

It should be observed that the factorization computed in (4) is not unique. Thus, for instance, we could try the form

$$A = LU \equiv \left[ \begin{array}{cccccc} 1 & & & & & & \\ \beta_2 & 1 & & & & & \\ \beta_3 & & 1 & & & & \\ \cdot & \cdot & & & & & \\ & \cdot & \cdot & & & & \\ & & \cdot & \cdot & & & \\ & & & \beta_n & 1 & & \end{array} \right] \left[ \begin{array}{ccccc} \alpha_1 & c_1 & & & \\ \alpha_2 & c_2 & & & \\ \cdot & \cdot & \ddots & & \\ & & & \ddots & \\ & & & & c_{n-1} \\ & & & & \alpha_n \end{array} \right]$$

The reader should derive the recursions analogous to (4)–(6) for this case and prove the corresponding version of Theorem 5. We give the reader leave to develop a treatment and operational count for the general *band matrix*. A matrix  $(c_{ij})$  of order  $n$  is called a band matrix of width  $(b, a)$  iff

$$c_{ij} = 0 \quad \text{for } j - i \geq a \quad \text{or} \quad i - j \geq b.$$

### 3.3. Block-Tridiagonal Matrices

Another form which is encountered frequently, especially in the numerical solution of partial differential equations and integral equations, is the so-called *block-tridiagonal matrix*

$$(10) \quad A = \left[ \begin{array}{ccccc} A_1 & C_1 & & & \\ B_2 & A_2 & C_2 & & \\ \cdot & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \\ & & & B_i & A_i & C_i \\ & & & \cdot & \cdot & \cdot \\ & & & & \cdot & \cdot \\ & & & & & \cdot & \cdot \\ & & & & & & B_n & A_n \end{array} \right]$$

Here, each of the  $A_i$  represents a square matrix, of order  $m_i$ , and each of the  $B_i$  and  $C_i$  are rectangular matrices which just “fit” the indicated pattern. That is,  $B_i$  must have  $m_i$  rows and  $m_{i-1}$  columns, and  $C_i$  must have  $m_i$  rows and  $m_{i+1}$  columns. Note that if all  $m_i = m$ , then all

the submatrices are square and of order  $m$ . The order of the matrix  $A$  is  $\sum_{i=1}^n m_i$ , or again if all  $m_i = m$  then the order is  $(mn)$ .

A system with coefficient matrix of the form (10) may be solved by a procedure formally analogous to the previous factorization of a Jacobi matrix. Thus, let the system be

$$(11) \quad A\mathbf{x} = \mathbf{f}$$

where now

$$(12) \quad \mathbf{x} = \begin{pmatrix} \mathbf{x}^{(1)} \\ \vdots \\ \mathbf{x}^{(n)} \end{pmatrix}, \quad \mathbf{f} = \begin{pmatrix} \mathbf{f}^{(1)} \\ \vdots \\ \mathbf{f}^{(n)} \end{pmatrix},$$

and each  $\mathbf{x}^{(i)}$  and  $\mathbf{f}^{(i)}$  are  $m_i$ -component column vectors. That is, the components of the vector  $\mathbf{x}$  are grouped into subsets,  $\mathbf{x}^{(i)}$ , and these subsets are to be “eliminated,” as in the Gaussian procedure, a group at a time. Thus, the method to be described is a special case of more general methods known as *group-* or *block-elimination*.

Exactly as in Subsection 3.2 we seek a factorization of the form

$$(13) \quad A = LU =$$

$$\left[ \begin{array}{ccccc} A_1 & & & & \\ B_2 & A_2 & & & \\ B_3 & A_3 & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ \cdot & \cdot & & & \\ B_n & A_n & & & \end{array} \right] \left[ \begin{array}{ccccc} I_1 & \Gamma_1 & & & \\ & I_2 & \Gamma_2 & & \\ & & I_3 & \ddots & \\ & & & \ddots & \ddots \\ & & & & \ddots & \Gamma_{n-1} \\ & & & & & I_n \end{array} \right]$$

where the  $I_j$  are identity matrices of order  $m_j$ , the  $A_j$  are square matrices of order  $m_j$ , and the  $\Gamma_j$  are rectangular matrices with  $m_j$  rows and  $m_{j+1}$  columns. Proceeding formally, we find that

$$(14a) \quad A_1 = A_1, \quad \Gamma_1 = A_1^{-1}C_1;$$

$$(14b) \quad A_i = A_i - B_i \Gamma_{i-1}, \quad i = 2, 3, \dots, n;$$

$$(14c) \quad \Gamma_i = A_i^{-1}C_i, \quad i = 2, 3, \dots, n-1.$$

From the definitions of the matrices involved, it is clear that each  $\Gamma_i$  is rectangular of the indicated order and that the product  $B_i \Gamma_{i-1}$  and hence  $A_i$  is square of order  $m_i$ . The system (11) is now equivalent to

$$(15) \quad Ly = \mathbf{f}, \quad U\mathbf{x} = \mathbf{y}$$

where  $\mathbf{y}$  also has the compound form indicated in (12). We thus obtain formally, from (13) in (15),

$$(16) \quad \begin{aligned} \mathbf{y}^{(1)} &= \mathbf{A}_1^{-1}\mathbf{f}^{(1)} \\ \text{and} \quad \mathbf{y}^{(i)} &= \mathbf{A}_i^{-1}(\mathbf{f}^{(i)} - \mathbf{B}_i\mathbf{y}^{(i-1)}), \quad i = 2, 3, \dots, n, \\ (17) \quad \mathbf{x}^{(n)} &= \mathbf{y}^{(n)} \\ \mathbf{x}^{(i)} &= \mathbf{y}^{(i)} - \Gamma_i \mathbf{x}^{(i+1)}, \quad i = n-1, n-2, \dots, 1. \end{aligned}$$

This method requires (or rather seems to require) the inversion of the  $n$  matrices  $\mathbf{A}_i$  and the formation of the  $2(n-1)$  matrix products  $\mathbf{A}_i^{-1}\mathbf{C}_i$ ,  $\mathbf{B}_i\Gamma_{i-1}$ . To estimate the total number of operations used, we consider the cases where all  $m_i = m$ . Then with Gaussian elimination to obtain the inverses, we require from (1.10) (see discussion below on improving efficiency by not computing inverses explicitly),

$$nm^3 \text{ ops. for all } \mathbf{A}_i^{-1}.$$

The product of two square matrices of order  $m$  requires  $m^3$  operations, hence, we have

$$2(n-1)m^3 \text{ ops. for all } \mathbf{A}_i^{-1}\mathbf{C}_i \text{ and } \mathbf{B}_i\Gamma_{i-1}.$$

Thus, the evaluation of (14) involves not more than

$$(18) \quad (3n-2)m^3 \text{ ops.}$$

The evaluation of (16) and (17) involves only products of  $m$ -component vectors by square matrices and we find

$$\begin{aligned} (16) \text{ requires } &(2n-1)m^2 \text{ ops.;} \\ (17) \text{ requires } &(n-1)m^2 \text{ ops.} \end{aligned}$$

The total for (14), (16), and (17) is thus

$$(19) \quad (3n-2)(m^3 + m^2) \text{ ops.,}$$

*to solve the system (11) with coefficient matrix (10).*

Notice that this number is much superior to estimates of the form  $\frac{1}{3}(nm)^3$  which are appropriate for direct elimination methods applied to arbitrary systems of order  $(nm)$ . In fact, if  $n = m$  the block-elimination scheme requires about  $3m^4$  operations, while from (1.9) straightforward Gaussian elimination uses on the order of  $\frac{1}{3}m^6$  operations. The great gain in economy of operations is again due to the careful account taken of the large number of zero elements in  $A$ . In fact, even greater efficiency is attained if each  $\Gamma_i$  is computed by solving the  $m$  linear systems,  $\mathbf{A}_i\Gamma_i = \mathbf{C}_i$ , and not by computing  $\mathbf{A}_i^{-1}$ ; and if similarly (15) is solved for  $\mathbf{y}^{(i)}$ .

The count in (19) is then reduced from the order of  $3nm^3$  operations to the order of  $\frac{5}{3}nm^3$  operations when we do not compute inverses.

The justification of the block-factorization method is given in

**THEOREM 6.** *If the leading diagonal submatrices*

$$A^{(k)} \equiv \begin{bmatrix} A_1 & C_1 & & & \\ B_2 & A_2 & \ddots & & \\ \vdots & \ddots & \ddots & \ddots & \\ & \ddots & \ddots & \ddots & C_{k-1} \\ & & B_k & A_k & \end{bmatrix}, \quad k = 1, 2, \dots, n,$$

*of the original matrix (10) are non-singular, then the block-factorization in (14) may be carried out (i.e., the  $A_i$  are non-singular).*

*Proof.* This is left to Problem 2. ■

### PROBLEMS, SECTION 3

1. If  $LU = A$  is a factorization of  $A$  satisfying (2), show that  $l_{ik}u_{kj}$  is independent of the choice of  $l_{kk}$  and  $u_{kk}$  that satisfy (2a).
2. Prove Theorem 6.

### 4. ITERATIVE METHODS

The previous direct methods for solving general systems of order  $n$  require about  $n^3/3$  operations. In addition, it has been indicated that, in practical computations with these methods, the errors which are necessarily introduced through rounding may become quite large for large  $n$ . Now we consider iterative methods in which an approximate solution is sought by using fewer operations per iteration. In general, these may be described as methods which proceed from some initial "guess,"  $\mathbf{x}^{(0)}$ , and define a sequence of successive approximations  $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \dots$  which, in principle, converge to the exact solution. If the convergence is sufficiently rapid, the procedure may be terminated at an early stage in the sequence and will yield a good approximation. One of the intrinsic advantages of such methods is the fact that errors, due to roundoff or even blunders, may be damped out as the procedure continues. In fact, special iterative methods are frequently used to improve "solutions" obtained by direct methods (see Subsection 4.3).

A large class of iterative methods may be defined as follows: Let the system to be solved be

$$(1) \quad A\mathbf{x} = \mathbf{f}$$

where  $\det |A| \neq 0$ . Then the coefficient matrix can be *split*, in an infinite number of ways, into the form

$$(2) \quad A = N - P$$

where  $N$  and  $P$  are matrices of the same order as  $A$ . The system (1) is then written as

$$(3) \quad N\mathbf{x} = P\mathbf{x} + \mathbf{f}.$$

Starting with some *arbitrary* vector  $\mathbf{x}^{(0)}$ , we define a sequence of vectors  $\{\mathbf{x}^{(\nu)}\}$ , by the recursion

$$(4) \quad N\mathbf{x}^{(\nu)} = P\mathbf{x}^{(\nu-1)} + \mathbf{f}, \quad \nu = 1, 2, \dots$$

It is now clear that one of the restrictions to be placed on the *splitting* (2) is that

$$(5) \quad \det N \neq 0,$$

in which case the recursions (4) define a unique sequence of vectors for all  $\mathbf{x}^{(0)}$  and  $\mathbf{f}$ . As a practical matter, it is also clear that  $N$  should be chosen such that a system of the form

$$(6) \quad Ny = z$$

can be “easily” solved. Furthermore, if greater accuracy is desired, it would be better to calculate with (4) in the equivalent form

$$N(\mathbf{x}^{(\nu)} - \mathbf{x}^{(\nu-1)}) = \mathbf{f} - A\mathbf{x}^{(\nu-1)}.$$

This point will be discussed further in Subsection 4.3.

The convergence of the sequence  $\{\mathbf{x}^{(\nu)}\}$  to the solution  $\mathbf{x}$  of (1) is studied by introducing the matrix

$$(7) \quad M = N^{-1}P$$

and the error vectors

$$(8) \quad \mathbf{e}^{(\nu)} = \mathbf{x}^{(\nu)} - \mathbf{x}, \quad \nu = 0, 1, 2, \dots$$

Subtract (3) from (4) to obtain, upon multiplication by  $N^{-1}$ ,

$$(9) \quad \begin{aligned} \mathbf{e}^{(\nu)} &= M\mathbf{e}^{(\nu-1)} \\ &= M^2\mathbf{e}^{(\nu-2)} \\ &\vdots \\ &= M^\nu\mathbf{e}^{(0)}, \quad \nu = 1, 2, \dots, \end{aligned}$$

where  $\mathbf{e}^{(0)}$  is the *arbitrary* initial error. Thus, it is clear that a *sufficient condition for convergence*, i.e., that  $\lim_{v \rightarrow \infty} \mathbf{e}^{(v)} = \mathbf{0}$ , is that  $\lim_{v \rightarrow \infty} M^v = O$ , and this is also *necessary if the method is to converge for all  $\mathbf{e}^{(0)}$* .

A matrix,  $M$ , that satisfies this condition is called a *convergent matrix*. The basic results characterizing convergent matrices have been established in Chapter 1, Theorem 1.4 and its Corollary, which we restate here as

**THEOREM 1.** *The matrix  $M$  is convergent, i.e.,*

$$\lim_{v \rightarrow \infty} M^v = O,$$

*iff all eigenvalues of  $M$  are less than one in absolute value.* ■

(This condition is frequently stated as  $\rho(M) < 1$  where  $\rho(M)$  is the *spectral radius* of  $M$  defined by

$$\rho(M) \equiv \max_i |\lambda_i|$$

where the  $\lambda_i$  are the eigenvalues of  $M$ .)

**COROLLARY 1.** *The matrix  $M$  is convergent if, for any matrix norm,  $\|M\| < 1$ .* ■

It is, in general, difficult to verify the conditions of Theorem 1. However, Corollary 1 may frequently be used to show that  $\rho(M) < 1$ . We have (see Chapter 1, Section 1, for the notation)

**COROLLARY 2.** *The matrix  $M \equiv (m_{ij})$  is convergent if either*

$$(10a) \quad \|M\|_\infty \equiv \max_i \sum_{j=1}^n |m_{ij}| < 1;$$

or

$$(10b) \quad \|M\|_1 \equiv \max_j \sum_{i=1}^n |m_{ij}| < 1.$$

*Proof.* We have shown in Chapter 1, Section 1, that  $\|M\|_\infty$  and  $\|M\|_1$  are matrix norms. ■

Let us return to the iteration scheme (4)–(9) and assume it to be a convergent one. We introduce the notion of the rate of convergence,  $R$ , of the iterative scheme by setting

$$(11) \quad R \equiv -\log \rho(M).$$

The significance of this quantity is most easily seen if we recall the corollary to Theorem 1.3 of Chapter 1, which states that:  $\rho(M) = \inf_{\{\|\cdot\|\}} \|M\|$  (here  $\{\|\cdot\|\}$  is the set of natural norms). Given the initial error,  $\mathbf{e}^{(0)}$ , (9) permits the estimate in terms of any natural norm

$$\|\mathbf{e}^{(v)}\| \leq \|M\|^v \|\mathbf{e}^{(0)}\|.$$

Then for a given  $\epsilon > 0$ , there is some norm such that

$$(12) \quad \|\mathbf{e}^{(v)}\| \leq [\rho(M) + \epsilon]^v \|\mathbf{e}^{(0)}\|.$$

On the other hand, again from (9), if  $\mathbf{e}^{(0)}$  is an eigenvector of  $M$  corresponding to the largest eigenvalue,  $\|\mathbf{e}^{(v)}\| = [\rho(M)]^v \|\mathbf{e}^{(0)}\|$ . Let it be required to reduce the amplitude of the error by a factor of at least  $10^{-m}$ ,  $m > 0$ . From (12), we see that, in some norm, the error amplitude is reduced by a factor close to  $[\rho(M)]^v$ . The number of iterations required is the least value of  $v$  for which

$$[\rho(M)]^v \leq 10^{-m}.$$

By taking logs and recalling that  $0 \leq \rho(M) < 1$ , we obtain

$$(13) \quad v \geq \frac{m}{-\log \rho(M)} = \frac{m}{R}.$$

Thus, the number of iterations required to reduce the initial error by the factor  $10^{-m}$  is inversely proportional to  $R$ , the rate of convergence.

#### 4.1. Jacobi or Simultaneous Iterations

A special case (attributed to *Jacobi*) of the previous general theory is

$$(14) \quad N \equiv (a_{ii}\delta_{ij}), \quad P \equiv N - A = (a_{ii}\delta_{ij} - a_{ij}).$$

From (14) in (4), it is seen that the components  $x_i^{(v)}$  of the  $v$ th iterate are simply computed with  $\mathbf{x}^{(0)}$  arbitrary by

$$(15) \quad x_i^{(v)} = \frac{1}{a_{ii}} \left( f_i - \sum_{\substack{j=1 \\ (j \neq i)}}^n a_{ij} x_j^{(v-1)} \right)$$

$$i = 1, 2, \dots, n; \quad v = 1, 2, \dots$$

Thus, this procedure may be employed provided only that  $a_{ii} \neq 0$  for all  $i = 1, 2, \dots, n$ . However, for the convergence of these iterations, Theorem 1 requires that all roots of  $\det |\lambda I - N^{-1}P| = 0$  satisfy  $|\lambda| < 1$ . This equation can be written as, assuming  $\det |N| \neq 0$ ,

$$(16) \quad \det |\lambda N - P| = \det \begin{vmatrix} \lambda a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & \lambda a_{22} & & \vdots \\ \vdots & & \ddots & a_{n-1,n} \\ a_{n1} & a_{n2} & & \lambda a_{nn} \end{vmatrix} = 0.$$

In general, the roots of such an equation, for large  $n$ , are not easily obtained and so we seek simpler sufficient conditions for convergence, as given in Corollary 2. The relevant matrix  $M$  is easily obtained since  $N^{-1} = (a_{ii}^{-1}\delta_{ij})$  and thus

$$(17) \quad M \equiv N^{-1}P = \left( \delta_{ij} - \frac{a_{ij}}{a_{ii}} \right)$$

Now conditions (10a) and (10b) of Corollary 2 become

$$(18a) \quad \|M\|_\infty = \max_i \sum_{\substack{j=1 \\ (j \neq i)}}^n \left| \frac{a_{ij}}{a_{ii}} \right| < 1,$$

$$(18b) \quad \|M\|_1 = \max_j \sum_{\substack{i=1 \\ (i \neq j)}}^n \left| \frac{a_{ij}}{a_{ii}} \right| < 1.$$

These tests are easily applied in practice. Since  $\rho(M) \leq \|M\|$  we obtain a lower bound on the rate of convergence

$$(19) \quad R = \log \frac{1}{\rho(M)} \geq \log \frac{1}{\min(\|M\|_1, \|M\|_\infty)}.$$

The operational count for the Jacobi iteration is simply obtained from (15); it is

$$(20) \quad n^2 \text{ ops. per iteration.}$$

Thus by (13), if these iterations converge they require a total of about

$$\frac{m \times n^2}{R} \text{ ops.,}$$

to reduce the initial error by at least  $10^{-m}$ . We see that if such an iterative method is to be at least as efficient as the direct elimination method it should have a rate of convergence and required accuracy factor, say  $10^{-m}$ , satisfying

$$\frac{m}{R} \leq \frac{n}{3}.$$

(We assume here that  $m$  has been so chosen that the iterative solution will have an accuracy comparable to the accuracy obtained by the direct elimination method using the same number of digits in the arithmetic.)

#### 4.2. Gauss-Seidel or Successive Iterations

It is clear from (15) that in the ordinary Jacobi iterations some components of  $\mathbf{x}^{(v)}$  are known, but not used, while computing the remaining components. The Gauss-Seidel method is a modification of the Jacobi method in which all of the latest known components are used. The term "successive" which is frequently applied to this method refers to the fact that "new" components are successively used as they are obtained. (In contrast, the previous scheme was called "simultaneous" since new components were not employed as found, i.e., the "new" components were introduced simultaneously at the end of the iterative cycle.)

The obvious modification of (15) suggested by the above remarks is, with  $\mathbf{x}^{(0)}$  arbitrary,

$$(21) \quad x_i^{(v)} = \frac{1}{a_{ii}} \left( f_i - \sum_{j=1}^{i-1} a_{ij}x_j^{(v)} - \sum_{j=i+1}^n a_{ij}x_j^{(v-1)} \right), \\ i = 1, 2, \dots, n; \quad v = 1, 2, \dots.$$

The splitting of  $A$  that yields this iterative scheme is

$$(22) \quad N \equiv \begin{bmatrix} a_{11} & & & \\ a_{21} & a_{22} & & \\ \cdot & \cdot & \ddots & \\ \cdot & & \ddots & \\ \cdot & & & \\ a_{n1} & a_{n2} & \cdots & a_{nn} \end{bmatrix}, \quad P \equiv N - A.$$

Since  $N$  is triangular,  $\det |N| \neq 0$  is assured again if  $a_{ii} \neq 0$ ;  $i = 1, 2, \dots, n$ . The characteristic equation, whose roots must be in absolute value less than unity for convergence, is now of the form

$$(23) \quad \det \begin{vmatrix} \lambda a_{11} & a_{12} & \cdots & a_{1n} \\ \lambda a_{21} & \lambda a_{22} & & \\ \vdots & & \ddots & \\ & & & a_{n-1, n} \\ \lambda a_{n1} & \lambda a_{n2} & \cdots & \lambda a_{nn} \end{vmatrix} = 0.$$

The roots of this equation are just as difficult to find as are those of (16), but the sufficient conditions of Corollary 2 are now much more complicated than those for the Jacobi iteration. However, a simple sufficient condition for convergence of the Gauss-Seidel method can be obtained. To derive this condition, we introduce the error vectors defined in (8) and find from (1) and (21) that the components of these vectors must satisfy

$$(24) \quad e_i^{(v)} = - \sum_{j=1}^{i-1} \frac{a_{ij}}{a_{ii}} e_j^{(v)} - \sum_{j=i+1}^n \frac{a_{ij}}{a_{ii}} e_j^{(v-1)}, \\ i = 1, 2, \dots, n; \quad v = 1, 2, \dots$$

The result to be proved may now be stated as

**LEMMA 1.** *Let the vectors  $\mathbf{e}^{(v)}$ ,  $v = 1, 2, \dots$ , be defined by (24) with  $\mathbf{e}^{(0)}$  arbitrary. Define the maximum norm,  $\|\cdot\|_\infty$ , and factors,  $r_i$ , by*

$$(25a) \quad \|\mathbf{e}^{(v)}\|_\infty \equiv \max_j |e_j^{(v)}|,$$

$$(25b) \quad r_i \equiv \sum_{\substack{j=1 \\ (j \neq i)}}^n \left| \frac{a_{ij}}{a_{ii}} \right|;$$

and let the matrix  $A$  satisfy

$$(26) \quad r \equiv \max_i r_i < 1.$$

Then

$$(27) \quad \|\mathbf{e}^{(v)}\|_\infty \leq r^v \|\mathbf{e}^{(0)}\|_\infty,$$

and  $\mathbf{e}^{(v)} \rightarrow \mathbf{0}$  as  $v \rightarrow \infty$ .

*Proof.* The lemma clearly follows from the inequalities

$$(28) \quad \|\mathbf{e}^{(v)}\|_\infty \leq r \|\mathbf{e}^{(v-1)}\|_\infty, \quad v = 1, 2, \dots;$$

which we shall prove by induction (on the components of  $\mathbf{e}^{(v)}$ ). From (24) with  $i = 1$  we obtain, using (25) and (26),

$$\begin{aligned} |e_1^{(v)}| &\leq \sum_{j=2}^n \left| \frac{a_{1j}}{a_{11}} \right| \cdot |e_j^{(v-1)}| \\ &\leq \|\mathbf{e}^{(v-1)}\|_\infty \sum_{j=2}^n \left| \frac{a_{1j}}{a_{11}} \right| = \|\mathbf{e}^{(v-1)}\|_\infty \cdot r_1 \\ &\leq r \|\mathbf{e}^{(v-1)}\|_\infty. \end{aligned}$$

Now assume  $|e_k^{(v)}| \leq r \|\mathbf{e}^{(v-1)}\|_\infty$  for  $k = 1, 2, \dots, i - 1$ . Then again from (24), recalling that  $r < 1$

$$\begin{aligned} |e_i^{(v)}| &\leq \sum_{j=1}^{i-1} \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |e_j^{(v)}| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \cdot |e_j^{(v-1)}| \\ &\leq \|\mathbf{e}^{(v-1)}\|_\infty \left\{ \sum_{j=1}^{i-1} r \left| \frac{a_{ij}}{a_{ii}} \right| + \sum_{j=i+1}^n \left| \frac{a_{ij}}{a_{ii}} \right| \right\} \\ &\leq \|\mathbf{e}^{(v-1)}\|_\infty \sum_{\substack{j=1 \\ (j \neq i)}}^n \left| \frac{a_{ij}}{a_{ii}} \right| = r_i \|\mathbf{e}^{(v-1)}\|_\infty \\ &\leq r \|\mathbf{e}^{(v-1)}\|_\infty. \end{aligned}$$

Thus, the induction argument is complete and since the above inequality is valid for all  $i = 1, 2, \dots, n$ , it follows by (25) that (28) is valid. ■

The convergence test of this lemma is easily applied and is formally identical to that of (18a) for the Jacobi method. However, it is not generally true that if the Gauss-Seidel method converges then the Jacobi method will converge, nor is the converse generally true.

See Subsection 4.4 for other convergence tests.

### 4.3. Method of Residual Correction

This iterative scheme improves upon the accuracy of the approximate solution of (1) (obtained for example by the Gaussian elimination method), by using the approximate numerical triangular factorization of  $A$ . That is, the triangularization of (1), performed with  $t$  digits, produces  $\mathcal{L}$  (lower),  $\mathcal{U}$  (upper), and  $\mathbf{x}^{(0)}$ . Now define

$$(29) \quad \begin{aligned} N &\equiv \mathcal{L}\mathcal{U}, \quad P \equiv N - A, \\ \mathbf{r}^{(0)} &\equiv \mathbf{f} - Ax^{(0)}. \end{aligned}$$

Observe that  $N$  is easily invertible, or rather that the equation

$$\mathcal{L}\mathcal{U}\mathbf{y} = \mathbf{z}$$

may be readily “solved,” since  $\mathcal{L}$  and  $\mathcal{U}$  are triangular, by using  $n(n + 1)$  operations. [If  $\mathcal{L} \equiv (s_{ij})$  has  $s_{ii} = 1$  for all  $i$ , then the number of operations used to solve  $\mathcal{L}\mathbf{w} = \mathbf{z}$  is  $n(n - 1)/2$ , while the number for solving  $\mathcal{U}\mathbf{y} = \mathbf{w}$  is  $n(n + 1)/2$ . Hence, in this case  $n^2$  is the operational count for solving  $N\mathbf{y} = \mathbf{z}$ .]

Now, the iteration scheme given by (4) is convergent if  $M$ , defined by (7), satisfies

$$\|M\| \equiv \|I - N^{-1}A\| < 1.$$

This inequality is satisfied if  $\|P\| \cdot \|A^{-1}\| < \frac{1}{2}$  (see Problem 5). In practice, (4) is not solved in the form

$$(30) \quad \mathcal{L}\mathcal{U}\mathbf{x}^{(v)} = (\mathcal{L}\mathcal{U} - A)\mathbf{x}^{(v-1)} + \mathbf{f}.$$

Rather, we introduce the *change* in the iterate by

$$\delta\mathbf{x}^{(v-1)} \equiv \mathbf{x}^{(v)} - \mathbf{x}^{(v-1)}$$

and the residual of the iterate by

$$(31) \quad \mathbf{r}^{(v-1)} \equiv \mathbf{f} - A\mathbf{x}^{(v-1)}.$$

Then (30) can be written simply as

$$(32) \quad \mathcal{L}\mathcal{U}[\delta\mathbf{x}^{(v-1)}] = \mathbf{r}^{(v-1)},$$

and the computations are done with these equations.

The evaluation of  $\mathbf{r}^{(v)}$  involves  $n^2$  operations; hence each iteration step, (31) and (32), requires  $n(2n + 1)$  operations (only  $2n^2$  operations if  $s_{ii} = 1$  for all  $i$ ). By using (1) and (31) the error satisfies

$$(33) \quad \|\mathbf{x}^{(v)} - \mathbf{x}\| = \|A^{-1}\mathbf{r}^{(v)}\| \leq \|A^{-1}\| \|\mathbf{r}^{(v)}\|.$$

But from (32), the definition of  $M$ , and the corollary to Theorem 1.5 of Chapter 1,

$$\begin{aligned} \|A^{-1}\mathbf{r}^{(v)}\| &= \|A^{-1}N\delta\mathbf{x}^{(v)}\| = \|(I - M)^{-1}\delta\mathbf{x}^{(v)}\| \\ &\leq \frac{\|\delta\mathbf{x}^{(v)}\|}{1 - q}, \end{aligned}$$

provided  $q \equiv \|M\| = \|N^{-1}(N - A)\| < 1$ .

As described in Subsection 1.2, the numerical solution of (32) produces a vector  $\delta\mathbf{x}^{(v-1)}$  that satisfies

$$(34) \quad \mathcal{L}_{v-1}\mathcal{U}_{v-1}\delta\mathbf{x}^{(v-1)} = \mathbf{r}^{(v-1)},$$

where

$$\mathcal{L}_{v-1} \equiv \mathcal{L} + \delta\mathcal{L}_{v-1}, \quad \mathcal{U}_{v-1} \equiv \mathcal{U} + \delta\mathcal{U}_{v-1}.$$

The perturbations  $\delta\mathcal{L}_v$  and  $\delta\mathcal{U}_v$  are small relative to  $\mathcal{L}$  and  $\mathcal{U}$  respectively if the number of digits carried in the arithmetic calculations is large enough. Set

$$N_v \equiv \mathcal{L}_v\mathcal{U}_v, \quad P_v \equiv N_v - A$$

and

$$M_v = N_v^{-1}P_v.$$

Then the error

$$\mathbf{e}^{(v)} \equiv \mathbf{x}^{(v)} - \mathbf{x},$$

satisfies

$$\begin{aligned}\mathbf{e}^{(v)} &= M_{v-1} \mathbf{e}^{(v-1)} \\ &= M_{v-1} M_{v-2} \mathbf{e}^{(v-2)} \\ &\vdots \\ &= M_{v-1} M_{v-2} \cdots M_0 \mathbf{e}^{(0)}.\end{aligned}$$

If  $\|M_i\| \leq q < 1$  for all  $i$ , then  $\|\mathbf{e}^{(v)}\| \leq q^v \|\mathbf{e}^{(0)}\|$ , and the scheme is convergent for any  $\mathbf{e}^{(0)}$ .

As a practical matter, from equations (31) and (32), we see that  $\mathbf{r}^v \rightarrow \mathbf{0}$  may occur only if the right-hand side of (31) is calculated to ever higher precision as  $v$  increases. On the other hand, equation (32) or equation (34) requires only single precision accuracy for  $\mathbf{r}^{(v-1)}$ , in order to determine  $\delta \mathbf{x}^{(v-1)}$  by using single precision arithmetic.

#### 4.4. Positive Definite Systems

Many of the large order linear systems that arise in practice have real symmetric matrices which are positive definite. In such cases we can show that a quite general class of iteration methods converges. We state this result as

**THEOREM 2.** *Let  $A$  be Hermitian (of order  $n$ ) and  $N$  be any non-singular matrix (of order  $n$ ) for which*

$$(35) \quad Q \equiv N + N^* - A$$

*is positive definite. Then the matrix*

$$M \equiv I - N^{-1}A$$

*is convergent iff  $A$  is positive definite.*

*Proof.* For any eigenvalue,  $\lambda$ , and corresponding eigenvector,  $\mathbf{u}$ , of  $M$  we have

$$M\mathbf{u} = \lambda\mathbf{u}.$$

But since  $N$  is non-singular this implies

$$(36) \quad A\mathbf{u} = (1 - \lambda)N\mathbf{u},$$

and so  $\lambda = 1$  iff  $A\mathbf{u} = \mathbf{0}$ .

Now let  $A$  be positive definite (i.e.,  $\mathbf{v}^* A \mathbf{v} > 0$  if  $\mathbf{v} \neq \mathbf{0}$ ) and  $\mathbf{u}$  be any eigenvector of  $M$ . Then  $A\mathbf{u} \neq \mathbf{0}$  so that the corresponding eigenvalue  $\lambda$

of  $M$  satisfies  $\lambda \neq 1$ . By taking the complex inner product of each side of (36) with  $\mathbf{u}$  we then obtain

$$\frac{1}{1 - \lambda} = \frac{\mathbf{u}^* N \mathbf{u}}{\mathbf{u}^* A \mathbf{u}}.$$

The complex conjugate of this expression is, since  $\mathbf{u}^* A \mathbf{u}$  is real,

$$\frac{1}{1 - \bar{\lambda}} = \frac{\mathbf{u}^* N^* \mathbf{u}}{\mathbf{u}^* A \mathbf{u}}.$$

If we add these two equations we get

$$2 \operatorname{Re} \frac{1}{1 - \lambda} = \frac{\mathbf{u}^* (N + N^*) \mathbf{u}}{\mathbf{u}^* A \mathbf{u}}.$$

Now set  $\lambda = \alpha + i\beta$  and recall (35) to write this as

$$\frac{2(1 - \alpha)}{(1 - \alpha)^2 + \beta^2} = 1 + \frac{\mathbf{u}^* Q \mathbf{u}}{\mathbf{u}^* A \mathbf{u}} > 1,$$

since by hypothesis  $Q$  is positive definite. Hence, we have the inequality

$$|\lambda|^2 = \alpha^2 + \beta^2 < 1.$$

The sufficiency is thus demonstrated. The necessity part of the proof is indicated in Problems 1 and 2. ■

As an immediate corollary of this theorem, we have a result on the convergence of the Gauss-Seidel method for Hermitian matrices.

**COROLLARY 1.** *Let  $A$  be Hermitian with positive diagonal elements. Then the Gauss-Seidel method for this matrix converges iff  $A$  is positive definite.*

*Proof.* By the hypothesis on  $A$  it can be written as

$$A = D + E + E^*$$

where  $D$  is a diagonal matrix of positive diagonal elements and  $E$  is strictly lower triangular (i.e., zeros on and above the diagonal). The Gauss-Seidel method applied to  $A$ , see (22), is equivalent to the splitting

$$N \equiv D + E, \quad P = -E^*.$$

However, with this choice for  $N$  we have

$$Q \equiv N + N^* - A = D$$

which is clearly positive definite. Thus the hypothesis of Theorem 2 applies and the proof is concluded. ■

Similarly, we obtain a result on the convergence of the Jacobi iterations as a special case of

**COROLLARY 2.** *Let  $D = D^*$  be non-singular and*

$$D - (E + E^*)$$

*be positive definite. Then*

$$D^{-1}(E + E^*)$$

*is convergent iff  $A \equiv D + (E + E^*)$  is positive definite.*

*Proof.* Use  $N \equiv D$  in the theorem. ■

In the special case that  $D$  is a diagonal matrix, Corollary 2 yields the convergence of the Jacobi method for the matrix  $A$ .

#### 4.5. Block Iterations

There are other splittings of  $A$  which in many important cases yield rapidly convergent iterations. In particular, since tridiagonal and block-tridiagonal systems are easily solved, it is natural to consider iterations in which  $N$  has either of these forms. Many of the large order systems which arise in the finite difference methods for partial differential equations suggest such block iterations. More generally, if the elements “close” to the diagonal of a matrix are large compared to the other elements, it is usually advantageous to include all of these large elements in  $N$  (assuming, of course, that the resulting systems which determine the iterates are still easily solved). Of course, in all applications of these block methods, attempts should be made to prove the convergence of the method and, if possible, to estimate the rate of convergence.

### PROBLEMS, SECTION 4

1. Let the sequence  $\{\mathbf{v}_\nu\}$  be defined, with  $\mathbf{v}_0$  arbitrary, by

$$\mathbf{v}_{\nu+1} = M\mathbf{v}_\nu, \quad \nu = 0, 1, \dots,$$

where  $M \equiv I - N^{-1}A$  and  $A$  is Hermitian. Then

- (a) Verify the identity

$$\mathbf{v}_\nu^* A \mathbf{v}_\nu - \mathbf{v}_{\nu+1}^* A \mathbf{v}_{\nu+1} = (\mathbf{v}_\nu - \mathbf{v}_{\nu+1})^* Q(\mathbf{v}_\nu - \mathbf{v}_{\nu+1})$$

where  $Q \equiv N + N^* - A$ ;

- (b) If  $Q$  is positive definite show that  $\{\mathbf{v}_\nu^* A \mathbf{v}_\nu\}$  is a non-increasing sequence. (In fact, the sequence is strictly decreasing if 1 is not an eigenvalue of  $M$ .)

2. Use part (b) of Problem 1 to show that if  $M$  is convergent then  $A$  is positive definite.

[Hint: Use proof by contradiction; assume  $\mathbf{v}_0^* A \mathbf{v}_0 \leq 0$  for some  $\mathbf{v}_0 \neq \mathbf{0}$ . Then, since  $M$  is convergent,  $\mathbf{v}_1 \neq \mathbf{v}_0$ . Therefore,

$$\mathbf{v}_v^* A \mathbf{v}_v \leq \mathbf{v}_1^* A \mathbf{v}_1 < \mathbf{v}_0^* A \mathbf{v}_0 \leq 0.$$

This is a contradiction, since the convergence of  $M$  implies  $\mathbf{v}_v \rightarrow \mathbf{0}$ .]

3. Analyze the convergence of the Jacobi and the Gauss-Seidel iterative methods for the second order matrix

$$A \equiv \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}, \quad |\rho| < 1, \mathbf{x}_0 \neq \mathbf{0}.$$

4. Determine when the Jacobi iterative method converges for the compound matrix

$$A \equiv \begin{pmatrix} I & S \\ S^T & I \end{pmatrix}, \text{ with } I \text{ and } S$$

of order  $n$ .

[Hint: Work with the compound vectors  $\begin{pmatrix} \mathbf{x}_v \\ \mathbf{y}_v \end{pmatrix}$ . Define the compound error vectors

$$\begin{pmatrix} \mathbf{e}_v \\ \mathbf{g}_v \end{pmatrix} \equiv \begin{pmatrix} \mathbf{x}_v \\ \mathbf{y}_v \end{pmatrix} - \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}.$$

Find a recursion formula for  $\{\mathbf{e}_v\}$  that doesn't involve  $\{\mathbf{g}_v\}$ .]

5. The convergence of the residual correction scheme defined by (30) is assured if  $\|I - N^{-1}A\| < 1$ . Verify that this inequality holds if

$$\|P\| \cdot \|A^{-1}\| < \frac{1}{2}.$$

[Hint: Let  $B = A^{-1}N$ . Then

$$\begin{aligned} I - N^{-1}A &= I - B^{-1} = B^{-1}(B - I) \\ &= B^{-1}(A^{-1}P). \end{aligned}$$

Note that  $B = A^{-1}P + I$  and therefore, by the remark following the corollary to Theorem 1.5 of Chapter 1, we have  $\|B^{-1}\| < 2$ .]

## 5. THE ACCELERATION OF ITERATIVE METHODS

Given any iteration procedure, for a specific system of equations, it may be possible to improve its rate of convergence by a simple device. Such modifications, which we call *acceleration*, are frequently termed “*extrapolation*,” “*over-relaxation*,” or various other names depending upon the problem to which they are applied or perhaps upon the particular form of device which is used. In any event, the general principle common to almost all acceleration procedures is the introduction of a splitting,

similar to (4.2), which depends upon some real parameter, say  $\alpha$ , in an “appropriate” manner. The splitting may be denoted by

$$A = N(\alpha) - P(\alpha)$$

and is still subject to the requirement that

$$\det |N(\alpha)| \neq 0.$$

(This will place some restriction on the permissible values of  $\alpha$ .) Now, as has been shown in Section 4, an iteration scheme based on the above splitting will converge, for arbitrary initial vectors, iff all eigenvalues of

$$M(\alpha) \equiv N^{-1}(\alpha)P(\alpha),$$

are in absolute value less than unity.

Let these eigenvalues be denoted by

$$\lambda_i(\alpha), \quad i = 1, 2, \dots, n;$$

where, as indicated, their values may depend upon the choice of the parameter  $\alpha$ . Now if a value of  $\alpha$  can be determined such that

$$\rho[M(\alpha)] \equiv \max_i |\lambda_i(\alpha)| < 1,$$

the scheme will converge. Furthermore, since the rate of convergence is

$$R(\alpha) = \log \frac{1}{\rho[M(\alpha)]},$$

the convergence is “best” for the value  $\alpha = \alpha_*$  such that

$$\rho[M(\alpha_*)] = \min_{\alpha} \rho[M(\alpha)].$$

The selection of an *optimal*  $\alpha_*$  is the most important feature of acceleration procedures.

Some acceleration procedures that are commonly used are described as follows: Let some definite splitting, (4.2), be given by

$$(1) \quad A = N_0 - P_0,$$

where  $N_0$  and  $P_0$  are fixed matrices with  $\det |N_0| \neq 0$ . Let the relevant eigenvalues of this scheme, i.e., those of

$$(2a) \quad M_0 = N_0^{-1}P_0,$$

be

$$(2b) \quad \lambda_i, \quad i = 1, 2, \dots, n.$$

Then we introduce the one-parameter family of splittings

$$(3) \quad \begin{aligned} N(\alpha) &= (1 + \alpha)N_0, \\ P(\alpha) &= (1 + \alpha)N_0 - A = P_0 + \alpha N_0. \end{aligned}$$

In order that  $\det |N(\alpha)| \neq 0$ , we need only require  $\alpha \neq -1$ . Then if the eigenvalues of  $M(\alpha) \equiv N^{-1}(\alpha)P(\alpha)$  are denoted by  $\mu_i(\alpha)$ ,  $i = 1, 2, \dots, n$ , we claim that

$$(4) \quad \mu_i(\alpha) = \frac{\lambda_i + \alpha}{1 + \alpha}, \quad i = 1, 2, \dots, n.$$

The verification of (4) requires only a simple application of the definitions of eigenvalue and eigenvector. Specifically from (2) and (3) we have

$$\begin{aligned} M(\alpha) &= N^{-1}(\alpha)P(\alpha) = \frac{1}{1 + \alpha} N_0^{-1}(P_0 + \alpha N_0) \\ &= \frac{\alpha}{1 + \alpha} I + \frac{1}{1 + \alpha} M_0. \end{aligned}$$

Thus, if  $\mathbf{u}$  is any eigenvector of  $M_0$  belonging to the eigenvalue  $\lambda$ , that is  $M_0\mathbf{u} = \lambda\mathbf{u}$ , we obtain from the above

$$M(\alpha)\mathbf{u} = \frac{\alpha}{1 + \alpha} \mathbf{u} + \frac{\lambda}{1 + \alpha} \mathbf{u} = \frac{\lambda + \alpha}{1 + \alpha} \mathbf{u}.$$

That is,  $\mathbf{u}$  must also be an eigenvector of  $M(\alpha)$  belonging to the eigenvalue  $(\lambda + \alpha)/(1 + \alpha)$ . Conversely, if  $M(\alpha)\mathbf{v} = \mu\mathbf{v}$  we obtain

$$\mu\mathbf{v} = M(\alpha)\mathbf{v} = \frac{\alpha}{1 + \alpha} \mathbf{v} + \frac{1}{1 + \alpha} M_0\mathbf{v},$$

or, since  $1 + \alpha \neq 0$  by assumption,

$$M_0\mathbf{v} = [\mu(1 + \alpha) - \alpha]\mathbf{v}.$$

Thus every eigenvector of  $M(\alpha)$  is an eigenvector of  $M_0$  and (4) is established for all  $\alpha \neq -1$ .

In order to determine convergent schemes of the form (3), we must study the relation (4). This is done first for a very important class of special cases in which the “best” such scheme can be obtained. These results may be stated as

**THEOREM 1.** *Let  $N_0$  and  $P_0$  be such that the eigenvalues  $\lambda_i$  of  $N_0^{-1}P_0$  are all real and satisfy*

$$(5) \quad \lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n < 1.$$

*Then the scheme (3) will converge for any  $\alpha$  such that*

$$(6) \quad \alpha > -\frac{1 + \lambda_1}{2} > -1.$$

Furthermore, the largest rate of convergence for these schemes is obtained when

$$(7) \quad \alpha = \alpha_* \equiv -\frac{\lambda_1 + \lambda_n}{2}$$

for which value

$$(8) \quad \rho[M(\alpha_*)] \equiv \min_{\alpha} \rho[M(\alpha)] = \min_{\alpha} \left( \max_{i=1}^n |\mu_i(\alpha)| \right) = \frac{\lambda_n - \lambda_1}{2 - \lambda_1 - \lambda_n} < 1.$$

*Proof.* A scheme of the form (3) will converge if  $|\mu_i(\alpha)| < 1$ ,  $i = 1, 2, \dots, n$ . Let us introduce

$$(9) \quad x \equiv \frac{1}{1 + \alpha}; \quad m_j \equiv \lambda_j - 1, \quad j = 1, 2, \dots, n.$$

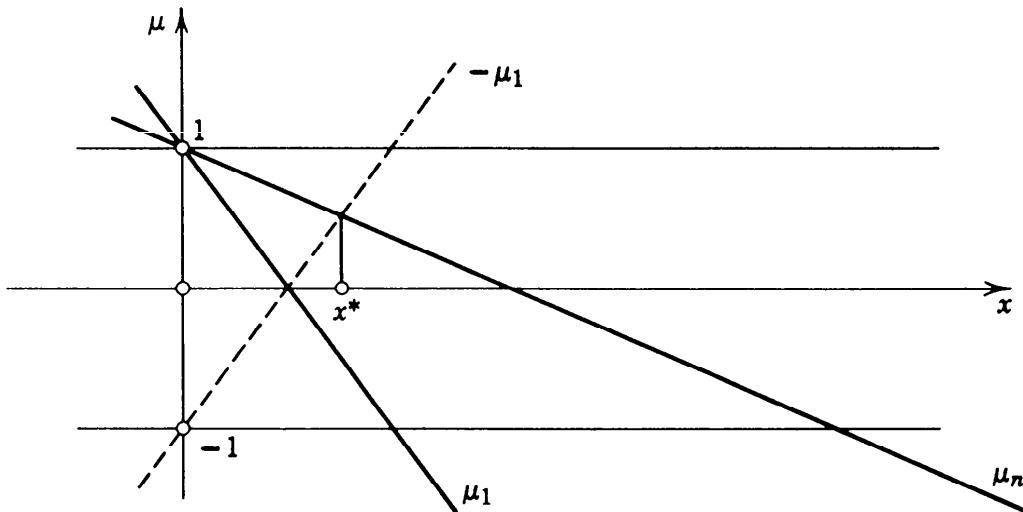


Figure 1

Then (4) can be written as

$$(10) \quad \mu_i = m_i x + 1, \quad i = 1, 2, \dots, n,$$

where by (5), all  $m_i < 0$ . The equations (10), for the  $\mu_i$  as functions of  $x$ , represent  $n$  straight lines with negative slopes. Let us assume that the  $\lambda_i$  have been ordered as in (5). Then by (9) we have

$$m_1 \leq m_2 \leq \dots \leq m_n < 0,$$

and all the lines (10) are bounded by those for  $i = 1$  and  $i = n$  (see Figure 1). Thus, we have for

$$(11) \quad \begin{aligned} x > 0: \quad & \mu_1 = m_1 x + 1 \leq \mu_i \leq m_n x + 1 = \mu_n; \\ x < 0: \quad & \mu_n = m_n x + 1 \leq \mu_i \leq m_1 x + 1 = \mu_1, \quad i = 1, 2, \dots, n. \end{aligned}$$

Clearly then, all  $\mu_i < 1$  iff  $x > 0$ . Similarly, all  $\mu_i > -1$  iff  $\mu_1 > -1$  or equivalently  $x < -2/m_1$ . Thus,  $|\mu_i| < 1$  iff  $0 < x < -2/m_1$ , and using (9) we obtain (6).

For  $x > 0$  we have by (11)

$$\mu = \max_{i=1}^n |\mu_i| = \max(|m_1x + 1|, |m_nx + 1|).$$

From Figure 1 it is then clear that

$$\min_{x>0} \mu = |m_1x_* + 1| = |m_nx_* + 1| = \frac{m_1 - m_n}{m_1 + m_n},$$

where  $x_* = -2/(m_1 + m_n)$ . Upon applying (9) again, we obtain (7) and (8) and the proof is complete. ■

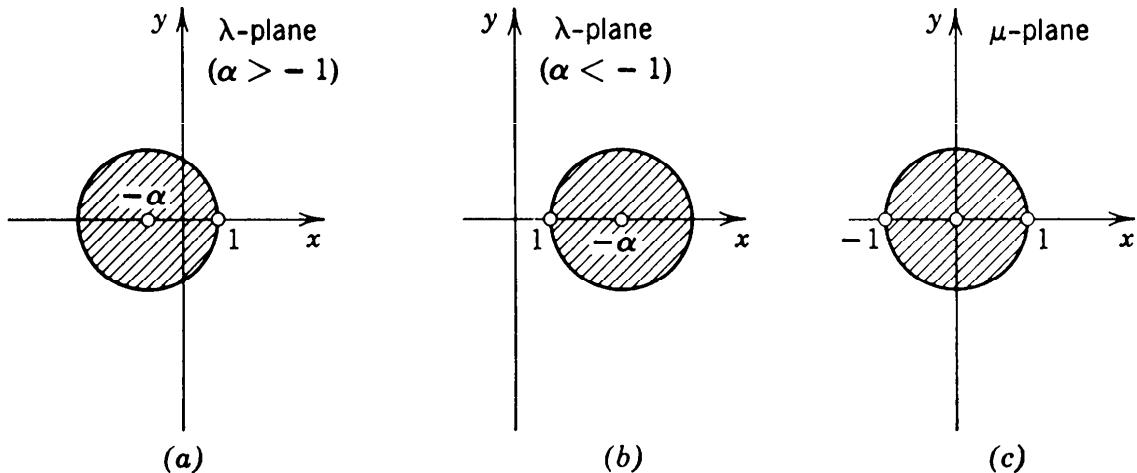


Figure 2

By an exactly analogous proof similar results can be obtained for the case where all  $\lambda_i > 1$  (see Problem 1).

In the general case, the  $\lambda_i$ , and hence also the  $\mu_j(\alpha)$  will be complex. Then the schemes (3) will be convergent if the complex numbers  $\mu_j(\alpha) \equiv \xi_j(\alpha) + i\eta_j(\alpha)$  all lie in the interior of the unit circle

$$(12) \quad |\mu|^2 \equiv \xi^2 + \eta^2 = 1,$$

of the  $(\xi, \eta)$ -plane. The relations (4) can now be considered as special points of the mapping of the  $\lambda = x + iy$  plane into the  $\mu$ -plane

$$(13) \quad \mu = \frac{\lambda + \alpha}{1 + \alpha}, \quad \alpha \neq -1.$$

This is a special case of the well-known Möbius transformations studied in function theory. If  $\alpha$  is real, we can easily verify directly that the unit

circle (12) in the  $\mu$ -plane corresponds to a circle in the  $\lambda$ -plane given by

$$(14) \quad (x + \alpha)^2 + y^2 = (1 + \alpha)^2.$$

It can also be shown that any interior point of the circle (14) is mapped by (13) into an interior point of (12). The transformation (13) is illustrated in Figure 2 for  $\alpha > -1$  and  $\alpha < -1$ .

From this figure it can be seen that convergent schemes can be found if the eigenvalues  $\lambda_i$  satisfy either

$$(15a) \quad \operatorname{Re}(\lambda_i) < 1, \quad i = 1, 2, \dots, n,$$

or

$$(15b) \quad \operatorname{Re}(\lambda_i) > 1, \quad i = 1, 2, \dots, n.$$

That is, a “convergent” value of  $\alpha$  can be obtained corresponding to any circle in the  $\lambda$ -plane which has the properties:

- (i) the center is on the real axis;
- (ii) it passes through the point  $(1, 0)$ ;
- (iii) all eigenvalues  $\lambda_i$  are interior to it.

If such a circle exists, then we call the coordinates of its center  $(-\alpha, 0)$  and this value of  $\alpha$  yields a convergent scheme. However, now it is not a simple matter to determine the best value of  $\alpha$ .

### 5.1. Practical Application of Acceleration Methods

It is assumed that the basic scheme determined by (1) can be efficiently computed. That is, to solve

$$Ax = f$$

we consider the iterates, with  $x^{(0)}$  arbitrary, given by

$$(16) \quad N_0 x^{(\nu)} = P_0 x^{(\nu-1)} + f, \quad \nu = 1, 2, \dots$$

We assume that such systems can be solved in an efficient manner. Now the iterates,  $x^{(\nu)}$ , satisfy the system of equations

$$(17) \quad x^{(\nu)} = M_0 x^{(\nu-1)} + g, \quad \nu = 1, 2, \dots,$$

where  $g \equiv N_0^{-1}f$  and  $M_0$  is defined in (2a).

The acceleration (3) corresponding to this procedure yields with  $y^{(0)}$  arbitrary

$$(18) \quad y^{(\nu)} = M(\alpha) y^{(\nu-1)} + \frac{1}{1+\alpha} g, \quad \nu = 1, 2, \dots,$$

since  $N^{-1}(\alpha)f = 1/(1 + \alpha)N_0^{-1}f$ . In terms of  $M_0$  these iterates can be written as

$$(19) \quad y^{(\nu)} = \frac{\alpha}{1+\alpha} y^{(\nu-1)} + \frac{1}{1+\alpha} (M_0 y^{(\nu-1)} + g), \quad \nu = 1, 2, \dots$$

A comparison of (17) and (19) yields some insight into the relationship between the basic scheme and its accelerated version:

If  $\alpha = 0$  the basic scheme results. If  $\alpha > 0$  then for any real numbers  $a$  and  $b$ ,

$$\min(a, b) \leq \frac{\alpha}{1 + \alpha} a + \frac{1}{1 + \alpha} b \leq \max(a, b).$$

So in this case, the acceleration scheme yields a vector on the line segment joining the previous iterate and what would have been the next iterate of the basic scheme. The term “interpolated iteration” is frequently employed to describe this type of acceleration. If  $-1 < \alpha < 0$  then

$$\frac{\alpha}{1 + \alpha} a + \frac{1}{1 + \alpha} b \begin{cases} \leq b & \text{if } b \leq a, \\ \geq b & \text{if } b \geq a; \end{cases}$$

and the acceleration scheme yields vectors with components whose values are definitely not between those of  $\mathbf{y}^{(v-1)}$  and  $M_0\mathbf{y}^{(v-1)} + \mathbf{f}$ . The scheme is now termed an “extrapolated iteration.” Similarly, the remaining case  $\alpha < -1$  is such an extrapolation method.

To compute using the scheme (19), we define the vectors  $\mathbf{z}^{(v)}$  by

$$(20a) \quad N_0 \mathbf{z}^{(v)} = P_0 \mathbf{y}^{(v-1)} + \mathbf{f}, \quad v = 1, 2, \dots,$$

and then write (18) as

$$(20b) \quad \mathbf{y}^{(v)} = \frac{\alpha}{1 + \alpha} \mathbf{y}^{(v-1)} + \frac{1}{1 + \alpha} \mathbf{z}^{(v)}.$$

Thus, as in the basic scheme, the calculations only require the solution of systems of the form (20a). [Note that in (20),  $\mathbf{z}^{(v)}$  is defined by a recursion which is similar, but not identical, to (16).]

In general, the eigenvalues  $\lambda_j$ , or in particular  $\lambda_1$  and  $\lambda_n$ , of the basic scheme will not be known. But it may be possible to approximate the value  $\alpha = \alpha_*$  which yields the fastest convergence. This is accomplished by some test calculations that are easily performed:

Since the rate of convergence is independent of the inhomogeneous term,  $\mathbf{f}$ , we seek the best scheme for solving

$$(21) \quad A\mathbf{y} = \mathbf{o}.$$

Since it is assumed that  $\det |A| \neq 0$ , this system has the unique solution  $\mathbf{y} = \mathbf{o}$ . Apply the scheme (18) with some fixed value of  $\alpha = \alpha_1$  to (21) and compute [actually use (20)]

$$(22) \quad \mathbf{y}^{(v)}(\alpha_1) = M^v(\alpha_1)\mathbf{y}^{(0)}, \quad v = 1, 2, \dots,$$

where  $\mathbf{y}^{(0)}$  is an arbitrary but fixed initial vector. If the value  $\alpha_1$  yields a convergent iteration, compute until

$$\|\mathbf{y}^{(\nu)}(\alpha_1)\| \equiv \max_j |y_j^{(\nu)}(\alpha_1)| \leq 10^{-m},$$

where  $m$  is a fixed positive number. This requires some minimum number of iterations which we may call  $\nu(\alpha_1)$ . Repeat this procedure with the same  $\mathbf{y}^{(0)}$  and  $m$ , for a sequence of values of  $\alpha = \alpha_2, \alpha_3, \dots$ , to obtain the corresponding sequence  $\{\nu(\alpha_i)\}$ . The approximate value for  $\alpha_*$  can be obtained by plotting the points  $(\nu(\alpha_i), \alpha_i)$  and choosing that  $\alpha$  which seems to minimize the “function”  $\nu(\alpha)$ .

An obvious alternative is to compute the sequence  $\{\|\mathbf{y}^{(N)}(\alpha_i)\|\}$  using a fixed number, say  $N$ , of iterations. Then an approximation to  $\alpha_*$  is that value which minimizes  $\|\mathbf{y}^{(N)}(\alpha)\|$ .

## 5.2. Generalizations of the Acceleration Method

There are numerous generalizations of the acceleration method which in fact are more powerful than the scheme described by (3). The simplest type of generalization proceeds from a single basic splitting of the form (1)–(2) but employs cyclically a fixed sequence of acceleration parameters, say  $\alpha_1, \alpha_2, \dots, \alpha_r$ . Specifically for  $i = 1, 2, \dots, r$ , define  $N(\alpha_i)$  and  $P(\alpha_i)$  as in (3) and the corresponding matrices  $M(\alpha_i)$  by

$$(23) \quad M(\alpha_i) \equiv N^{-1}(\alpha_i)P(\alpha_i) = \frac{\alpha_i}{1 + \alpha_i} I + \frac{1}{1 + \alpha_i} M_0, \quad i = 1, 2, \dots, r.$$

The iterations are defined as follows, with  $\mathbf{x}^{(0)}$  arbitrary, for  $\nu = 1, 2, \dots$ :

$$(24a) \quad \mathbf{y}^{(\nu, 0)} = \mathbf{x}^{(\nu - 1)};$$

$$(24b) \quad \mathbf{y}^{(\nu, s)} = M(\alpha_s)\mathbf{y}^{(\nu, s - 1)} + N^{-1}(\alpha_s)\mathbf{f}, \quad s = 1, 2, \dots, r;$$

$$(24c) \quad \mathbf{x}^{(\nu)} = \mathbf{y}^{(\nu, r)}.$$

Again, each of the  $r$  vectors of (24b) can be obtained by solving a linear system of the form

$$N(\alpha_s)\mathbf{y}^{(\nu, s)} = P(\alpha_s)\mathbf{y}^{(\nu, s - 1)} + \mathbf{f}.$$

With this notation, one iteration of this generalized acceleration scheme requires the same number of computations as  $r$  iterations in the ordinary acceleration scheme. The convergence of this method can be analyzed by means of the equivalent formulation

$$(25) \quad \mathbf{x}^{(\nu)} = M(\alpha_1, \alpha_2, \dots, \alpha_r)\mathbf{x}^{(\nu - 1)} + \mathbf{g}, \quad \nu = 1, 2, \dots,$$

where by (24) we find that

$$(26) \quad \begin{aligned} M(\alpha_1, \alpha_2, \dots, \alpha_r) &\equiv M(\alpha_r) \cdots M(\alpha_2)M(\alpha_1); \\ \mathbf{g} &\equiv [N^{-1}(\alpha_r) + M(\alpha_r)N^{-1}(\alpha_{r-1}) + \cdots \\ &\quad + M(\alpha_r) \cdots M(\alpha_2)N^{-1}(\alpha_1)]\mathbf{f}. \end{aligned}$$

As in the proof of (4), the eigenvalues of  $M(\alpha_1, \alpha_2, \dots, \alpha_r)$  can be determined, by using (23), in terms of the eigenvalues  $\lambda_i$  of  $M_0$ . We find in fact, that

$$(27) \quad \mu_i(\alpha_1, \alpha_2, \dots, \alpha_r) = \prod_{j=1}^r \frac{\lambda_i + \alpha_j}{1 + \alpha_j}, \quad i = 1, 2, \dots, n,$$

are the relevant eigenvalues. Now if we define the  $r$ th degree polynomial

$$(28) \quad P_r(\lambda) = \prod_{j=1}^r \frac{\lambda + \alpha_j}{1 + \alpha_j},$$

then convergence is implied by  $|P_r(\lambda_i)| < 1$  for  $i = 1, 2, \dots, n$ . In particular, if all the eigenvalues of  $M_0$  are real and lie in the interval

$$a \leq \lambda \leq b,$$

then convergence is implied by

$$|P_r(\lambda)| < 1, \quad a \leq \lambda \leq b.$$

In this case, the fastest convergence can be expected for that polynomial which has the smallest absolute magnitude in the indicated interval. Such problems are considered in Chapter 5, Section 4, and it is found that the Chebyshev polynomials can be used to find the polynomials of “least deviation from zero.” Hence, in principle, if  $a$  and  $b$  are known, the best acceleration parameters  $\alpha_1, \alpha_2, \dots, \alpha_r$  can be determined (see Problem 2).

Another type of generalization of the acceleration method is obtained by employing a sequence of different basic splittings, say

$$A = N_i - P_i, \quad i = 1, 2, \dots, r;$$

and their corresponding accelerated forms

$$N_i(\alpha_i), \quad P_i(\alpha_i).$$

An application of this technique is contained in subsection 2.2 of Chapter 9.

## PROBLEMS, SECTION 5

**1.** State and prove a theorem analogous to Theorem 1, for the case that (5) is replaced by

$$1 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_n.$$

2. If (5) is replaced by

$$-1 \leq \lambda_i \leq b < 1,$$

compare the efficiency of

- (a) the method using three acceleration parameters  $\{\alpha_i\}$  such that  $\{-\alpha_i\}$  are the zeros of the Chebyshev polynomial of third degree for the interval  $[-1, b]$ , with
- (b) the method using the single parameter  $\alpha = (1 - b)/2$ .

## 6. MATRIX INVERSION BY HIGHER ORDER ITERATIONS

The previous methods of this chapter have been primarily concerned with solving linear systems. Of course, as demonstrated in Section 0, they can all be employed to determine the inverse of any given non-singular matrix,  $A$ . We consider now an iterative method for directly computing  $A^{-1}$ . This method is a means for improving the accuracy of an approximate inverse, say  $R_0$ , obtained by other procedures. However, in many cases the present method is feasible when the initial approximate inverse is assumed to have the simple form  $R_0 = \omega I$ . Because of the large number of operations involved in matrix multiplication, these schemes are not generally used.

Assume that  $R_0$  is any approximation to  $A^{-1}$  and define the error in this approximation by

$$(1) \quad E_0 = I - AR_0.$$

Clearly, if  $R_0 = A^{-1}$  then  $E_0 = 0$ . Now with  $R_0$  as the initial approximation, we define a sequence of approximate inverses by

$$(2a) \quad R_v = R_{v-1}(I + E_{v-1} + E_{v-1}^2 + \cdots + E_{v-1}^{p-1}), \quad v = 1, 2, \dots,$$

$$(2b) \quad E_v \equiv I - AR_v, \quad v = 1, 2, \dots$$

Here,  $p$  is an arbitrary fixed integer not less than two. (This method is usually described for the case  $p = 2$  but, as will be shown, the “best” value for this integer is  $p = 3$ .) From (2) we obtain

$$(3) \quad \begin{aligned} E_v &= I - AR_v \\ &= I - AR_{v-1}(I + E_{v-1} + \cdots + E_{v-1}^{p-1}) \\ &= I - (I - E_{v-1})(I + E_{v-1} + \cdots + E_{v-1}^{p-1}) \\ &= E_{v-1}^p, \quad v = 1, 2, \dots \end{aligned}$$

Thus, in one iteration, the error matrix is raised to the  $p$ th power and the method is consequently called a  $p$ th order method. Apply (3) recursively, to find that

$$(4) \quad E_v = E_0^{p^v}, \quad v = 1, 2, \dots$$

Also from (2b) and the above we have

$$\begin{aligned}
 A^{-1} - R_v &= A^{-1}E_v \\
 (5) \qquad \qquad \qquad &= A^{-1}E_0^{p^v} \\
 &= R_0(I - E_0)^{-1}E_0^{p^v};
 \end{aligned}$$

where we have used (1) and the assumption that  $\det|I - E_0| \neq 0$ . It is now clear that the iterations converge when  $E_0$  is a convergent matrix (see Section 4).

Let us assume that  $E_0$  is a convergent matrix. Then its eigenvalues  $\lambda_i(E_0)$  satisfy  $|\lambda_i| < 1$ ,  $i = 1, 2, \dots, n$ , from Theorem 4.1. Let

$$\rho \equiv \rho(E_0) = \max_i |\lambda_i|.$$

Then, since the eigenvalues of  $E_0^p$  are  $\lambda_i^p$ , the error  $E_v$  of (4) vanishes like  $\rho^{p^v}$ .†

We now pose the problem of determining the “best” value of  $p$  to be used for any convergent  $E_0$ . By “best” we shall mean that procedure which for the least amount of computation yields an approximate inverse of desired accuracy. Alternatively, the best scheme could be defined as the one for which a given amount of computation yields the most accurate inverse. Adopting our usual convention we find from (2), since the product of two matrices requires  $n^3$  operations, that  $v$  iterations of a  $p$ th order scheme require

$$vpn^3 \text{ ops.}$$

If only  $K$  operations are to be permitted the number of iterations allowed is

$$v = \frac{K}{pn^3},$$

where we assume  $K/(pn^3)$  is an integer. Thus, the principal eigenvalue is reduced to

$$\rho^{p^v} = \rho^{(p^K/(pn^3))} = \rho^{(p^{1/p})^K/n^3}.$$

Since  $K$ ,  $n$ , and  $\rho < 1$  are independent of  $p$ , we find that the error is minimized when  $p^{1/p}$  is a maximum. Now it is easily shown that the maximum of  $x^{1/x}$  is at  $x = e = 2.718\dots$ . But a simple calculation (pointed out by M. Altman) shows that for integers  $p$  the maximum is at  $p = 3$ .

In order to apply the procedure (2), we must have an initial estimate  $R_0$  such that  $E_0 = I - AR_0$  is convergent. For a very important class of

† This is only rigorously true if the elementary divisors of  $E_0$  are simple. But, by the corollary to Theorem 1.3 of Chapter 1, the statement isn't very wrong.

matrices such an estimate is easily found. This result is contained in

**THEOREM 1.** *Let  $A$  have real eigenvalues in the interval*

$$0 < m \leq \lambda_j \leq M, \quad j = 1, 2, \dots, n.$$

*Then if  $R_0(\omega) \equiv \omega I$  and  $E_0(\omega) \equiv I - \omega A$ ,  $E_0(\omega)$  will be convergent for all  $\omega$  in*

$$(6) \quad 0 < \omega < \frac{2}{M}.$$

*Further if  $\rho(\omega)$  is the spectral radius of  $E_0(\omega)$ , i.e.,  $\rho(\omega) \equiv \rho[E_0(\omega)]$ , then*

$$(7) \quad \rho(\omega_*) = \min_{0 < \omega < 2/M} \rho(\omega) = \frac{M - m}{M + m}, \quad \omega_* = \frac{2}{M + m}.$$

*Proof.* This theorem is essentially a restatement of Theorem 5.1. If we make the association  $(x, m_j) \leftrightarrow (\omega, -\lambda_j)$  in the proof of that theorem, the above follows. ■

## PROBLEMS, SECTION 6

1. Newton's method for improving  $R_0$ , the approximate inverse of  $A$ , is formally obtained by setting

$$\begin{aligned} A &= (R_0 + \delta R_0)^{-1} \\ &= [R_0(I + R_0^{-1}\delta R_0)]^{-1} \\ &= (I + R_0^{-1}\delta R_0)^{-1}R_0^{-1}. \end{aligned}$$

Therefore,

$$AR_0 = (I + R_0^{-1}\delta R_0)^{-1} \simeq I - R_0^{-1}\delta R_0.$$

Solve for  $\delta R_0$ . Does this formula fit into the iteration scheme (2) for  $p = 2$ ?

2. Show that if  $A$  is non-singular, the choice  $R_0 \equiv aA^*$  with  $a \equiv 1/\text{tr}(AA^*)$  produces a convergent matrix  $E_0$  in (1).

# 3

## Iterative Solution of Non-Linear Equations

### 0. INTRODUCTION

In this chapter, we consider iterative methods for determining the roots of equations

$$\mathbf{f}(\mathbf{x}) = \mathbf{0}$$

where  $\mathbf{f}$  and  $\mathbf{x}$  are vectors of the same dimension  $k$ : i.e., if  $k = 1$  we have a single equation; if  $k = n$  we have a system of  $n$  equations. Most of the iterative methods can be written in the form  $\mathbf{x}_{n+1} = \mathbf{g}(\mathbf{x}_n)$  for some suitable function  $\mathbf{g}$  and initial approximation  $\mathbf{x}_0$ . The convergence of this iteration process is assured if the mapping  $\mathbf{g}(\mathbf{x})$  carries a closed and bounded set  $S \subset C_k$  into itself and if the *mapping is contracting*, i.e., if  $\|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq M \|\mathbf{x} - \mathbf{y}\|$  for some norm, “Lipschitz” constant  $M < 1$ , and all  $\mathbf{x}, \mathbf{y} \in S$ . Such an iteration scheme is sometimes referred to as the *Picard iteration method*, or as a *functional iteration method*. It can be easily shown under these conditions that  $\mathbf{g}(\mathbf{x})$  has a unique *fixed point*  $\alpha$  in  $S$  satisfying

$$\alpha = \mathbf{g}(\alpha).$$

We shall study this contracting mapping theorem in one or more dimensions and the related results which are basic to many of the iterative methods of this chapter.

Usually the iterative methods are valid for real and complex roots. However, in the latter case complex arithmetic must be incorporated into the appropriate digital computer codes and the initial estimate of the root must usually be complex (see Subsection 4.4 for an exception). The iterative methods require at least one initial estimate or guess at the location of the root being sought. If this initial estimate, say  $\mathbf{x}_0$ , is “sufficiently close” to a root, then, in general, the procedures will converge.

The problem of how to obtain such a “good”  $\mathbf{x}_0$  is unresolved in general. Frequently, a good estimate of the root is known to the problem formulator (i.e., the engineer, physicist, mathematician, or other scientist who is interested in the solution) or can be found by an analytical study. For many purposes merely graphical accuracy (about two decimal figures) is needed for the initial value. In these cases, one may tabulate the function and plot the data in one or two variables or “fit” linear forms  $a_{i0} + \sum_{j=1}^k a_{ij}x_j$  to  $f_i(\mathbf{x})$  to find the approximate starting values. If a digital computer is to be employed, this plotting method is quite convenient since all of the required function evaluations will be contained in the eventual machine code for the problem.

As a general empirical rule, the schemes which converge more rapidly (i.e., higher order methods) require closer initial estimates. In practice, these higher order schemes may require the use of more significant digits in order that they converge as theoretically predicted. Thus, it is frequently a good idea to use a simple method to start with and then, when fairly close to the root, to use some higher order method for just a few iterations.

For polynomial equations in one variable we know much about the roots. While the general iteration schemes apply to them there are also special methods which can be used to obtain the zeros of polynomials. Such considerations are to be found in Section 4.

## 1. FUNCTIONAL ITERATION FOR A SINGLE EQUATION

Let us consider a scalar equation of the special form

$$(1) \quad x - g(x) = 0, \quad \text{or} \quad x = g(x).$$

[It is clear that any equation  $f(x) = 0$  can be written equivalently in this form by defining  $g(x) \equiv x - f(x)$ .] If  $x_0$  is some initial estimate of a root of (1), a scheme naturally suggested is to form the sequence

$$(2) \quad x_{\nu+1} = g(x_{\nu}), \quad \nu = 0, 1, \dots$$

An important result concerning the convergence of this procedure and a proof of the existence of a unique root is contained in

**THEOREM 1.** *Let  $g(x)$  satisfy the Lipschitz condition*

$$(3a) \quad |g(x) - g(x')| \leq \lambda|x - x'|,$$

*for all values  $x, x'$  in the closed† interval  $I \equiv [x_0 - \rho, x_0 + \rho]$  where the*

† Unless otherwise specified:  $[a, b]$  denotes the *closed* interval,  $a \leq x \leq b$ ;  $(a, b)$  denotes the *open* interval,  $a < x < b$ ;  $(a, b]$  and  $[a, b)$  denote respectively the *half-open* intervals  $a < x \leq b$  and  $a \leq x < b$ .

*Lipschitz constant,  $\lambda$ , satisfies*

$$(3b) \quad 0 \leq \lambda < 1.$$

*Let the initial estimate,  $x_0$ , be such that*

$$(4) \quad |x_0 - g(x_0)| \leq (1 - \lambda)\rho.$$

*Then*

(i) *all the iterates  $x_v$ , defined by (2), lie within the interval  $I$ ; i.e.,*

$$(5) \quad x_0 - \rho \leq x_v \leq x_0 + \rho,$$

(ii) *(existence) the iterates converge to some point, say,*

$$\lim_{v \rightarrow \infty} x_v = \alpha, \quad (\text{in fact, } |x_v - \alpha| \leq \lambda^v \rho)$$

*which is a root of (1), and*

(iii) *(uniqueness)  $\alpha$  is the only root in  $[x_0 - \rho, x_0 + \rho]$ .*

*Proof.* We prove (i) by induction. Since  $x_1 = g(x_0)$ , we have by (3b) and (4)

$$(6) \quad |x_0 - x_1| \leq (1 - \lambda)\rho \leq \rho$$

and hence  $x_1$  is in the interval (5). Assume this true for the iterates  $x_1, x_2, \dots, x_v$ . Then from (2)

$$|x_{v+1} - x_v| = |g(x_v) - g(x_{v-1})|$$

and by the inductive assumption  $x_v$  and  $x_{v-1}$  are in the interval (5). Thus, by (3a), the Lipschitz condition yields

$$\begin{aligned} (7) \quad |x_{v+1} - x_v| &\leq \lambda|x_v - x_{v-1}| \\ &\leq \lambda^2|x_{v-1} - x_{v-2}| \\ &\vdots \\ &\leq \lambda^v|x_1 - x_0| \\ &\leq \lambda^v(1 - \lambda)\rho. \end{aligned}$$

Here we have used (2) and (3a) recursively and then applied (6). However,

$$\begin{aligned} |x_{v+1} - x_0| &= |(x_{v+1} - x_v) + (x_v - x_{v-1}) + \cdots + (x_1 - x_0)| \\ &\leq |x_{v+1} - x_v| + |x_v - x_{v-1}| + \cdots + |x_1 - x_0| \\ &\leq (\lambda^v + \lambda^{v-1} + \cdots + 1)(1 - \lambda)\rho = (1 - \lambda^{v+1})\rho \\ &\leq \rho, \end{aligned}$$

which completes the proof of (i).

To prove part (ii), we first show that the sequence  $\{x_v\}$  is a Cauchy sequence. Thus, for arbitrary positive integers  $m$  and  $p$ , we consider

$$\begin{aligned}
 |x_m - x_{m+p}| &= |(x_m - x_{m+1}) + (x_{m+1} - x_{m+2}) + \dots \\
 &\quad + (x_{m+p-1} - x_{m+p})| \\
 (8) \quad &\leq |x_m - x_{m+1}| + |x_{m+1} - x_{m+2}| + \dots \\
 &\quad + |x_{m+p-1} - x_{m+p}| \\
 &\leq (\lambda^m + \lambda^{m+1} + \dots + \lambda^{m+p-1})(1 - \lambda)\rho \\
 &\leq (1 - \lambda^p)\rho\lambda^m.
 \end{aligned}$$

Here we have used the inequalities (7) which are valid since (i) has been proved. Now given any  $\epsilon > 0$ , since  $\lambda$  in  $0 \leq \lambda < 1$  is fixed, we can find an integer  $N(\epsilon)$  such that  $|x_m - x_{m+p}| < \epsilon$  for all  $m > N(\epsilon)$  and  $p > 0$  (we need only take  $N$  such that  $\lambda^N < \epsilon/\rho$ ). Hence the sequence  $\{x_v\}$  is a Cauchy sequence and has a limit, say  $\alpha$ , in  $I$ . Since the function  $g(x)$  is continuous in the interval  $I$ , the sequence  $\{g(x_v)\}$  has the limit  $g(\alpha)$  and by (2) this limit must also be  $\alpha$ ; that is,  $\alpha = g(\alpha)$ . Now  $|x_v - \alpha| = |g(x_{v-1}) - g(\alpha)| \leq \lambda|x_{v-1} - \alpha|$ ; hence  $|x_v - \alpha| \leq \lambda^v|x_0 - \alpha| \leq \rho\lambda^v$ .

For part (iii), the uniqueness, let  $\beta$  be another root in  $[x_0 - \rho, x_0 + \rho]$ . Then, since  $\alpha$  and  $\beta$  are both in this interval, (3) holds and we have, if  $|\alpha - \beta| \neq 0$ ,

$$|\alpha - \beta| = |g(\alpha) - g(\beta)| \leq \lambda|\alpha - \beta| < |\alpha - \beta|.$$

This contradiction implies that  $\alpha = \beta$  and the proof of the theorem is concluded. ■

**COROLLARY.** *If  $|g'(x)| \leq \lambda < 1$  for  $|x - x_0| \leq \rho$  and (4) is satisfied, then the conclusion of Theorem 1 is valid.*

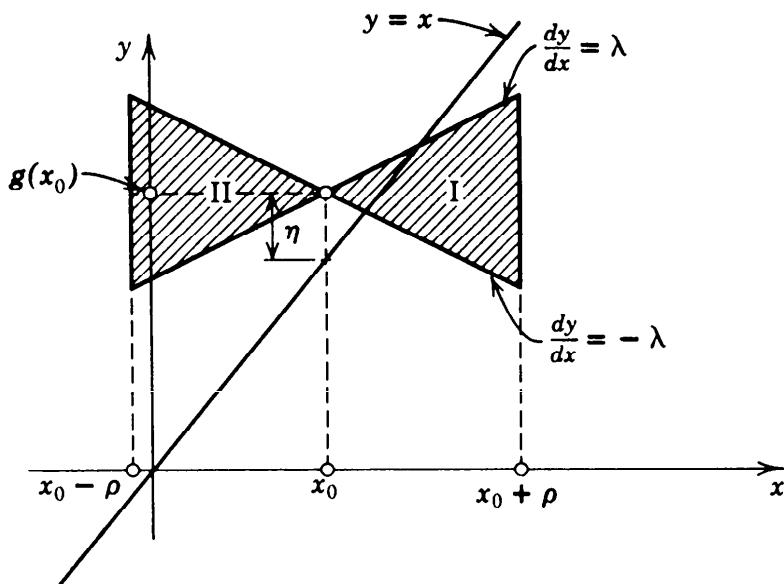


Figure 1

*Proof.* The mean value theorem implies  $g(x_1) - g(x_2) = g'(\xi)(x_1 - x_2)$ . Whence  $\lambda$  may serve as the Lipschitz constant in (3a and b). ■

A geometric interpretation of Theorem 1 is suggested by Figure 1. This illustrates the case with  $g(x_0) - x_0 = \eta > 0$  and the triangles I and II, determined by lines with slope  $\pm\lambda$  through  $(x_0, g(x_0))$ , are the regions in which the values of  $g(x)$  lie for  $x_0 - \rho \leq x \leq x_0 + \rho$ . It is easy to verify that

- (a) if  $\lambda \geq 1$ , the line  $y = x$  will not intersect the upper boundary of triangle I or if
- (b)  $\lambda < 1$  and  $\eta > (1 - \lambda)\rho$  that the line  $y = x$  will not intersect the upper boundary of triangle I and hence may not intersect an admissible function  $g(x)$  in the interval  $[x_0 - \rho, x_0 + \rho]$ .

In other words, the conditions  $\lambda < 1$ ,  $|\eta| \leq (1 - \lambda)\rho$  are necessary to insure the existence of a root for every function  $g(x)$  satisfying conditions (3a and b).

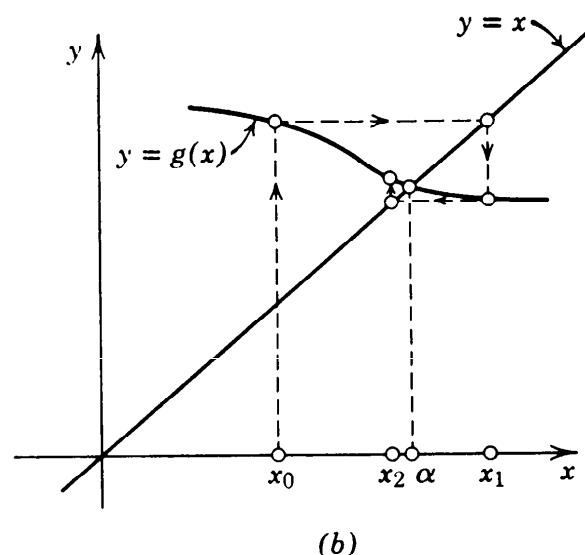
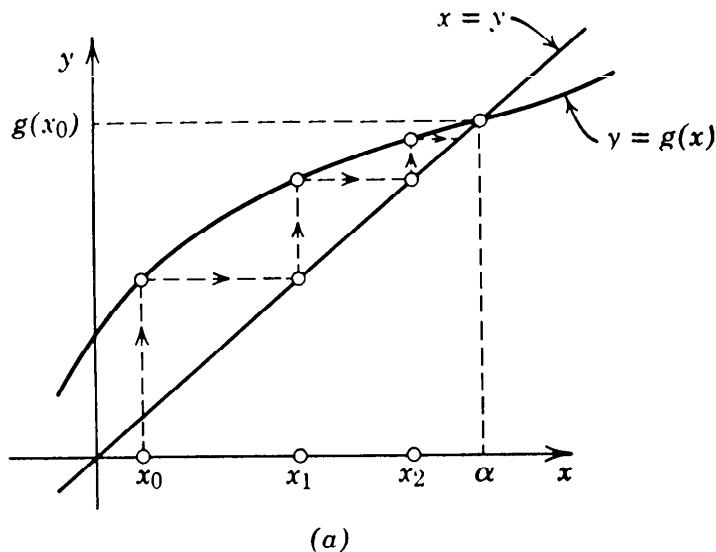


Figure 2

Figures 2a and 2b illustrate convergent iterative sequences for functions  $g(x)$  with positive and negative slope, respectively. Note that the sequence  $\{x_n\}$  converges to  $\alpha$  monotonically for  $g(x)$  of positive slope and converges with values alternately above and below  $\alpha$  for  $g(x)$  of negative slope.

Another convergence theorem is

**THEOREM 2.** *If  $x = g(x)$  has a root at  $x = \alpha$  and in the interval*

$$(9a) \quad |x - \alpha| < \rho$$

*$g(x)$  satisfies*

$$(9b) \quad |g(x) - g(\alpha)| \leq \lambda|x - \alpha|,$$

*with  $\lambda < 1$ , then for any  $x_0$  in (9a):*

- (i) *all the iterates  $x_v$  of (2) lie in the interval (9a),*
- (ii) *the iterates  $x_v$  converge to  $\alpha$ ,*
- (iii) *the root  $\alpha$  is unique in this interval.*

*Proof.* Part (i) is again proved by induction. By hypothesis  $x_0$  is in (9a) and we assume  $x_1, x_2, \dots, x_{v-1}$  are also. Then since  $\alpha = g(\alpha)$  we have from (2)

$$(10a) \quad \begin{aligned} |\alpha - x_v| &= |g(\alpha) - g(x_{v-1})| \\ &\leq \lambda|\alpha - x_{v-1}|, \end{aligned}$$

whence  $\lambda < 1$  implies (i). Furthermore,

$$(10b) \quad \begin{aligned} |\alpha - x_v| &\leq \lambda|\alpha - x_{v-1}|, \\ &\leq \lambda^2|\alpha - x_{v-2}|, \\ &\vdots \\ &\leq \lambda^v|\alpha - x_0|. \end{aligned}$$

By letting  $v \rightarrow \infty$ , we see that  $x_v \rightarrow \alpha$ , since  $\lambda < 1$ . The uniqueness follows as in Theorem 1. ■

Notice that condition (9b) is weaker than the general Lipschitz condition for the interval (9a), since the one point  $\alpha$  is fixed. This feature is applicable in Problem (1).

We can now prove a corollary (with a hypothesis which is oftentimes more readily verifiable).

**COROLLARY.** *If we replace (9b) by*

$$(9b)' \quad |g'(x)| \leq \lambda < 1,$$

*then the conclusions (i), (ii), and (iii) follow.*

*Proof.* From the mean value theorem and (2)

$$(10a)' \quad \alpha - x_v = g(\alpha) - g(x_{v-1}) = g'(\xi_{v-1})(\alpha - x_{v-1}).$$

Hence (10a) follows from (9b)'. Therefore (10b) and the rest of the proof of Theorem 2 apply. ■

It is clear from (10a)' that, if the iterations converge,  $\xi_v \rightarrow \alpha$  and thus “asymptotically” (as  $v \rightarrow \infty$ )

$$(11) \quad |\alpha - x_{k+v}| \approx |g'(\alpha)|^v |\alpha - x_k|,$$

for large enough  $k$ . The quantity

$$(12) \quad \hat{\lambda} = |g'(\alpha)|$$

is frequently called the (asymptotic) *convergence factor* and in analogy with the iterative solution of linear systems

$$(13) \quad R \equiv \log \frac{1}{\hat{\lambda}}$$

may be called the rate of convergence (if  $\hat{\lambda} < 1$ ). The number of additional iterations required to reduce the error at the  $k$ th step by the factor  $10^{-m}$  is then, asymptotically,

$$(14) \quad v = \frac{m}{R}.$$

We assume, in these definitions, that  $\hat{\lambda} = |g'(\alpha)| \neq 0$  and define such an iteration scheme (2) to be a *first order method*; higher order methods are considered in Subsection 1.2.

### 1.1. Error Propagation

In actual computations it may not be possible, or practical, to evaluate the function  $g(x)$  exactly (i.e., only a finite number of decimals may be retained after rounding or  $g(x)$  may be given as the numerical solution of a differential equation, etc.). For any value  $x$  we may then represent our approximation to  $g(x)$  by  $G(x) = g(x) + \delta(x)$  where  $\delta(x)$  is the error committed in evaluating  $g(x)$ . Frequently we may know a bound for  $\delta(x)$ , i.e.,  $|\delta(x)| < \delta$ . Thus the actual iteration scheme which is used may be represented as

$$(15) \quad X_{v+1} = g(X_v) + \delta_v, \quad v = 0, 1, 2, \dots,$$

where the  $X_v$  are the numbers obtained from the calculations and the  $\delta_v \equiv \delta(X_v)$  satisfy

$$(16) \quad |\delta_v| \leq \delta, \quad v = 0, 1, \dots$$

We cannot expect the computed iterates  $X_v$  of (15) to converge. However, under proper conditions, it should be possible to approximate a root to an accuracy determined essentially by the accuracy of the computations,  $\delta$ .

For example, from Figure 3 it is easy to see that for the special case of  $g(x) \equiv \alpha + \lambda(x - \alpha)$ , the uncertainty in the root  $\alpha$  is bounded by  $\pm \delta/(1 - \lambda)$ . We note that, if the slope  $\lambda$  is close to unity the problem is not "properly posed." We now establish Theorem 3 which states quite generally that when the functional iteration scheme is convergent, the presence of errors in computing  $g(x)$ , of magnitudes bounded by  $\delta$ , causes the scheme to estimate the root  $\alpha$  with an uncertainty bounded by  $\pm \delta/(1 - \lambda)$ .

**THEOREM 3.** *Let  $x = g(x)$  satisfy the conditions of Theorem 2. Let  $X_0$  be any point in the interval*

$$(17a) \quad |\alpha - x| \leq \rho_0,$$

where

$$(17b) \quad 0 < \rho_0 \leq \rho - \frac{\delta}{1 - \lambda}.$$

Then the iterates  $X_v$  of (15), with the errors bounded by (16), lie in the interval

$$|\alpha - X_v| \leq \rho,$$

and

$$(18) \quad |\alpha - X_k| \leq \frac{\delta}{1 - \lambda} + \lambda^k \left( \rho_0 - \frac{\delta}{1 - \lambda} \right),$$

where  $\lambda^k \rightarrow 0$  as  $k \rightarrow \infty$ .

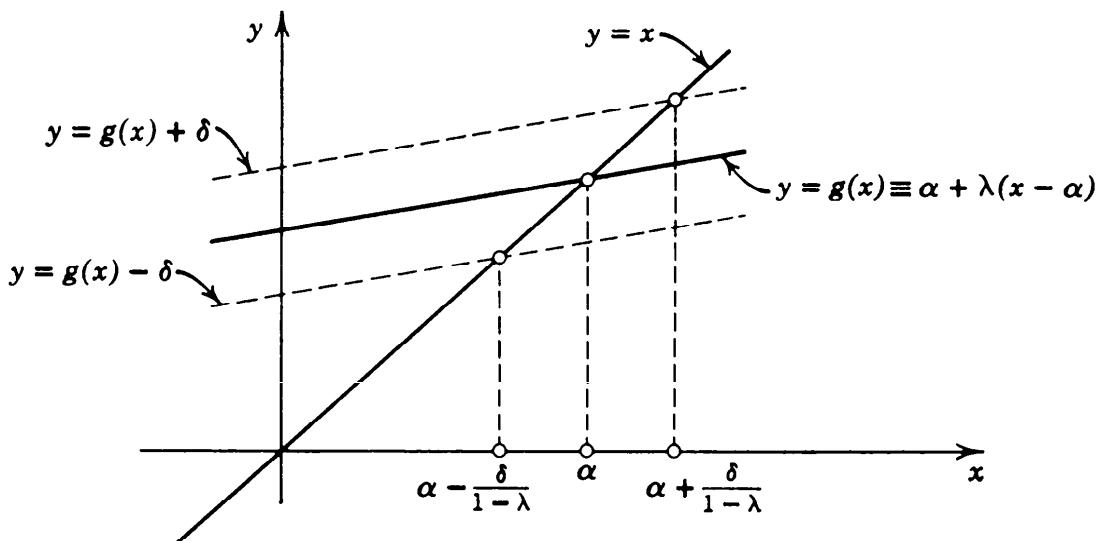


Figure 3

*Proof.* It is clear that  $|\alpha - X_0| \leq \rho_0 \leq \rho_0 + \delta/(1 - \lambda) \leq \rho$ . Then for an inductive proof assume  $X_0, X_1, \dots, X_{v-1}$  are in  $|\alpha - x| \leq \rho$ . By (15) and (16)

$$|\alpha - X_v| \leq |[g(\alpha) - g(X_{v-1})] - \delta_{v-1}| \leq |g(\alpha) - g(X_{v-1})| + \delta.$$

From (9b), we then have

$$\begin{aligned} |\alpha - X_v| &\leq \lambda|\alpha - X_{v-1}| + \delta \\ &\leq \lambda^2|\alpha - X_{v-2}| + \lambda\delta + \delta \\ &\leq \lambda^3|\alpha - X_{v-3}| + \lambda^2\delta + \lambda\delta + \delta \\ &\vdots \\ &\leq \lambda^v|\alpha - X_0| + \lambda^{v-1}\delta + \dots + \lambda\delta + \delta \\ &\leq \lambda^v\rho_0 + \frac{1 - \lambda^v}{1 - \lambda}\delta \\ &\leq \lambda^v\rho_0 + \frac{\delta}{1 - \lambda} - \lambda^v \frac{\delta}{1 - \lambda} \\ &\leq \rho_0 + \frac{\delta}{1 - \lambda} \\ &\leq \rho. \end{aligned}$$

Thus all iterates lie in  $|\alpha - x| \leq \rho$  and the iteration process is defined. Moreover, from the last inequality involving  $v$  we find the estimate (18) which completes the proof. ■

Theorem 3 shows that the method is “as convergent as possible,” that is, the computational errors which arise from the evaluations of  $g(x)$  may cumulatively produce an error of magnitude at most  $\delta/(1 - \lambda)$ . It is also clear that such errors limit the size of the error bound *independently of the number of iterations*. Thus in actual calculations, it is not worthwhile to iterate until  $\lambda^v\rho_0 \ll \delta/(1 - \lambda)$ . In fact, if reasonable estimates of  $\lambda$ ,  $\delta$ , and  $\rho_0$  are known, it is an efficient procedure to have the two types of error term in (18) of the same magnitude; i.e.,

$$\lambda^v\rho_0 \cong \frac{\delta}{1 - \lambda}. \dagger$$

The number of iterations required is then about

$$\begin{aligned} v &\cong \log \left[ \frac{\delta}{(1 - \lambda)\rho_0} \right] / \log \lambda \\ (19) \quad &\cong \log \left[ \frac{(1 - \lambda)\rho_0}{\delta} \right] / \log \frac{1}{\lambda}. \end{aligned}$$

†  $\cong$  reads “approximately equals” and we have tacitly assumed that  $\delta/(1 - \lambda) \ll \rho_0$ .

Of course, if the acceptable error is much greater than  $\delta/(1 - \lambda)$  the number of iterations given by (13) and (14) is the relevant estimate. It is essential to estimate  $\delta$  from the arithmetic calculations involved in evaluating  $g(x)$ . Also, note that since the  $X_v$  need not converge any tests for the termination of the iterations should allow for this roundoff effect.

### 1.2. Second and Higher Order Iteration Methods

It is clear from the corollary to Theorem 2, that if at the root  $x = \alpha$ ,

$$(20) \quad g'(\alpha) = 0,$$

then the convergence should be quite rapid. Let this be the case and assume further that  $g''(x)$  exists and is bounded in some interval,  $|\alpha - x| \leq \rho$ , in which (9) is satisfied. Then for any  $x$  in this interval we have by Taylor's theorem:

$$\begin{aligned} g(x) &= g(\alpha) + 0 + \frac{(x - \alpha)^2}{2} g''(\xi), \\ &= \alpha + \frac{(x - \alpha)^2}{2} g''(\xi). \end{aligned}$$

Here  $\xi$  is some value between  $x$  and  $\alpha$ . By using this result we obtain for any iterate (2) (assuming  $|\alpha - x_0| < \rho$  and  $\frac{1}{2}\rho|g''(x)| \leq \lambda < 1$ ):

$$(21) \quad |x_v - \alpha| = |g(x_{v-1}) - g(\alpha)| = |\frac{1}{2}g''(\xi_{v-1})| \cdot |x_{v-1} - \alpha|^2, \quad v = 1, 2, 3, \dots$$

Thus, the error in any iterate is proportional to the *square* of the previous error and if  $g''(\alpha) \neq 0$  the procedure (2) is now called a *second order method*.

Let the bound on  $g''(x)$  be denoted by

$$(22) \quad |g''(x)| \leq M, \quad |\alpha - x| < \rho.$$

Then from (21)

$$\begin{aligned} (23) \quad |x_v - \alpha| &\leq M|x_{v-1} - \alpha|^2 \\ &\leq M \cdot M^2|x_{v-2} - \alpha|^4 \\ &\leq M \cdot M^2 \cdot M^4|x_{v-3} - \alpha|^8 \\ &\vdots \\ &\leq M^{(2^{v-1})}|x_0 - \alpha|^{2^v} \\ &\leq (M|x_0 - \alpha|)^{2^{v-1}}|x_0 - \alpha|. \end{aligned}$$

Thus, if  $M|x_0 - \alpha| < 1$ , the second order method converges and reduces the initial error by at least  $10^{-m}$  when

$$(M|x_0 - \alpha|)^{2^{v-1}} \cong 10^{-m}.$$

The number of iterations required is now obtained from

$$(24) \quad \begin{aligned} 2^\nu &\cong \frac{-m}{\log(M|x_0 - \alpha|)}, \\ \nu &\cong \frac{1}{\log 2} \log \frac{m}{\log 1/(M|x_0 - \alpha|)}. \end{aligned}$$

A comparison with first order schemes is possible, i.e., the estimates in (12)–(14), if we assume  $\lambda = M|x_0 - \alpha|$ . That is, by letting  $\nu_{(1)}$  and  $\nu_{(2)}$  represent the exponents  $\nu$  in (10b) and (23), respectively, we have equal reduction of the error if

$$(25) \quad 2^{\nu_{(2)}} = 1 + \nu_{(1)}.$$

For instance, 130 iterations of the first order scheme are equivalent, under the above assumptions, to about 7 iterations of a second order scheme! A further striking property of second order schemes can be obtained by assuming for all  $\nu \geq \nu_0$  that

$$|x_\nu - \alpha| = 10^{-p_\nu}, \quad p_\nu > 0;$$

i.e.,  $p_\nu$  is essentially the number of correct decimals in the  $\nu$ th iterate. Then from the first line of (23):

$$10^{-p_\nu} \leq M 10^{-2p_{\nu-1}},$$

and upon taking the logarithm of both sides

$$(26) \quad p_\nu \geq 2p_{\nu-1} - \log M.$$

Thus, if  $M < 1$ , then  $-\log M > 0$  and the number of correct decimals *more than doubles* on each iteration. (If  $M > 1$  the number does not quite double but, since  $p_\nu \gg \log M$  for large  $\nu$ , this doubling is at least asymptotically true.)

Schemes which are more quickly convergent than second order ones are now easily described. Let us assume that at a root  $x = \alpha$  of (1):

$$(27a) \quad g'(\alpha) = g''(\alpha) = \dots = g^{(n-1)}(\alpha) = 0;$$

$$(27b) \quad g^{(n)}(\alpha) \neq 0; \quad |g^{(n)}(x)| \leq n!M \quad \text{in} \quad |x - \alpha| \leq \rho.$$

Then by Taylor's theorem,

$$g(x) = \alpha + \frac{(x - \alpha)^n}{n!} g^{(n)}(\xi), \quad |x - \alpha| < \rho$$

where  $\xi$  is between  $x$  and  $\alpha$ . Again from (2) and the above

$$(28) \quad \begin{aligned} |x_{v+1} - \alpha| &= \frac{|g^{(n)}(\xi_v)|}{n!} |x_v - \alpha|^n \\ &\leq M \cdot |x_v - \alpha|^n. \end{aligned}$$

The method (2) under conditions (27) is now called an  $n$ th order procedure and one can easily deduce the results corresponding to (23)–(26) for such methods. In the event that  $g(x)$  is calculated with an error of magnitude  $\delta$  as in (15), the root  $\alpha$  may be determined only to within an uncertainty of at best  $\pm \delta$ . This conclusion follows by letting  $\lambda \rightarrow 0$  in (18).

## PROBLEMS, SECTION 1

1. Given  $g(x) \equiv x^2 - 2x + 2$ . For what values  $x_0$  does (2) converge?

[Hint: Use Theorem 2.] What is the order of the convergence? Sketch a graph analogous to Figures 2a and b.

2. For  $g(x) \equiv \cos x$ , show that  $x_{n+1} = g(x_n)$  defines a convergent sequence for arbitrary  $x_0$ . Calculate the root  $\alpha = \cos \alpha$  to three decimal places.

## 2. SOME EXPLICIT ITERATION PROCEDURES

The general problem to which the previous iteration methods are to be applied is that of finding the root (or roots) of

$$(1) \quad f(x) = 0$$

in some interval, say  $a \leq x \leq b$ . Let  $\phi(x)$  be any function such that

$$(2) \quad 0 < |\phi(x)| < \infty, \quad a \leq x \leq b.$$

Then the equation

$$(3) \quad x = g(x) \equiv x - \phi(x)f(x),$$

has roots which coincide with those of (1) in the interval  $[a, b]$  and no others. Many of the standard iterative methods are obtained for special choices of  $\phi(x)$ .

Another procedure for defining the function  $g(x)$  is to use

$$(4) \quad g(x) \equiv x - F(f(x)),$$

where  $F(y)$  is a function such that

$$F(0) = 0; \quad F(y) \neq 0, \quad y \neq 0.$$

Such methods more naturally describe many higher order schemes.

### 2.1. The Simple Iteration or Chord Method (First Order)

The simplest choice for  $\phi(x)$  in (3) is to take

$$(5) \quad \phi(x) \equiv m \neq 0.$$

If  $f(x)$  is differentiable, we note that

$$(6) \quad g'(x) = 1 - mf'(x),$$

and the scheme will be convergent, by the corollary to Theorem 1.2, in some interval about  $\alpha$  provided that  $m$  is chosen such that

$$(7) \quad 0 < mf'(\alpha) < 2.$$

Thus  $m$  must have the same sign as  $f'(\alpha)$ , while if  $f'(\alpha) = 0$ , (7) cannot be satisfied.

The choice (5) yields the iteration equations

$$x_{v+1} = x_v - mf(x_v).$$

These iterates have a geometric realization in which the value  $x_{v+1}$  is the  $x$  intercept of the line with slope  $1/m$  through  $(x_v, f(x_v))$ . (See Figure 1.) The inequality (7) implies that this slope should be between  $\infty$  (i.e., vertical) and  $\frac{1}{2}f'(\alpha)$  [i.e., half the slope of the tangent to the curve  $y = f(x)$  at the root]. It is from this geometric description that the name chord method is derived—the next iterate is determined by a chord of constant slope joining a point on the curve to the  $x$ -axis.

### 2.2. Newton's Method (Second Order)

If the slope of the chord is changed at each iteration so that

$$(8) \quad g'(x_v) = 1 - m_v f'(x_v) = 0,$$

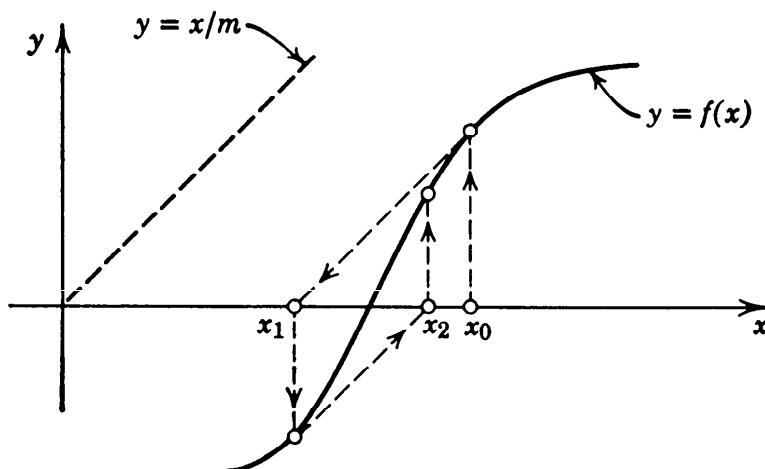


Figure 1

then a second order procedure may be obtained. From (8) we find

$$(9) \quad m_v = \frac{1}{f'(x_v)},$$

which suggests the choice in (3) of

$$(10) \quad \phi(x) = \frac{1}{f'(x)}, \quad \text{or} \quad g(x) \equiv x - \frac{f(x)}{f'(x)}.$$

The resulting iteration procedure is now

$$(11) \quad x_{v+1} = x_v - \frac{f(x_v)}{f'(x_v)},$$

and it is at least of second order if  $f'(\alpha) \neq 0$  and  $f''(x)$  exists, since

$$(12) \quad g'(\alpha) = \frac{f(\alpha)f''(\alpha)}{[f'(\alpha)]^2} = 0.$$

The geometrical interpretation of the scheme (11) simply replaces the chord in Figure 1 by the tangent line to  $y = f(x)$  at  $(x_v, f(x_v))$ .

In applying Newton's method, we are required to evaluate  $f'(x_v)$  as well as  $f(x_v)$  at each step of the procedure. For sufficiently simple functions, which are given explicitly, this may offer no serious difficulty. (This is especially true for polynomials whose derivatives are easily evaluated by synthetic division; see Subsection 4.1.) However, if  $f(x)$  is known only implicitly (say as the solution of some differential equation in which  $x$  is a parameter in the initial data), it may be impractical to evaluate  $f'(x_v)$  at each iteration. In such cases the derivative may be approximated by various methods, the most obvious approximation being

$$(13) \quad f'(x_v) \cong \frac{f(x_v) - f(x_{v-1})}{x_v - x_{v-1}}.$$

If this approximation is used, the procedure is no longer Newton's method but is the method of *false position* discussed in the next subsection.

A useful observation on the application of Newton's method, or the false position variation of it, is based on the fact that as the iterations converge,  $f'(x_v)$  or its approximations converge to  $f'(\alpha)$ . Thus, for all iterates  $v \geq v_0$ , say, it may suffice to use  $f'(x_{v_0})$  in place of  $f'(x_v)$  in (11). The iteration method from this point on is then just the chord method with  $1/m = f'(x_{v_0})$ .

It should be noted that Newton's method may be undefined and condition (2) violated if  $f'(x) = 0$  for some  $x$  in  $[a, b]$ . In particular, if at the

root  $x = \alpha$ ,  $f'(\alpha) = 0$ , the procedure may no longer be of second order since (12) is not satisfied. To examine this case we assume that

$$(14a) \quad f(x) \equiv (x - \alpha)^p h(x), \quad p > 1$$

where the function  $h(x)$  has a second derivative and

$$(14b) \quad h(\alpha) \neq 0.$$

From (14) in (10) we find that

$$g'(x) = \frac{\left(1 - \frac{1}{p}\right) + (x - \alpha) \frac{2h'(x)}{ph(x)} + (x - \alpha)^2 \frac{h''(x)}{p^2 h(x)}}{\left[1 + (x - \alpha) \frac{h'(x)}{ph(x)}\right]^2}.$$

Thus for  $x_0$  sufficiently close to  $\alpha$  we have  $|g'(x)| < 1$  for  $x \in [x_0, \alpha]$  and the iterations (11) will converge. The asymptotic convergence factor is now

$$|g'(\alpha)| = 1 - \frac{1}{p}.$$

So *only in the case of a linear root, i.e.,  $p = 1$ , is Newton's method second order, but it will converge as a first order method in the general case (14).* If the order of the root,  $p$ , is known (or can be closely estimated) quadratic convergence can be retained or approximated by the modification

$$g(x) \equiv x - p \frac{f(x)}{f'(x)}.$$

The details of this procedure are left to the reader. A convergence proof for Newton's method which *does not require  $f'(\alpha) \neq 0$*  is contained in Theorem 3.3 [see also Problems (3)–(6) of Section 3].

### 2.3. Method of False Position (Fractional Order)

If the difference quotient approximation to the derivative, given by (13), is employed in (11) we obtain the iterative procedure:

$$(15) \quad x_{v+1} = x_v - f(x_v) \frac{x_v - x_{v-1}}{f(x_v) - f(x_{v-1})}; \quad v = 1, 2, \dots$$

It should be noted that two successive iterates,  $x_0$  and  $x_1$ , must be estimated before the recursion formula can be used. However, only one function evaluation,  $f(x_v)$ , is required at each step since the previous value,  $f(x_{v-1})$ , may be retained. [This is an advantage over Newton's method where two evaluations,  $f(x_v)$  and  $f'(x_v)$ , are required.] The order of this procedure cannot be deduced by the analysis of Section 1 since (15)

cannot be written in the scalar form  $x_{v+1} = g(x_v)$ . To examine this question let  $x = \alpha$  be a root of  $f(x) = 0$ . Then we may write, by subtracting each side of (15) from  $\alpha$ ,

$$\begin{aligned}\alpha - x_{v+1} &= (\alpha - x_v) + f(x_v) \frac{x_v - x_{v-1}}{f(x_v) - f(x_{v-1})} \\ &= (\alpha - x_v) \frac{f[x_{v-1}, x_v] - f[x_v, \alpha]}{f[x_{v-1}, x_v]}\end{aligned}$$

where we define  $f[a, b] \equiv [f(b) - f(a)]/(b - a)$ . This can be further simplified to the form

$$(16) \quad \alpha - x_{v+1} = (\alpha - x_v)(\alpha - x_{v-1}) \left\{ -\frac{f[x_{v-1}, x_v, \alpha]}{f[x_{v-1}, x_v]} \right\},$$

by introducing

$$f[x_{v-1}, x_v, \alpha] \equiv \frac{f[x_{v-1}, x_v] - f[x_v, \alpha]}{x_{v-1} - \alpha}.$$

Here we have anticipated the divided difference notation to be studied in Chapter 6 and in Problems 2 and 3 of this section. If the function  $f(x)$  has a continuous second derivative in an interval including the points  $x_v$ ,  $x_{v-1}$ , and  $\alpha$ , then it is shown in Theorem 1.1 of Chapter 6 that

$$(17) \quad \begin{aligned}f[x_{v-1}, x_v] &= f'(\xi_v) \\ f[x_{v-1}, x_v, \alpha] &= \frac{1}{2}f''(\eta_v)\end{aligned}$$

for some points  $\xi_v$  and  $\eta_v$  in the obvious intervals. (See also Problem 3.) Thus we deduce that

$$(18) \quad \alpha - x_{v+1} = \frac{-f''(\eta_v)}{2f'(\xi_v)} (\alpha - x_v)(\alpha - x_{v-1}); \quad v = 1, 2, \dots$$

Let us assume that all the iterates are confined to some interval about the root  $\alpha$  and that for all  $\xi$ ,  $\eta$  in this interval

$$(19) \quad \left| \frac{f''(\eta)}{2f'(\xi)} \right| \leq M.$$

Then by setting  $M|\alpha - x_v| \equiv e_v$  we obtain the inequalities

$$(20) \quad e_{v+1} \leq e_v e_{v-1}; \quad v = 1, 2, \dots,$$

upon multiplication of (18) with  $M$  and the use of (19). If we define  $\max(e_0, e_1) = \delta$  the inequalities (20) imply

$$\begin{aligned}e_2 &\leq \delta^2 \\ e_3 &\leq \delta^3 \\ e_4 &\leq \delta^5 \\ &\vdots \\ e_v &\leq \delta^{m_v}\end{aligned}$$

where  $m_0 = m_1 = 1$  and  $m_{v+1} = m_v + m_{v-1}$ ,  $v = 1, 2, \dots$ . The numbers  $m_v$  form what is known as a *Fibonacci sequence*. It may be shown (see Problem 1) that

$$(21) \quad m_v = \frac{1}{\sqrt{5}} (r_+^{v+1} - r_-^{v+1}), \quad r_{\pm} = \frac{1 \pm \sqrt{5}}{2}.$$

Thus for large  $v$ :

$$m_v \approx \frac{1}{\sqrt{5}} (r_+)^{v+1} \cong 0.447(1.618)^{v+1}.$$

If  $\delta < 1$ , then the initial error is reduced by  $10^{-m}$  when  $\delta^{(m_v-1)} \cong 10^{-m}$  and we may compare this number,  $v$ , of iterations with the corresponding numbers  $v_{(2)}$  for the second order method and  $v_{(1)}$  for the first order method (see equation 1.25) in the case  $\delta = M|x_0 - \alpha|$  by noting that

$$(22) \quad m_v - 1 = 2^{v_{(2)}} - 1 = v_{(1)},$$

or

$$\frac{1}{\sqrt{5}} (r_+)^{v+1} \cong 2^{v_{(2)}}.$$

Hence

$$(23) \quad v = c + d v_{(2)}$$

where

$$c = \frac{\log \sqrt{5}}{\log r_+} - 1 \cong 0.672 \quad \text{and} \quad d = \frac{\log 2}{\log r_+} \cong 1.440.$$

We see that somewhat more of the current iterations are needed for a given accuracy than is the case for the second order methods (but it should be recalled that only one function evaluation per iteration is used).

If we were to postulate that as  $v \rightarrow \infty$ :  $|\alpha - x_{v+1}| \approx K|\alpha - x_v|^r$ , then (18), with the coefficient  $|f''/(2f')| = M$ , would yield  $K = M^{1/r}$ ,  $r = r_+$ . In other words, we might say that the false position method is of order  $\cong 1.618$ . Hence, two steps of *Regula Falsi* have an order  $\cong (1.6)^2 > 2.5$  and require only two evaluations of  $f(x)$ .

A geometric interpretation of the scheme (15) is easily given as follows: in the  $x, f(x)$  plane let the line through  $(x_v, f(x_v))$  and  $(x_{v-1}, f(x_{v-1}))$  intersect the  $x$ -axis at a point called  $x_{v+1}$ . In other words  $f(x)$  is approximated by a linear function through the indicated pair of points and the zero of this linear function is taken as the next approximation to the desired root. Depending upon the location of the points in question this procedure may be an interpolation or an extrapolation at each iteration.

In the *classical Regula Falsi* method, the point  $(x_0, f(x_0))$  is used in (15) in place of  $(x_{v-1}, f(x_{v-1}))$  for all  $v = 1, 2, \dots$ . The geometric interpretation of this scheme is quite clear, i.e., all lines pass through the original estimate, which is a poor strategy in general. Again either interpolation or extrapolation may occur.

Finally, we may use in (15), in place of  $(x_{v-1}, f(x_{v-1}))$ , the latest point for which the function value has sign opposite that of  $f(x_v)$ . In this latter method, only interpolation occurs, and furthermore, upper and lower bounds for the root are obtained, which is ideal for estimating the error. However, to start this scheme we must initially obtain such upper and lower estimates of the root and, of course, it is only applicable if  $f(x)$  changes sign at the root in question. This latter variation requires some additional testing and storage of data and hence is slightly more complicated to employ on a digital computer.

From the geometric description of the method of false position a natural generalization is suggested. That is, we set

$$x_{v+1} = P_{k,v}(0), \quad v = k, k + 1, \dots$$

where  $P_{k,v}(f)$  is the polynomial in  $f$  of degree  $k$  which passes through the  $k + 1$  points  $(f(x_v), x_v), (f(x_{v-1}), x_{v-1}), \dots, (f(x_{v-k}), x_{v-k})$ . Clearly, for  $k = 1$ , this is just the scheme (15). The construction of such interpolation polynomials is, in general, treated in Chapter 6, and Section 2 of that chapter is particularly suited for the present purpose. [We must interchange  $x$  and  $f$  or else use inverse interpolation. Also, it is assumed that the function values  $f(x_j)$  are distinct.] It can be shown that these “multi-point” methods have orders  $\eta_k$  which increase monotonically with  $k$  and that  $\lim_{k \rightarrow \infty} \eta_k = 2$ . We have seen that  $\eta_1 \cong 1.618$  so that no great improvement over the method of false position can be obtained. For  $k = 2$  or 3, the orders are close to 2.

Another possibility along the above lines is to use

$$x_{v+1} = P_{v,v}(0), \quad v = 1, 2, \dots;$$

that is a  $v$ th degree polynomial in  $f$  through all the previous iterates  $(f(x_j), x_j)$ ,  $j = 0, 1, \dots, v$  is used to determine the  $(v + 1)$ st iterate. Again, the iterative linear interpolation scheme of Chapter 6, Section 2, can be used for this purpose and it can be shown that the order of convergence is now 2 (for simple roots).

## 2.4. Aitken's $\delta^2$ -Method (Arbitrary Order)

This procedure is frequently presented as a means for *accelerating* the convergence of the functional iteration method based on (3). The method

can be described and motivated as follows: If  $x_v$  is any number approximating a root of (1) or (3), let  $\hat{x}_{v+1}$  be defined by

$$(24) \quad \hat{x}_{v+1} = g(x_v).$$

Then a measure of the “errors” in these two approximations,  $x_v$  and  $\hat{x}_{v+1}$ , can be defined by

$$(25) \quad \begin{aligned} e_v &\equiv g(x_v) - x_v = \hat{x}_{v+1} - x_v, \\ \hat{e}_{v+1} &\equiv g(\hat{x}_{v+1}) - \hat{x}_{v+1}. \end{aligned}$$

Since for a root this error should vanish, i.e.,  $e(\alpha) \equiv g(\alpha) - \alpha = -\phi(\alpha)f(\alpha) = 0$ , we may seek  $x_{v+1}$  by “extrapolating the errors to zero.” That is, the line segment joining the points  $(x_v, e_v)$  and  $(\hat{x}_{v+1}, \hat{e}_{v+1})$  is extended to intersect the  $x$ -axis and the point of intersection is taken as  $x_{v+1}$ . This yields the expression

$$(26a) \quad x_{v+1} = \frac{x_v \hat{e}_{v+1} - \hat{x}_{v+1} e_v}{\hat{e}_{v+1} - e_v} \equiv G(x_v).$$

For actual calculations (26a) is usually written as

$$(26b) \quad x_{v+1} = x_v - \frac{e_v^2}{\hat{e}_{v+1} - e_v},$$

and the evaluations proceed by using (24), (25), and (26b).

From (24)–(26) we see that the  $\delta^2$ -method can be viewed as functional iteration applied to

$$(27a) \quad x = G(x),$$

where

$$(27b) \quad G(x) \equiv \frac{xg(g(x)) - g^2(x)}{g(g(x)) - 2g(x) + x}.$$

[That is, from  $x_0$  we obtain the same sequence of iterates  $x_v$  by the procedure described in (24)–(26) as is obtained from  $x_{v+1} = G(x_v)$ .]

The functional iteration scheme applied to (27) is sometimes known as *Steffensen's method*. In fact, Aitken's  $\delta^2$ -method† was originally proposed to convert any convergent sequence (no matter how generated),  $\{x_n\}$ , into a more rapidly convergent sequence,  $\{x_n'\}$ , by using

$$(28) \quad x_n' \equiv x_n - \frac{(x_{n+1} - x_n)^2}{x_{n+2} - 2x_{n+1} + x_n}.$$

† The denominator in (28) suggests the second difference notation  $\delta^2$ . See equation (3.16a) of Chapter 6.

Several general applications of the  $\delta^2$ -process are illustrated in Problems 6 and 7.

The function (27b) is indeterminate at the root  $x = \alpha$  since  $g(\alpha) = \alpha$ . However, its value there is easily found by an application of L'Hospital's rule, assuming  $g(x)$  to be differentiable at the root and  $g'(\alpha) \neq 1$ :

$$\begin{aligned} G(\alpha) &= \frac{g(g(\alpha)) + \alpha g'(g(\alpha))g'(\alpha) - 2g(\alpha)g'(\alpha)}{g'(g(\alpha))g'(\alpha) - 2g'(\alpha) + 1} \\ &= \frac{\alpha + \alpha[g'(\alpha)]^2 - 2\alpha g'(\alpha)}{[g'(\alpha)]^2 - 2g'(\alpha) + 1} \\ &= \alpha. \end{aligned}$$

The case  $g'(\alpha) = 1$  corresponds to a multiple root of (1) at  $x = \alpha$ . However, in this case too, it can be shown from (33d) that  $\alpha = G(\alpha)$ . Thus, it follows that (27a) has roots wherever (3) has them. To show further that all roots of (27) are also roots of (3), assume that  $x$  is any finite root of (27). Then there are two cases, either  $g(g(x)) - 2g(x) + x$  vanishes or not. If not, then clearing fractions in (27) is legitimate and yields

$$[g(x) - x]^2 = 0.$$

Thus,  $x$  is also a root of (3). If the denominator in (27b) vanishes, the numerator must also vanish (since  $x$  was assumed finite). Now observe that since the denominator vanishes, we may use

$$xg(g(x)) = 2xg(x) - x^2$$

and substitute in the numerator to find that again  $[g(x) - x]^2 = 0$ . In other words (27) *has the same roots as (3)*.

The order of the  $\delta^2$ -method is simply related to the order of the functional iteration applied to  $x = g(x)$ . To derive this result, we assume that  $x = \alpha$  is a root and that:

$$(29a) \quad g'(\alpha) = g''(\alpha) = \dots = g^{(p-1)}(\alpha) = 0;$$

$$(29b) \quad g^{(p)}(\alpha) = p!A \neq 0;$$

$$(29c) \quad g^{(p+1)}(x) \text{ exists in } |x - \alpha| \leq \rho.$$

These conditions imply that  $g(x)$  determines a  $p$ th order method. By Taylor's theorem and (29) for every  $\epsilon$  such that  $|\epsilon| \leq \rho$ :

$$g(\alpha + \epsilon) = g(\alpha) + A\epsilon^p + \frac{g^{(p+1)}(\alpha + \theta\epsilon)}{(p+1)!} \epsilon^{p+1}, \quad 0 < \theta < 1;$$

$$(30a) \quad = \alpha + A\epsilon^p + B\epsilon^{p+1}$$

$$= \alpha + \delta.$$

Here we have introduced

$$(31a) \quad B \equiv \frac{g^{(p+1)}(\alpha + \theta\epsilon)}{(p+1)!}, \quad \delta \equiv (A + B\epsilon)\epsilon^p.$$

Since  $A$  and  $B$  are bounded, we can pick  $\epsilon$  sufficiently small such that  $|\delta| \leq \rho$  and then as in (30a)

$$(30b) \quad g(\alpha + \delta) = \alpha + A\delta^p + B'\delta^{p+1}$$

where

$$(31b) \quad B' = \frac{g^{(p+1)}(\alpha + \phi\delta)}{(p+1)!}, \quad 0 < \phi < 1.$$

From (30) in (27b) we obtain, with  $x = \alpha + \epsilon$  and  $\epsilon \neq 0$ ,

$$(32) \quad \begin{aligned} G(\alpha + \epsilon) &= \frac{(\alpha + \epsilon)g(\alpha + \delta) - (\alpha + \delta)^2}{g(\alpha + \delta) - 2(\alpha + \delta) + (\alpha + \epsilon)} \\ &= \alpha - \frac{\delta^2 - A\epsilon\delta^p - B'\epsilon\delta^{p+1}}{\epsilon - 2\delta + A\delta^p + B'\delta^{p+1}}. \end{aligned}$$

There are two cases,  $p \geq 2$  and  $p = 1$ , to be considered. First, with  $p \geq 2$  equation (32) can be written as

$$\begin{aligned} G(\alpha + \epsilon) &= \alpha - \epsilon^{2p-1}(A + B\epsilon)^2 \\ &\cdot \left\{ \frac{1 - A(A + B\epsilon)^{p-2}\epsilon^{(p-1)^2} - B'(A + B\epsilon)^{p-1}\epsilon^{(p-1)^2+p}}{1 - 2(A + B\epsilon)\epsilon^{p-1} + A(A + B\epsilon)^p\epsilon^{p^2-1} + B'(A + B\epsilon)^{p+1}\epsilon^{p^2+p-1}} \right\}. \end{aligned}$$

It is clear that the bracketed expression approaches 1 as  $\epsilon$  approaches 0, and so the above may be written as

$$(33a) \quad G(\alpha + \epsilon) = \alpha - A^2\epsilon^{2p-1} + \mathcal{O}(\epsilon^{2p}), \quad p \geq 2.$$

For the case  $p = 1$ , (32) becomes

$$(33b) \quad \begin{aligned} G(\alpha + \epsilon) &= \alpha - \epsilon^2(A + B\epsilon) \\ &\cdot \left\{ \frac{(B - B'A) - B'B\epsilon}{(1 - A)^2 - (2B - BA - B'A^2)\epsilon + 2B'BA\epsilon^2 + B'B^2\epsilon^3} \right\}. \end{aligned}$$

Now in general, if  $A \neq 1$ , the bracketed expression approaches  $B^*/(1 - A)$  as  $\epsilon$  approaches 0 since  $B'$  and  $B$  approach  $B^* \equiv g''(\alpha)/2$  and so (33b) can be written as

$$(33c) \quad G(\alpha + \epsilon) = \alpha - \frac{AB^*}{1 - A}\epsilon^2 + \mathcal{O}(\epsilon^3); \quad p = 1, \quad g'(\alpha) = A \neq 1.$$

But, if  $A = g'(\alpha) = 1$  and  $\alpha$  has multiplicity†  $m$ , then by Problem 4:

$$(33d) \quad G(\alpha + \epsilon) = \alpha + \left(1 - \frac{1}{m}\right)\epsilon + \mathcal{O}(\epsilon^2),$$

for

$$p = 1, \quad g'(\alpha) = 1,$$

$$g''(\alpha) = \dots = g^{(m-1)}(\alpha) = 0, \quad g^{(m)}(\alpha) \neq 0;$$

$$\text{for } m = 2, 3, \dots$$

We now invoke a lemma which shall enable us to determine the orders and convergence properties in the cases represented in (33a–d).

**LEMMA 1.** *Let  $G(x)$  be a function, with  $q + 1$  derivatives in a neighborhood of  $x = \alpha$ , such that*

$$G(\alpha) = \alpha,$$

*and for any  $\epsilon$  sufficiently small*

$$(34) \quad G(\alpha + \epsilon) = \alpha + C\epsilon^q + \mathcal{O}(\epsilon^{q+1}).$$

*Then*

$$G'(\alpha) = G''(\alpha) = \dots = G^{(q-1)}(\alpha) = 0, \quad G^{(q)}(\alpha) = q!C.$$

*Proof.* By Taylor's theorem we have, for sufficiently small  $\epsilon$ ,

$$(35) \quad \begin{aligned} G(\alpha + \epsilon) &= \alpha + \frac{\epsilon}{1!} G'(\alpha) + \dots + \frac{\epsilon^q}{q!} G^{(q)}(\alpha) \\ &\quad + \frac{\epsilon^{q+1}}{(q+1)!} G^{(q+1)}(\alpha + \theta\epsilon), \quad 0 < \theta < 1. \end{aligned}$$

The lemma follows by comparing, in the order  $k = 1, 2, \dots, q$ , the values obtained from (34) and (35) of:

$$\lim_{\epsilon \rightarrow 0} \left[ \frac{G(\alpha + \epsilon) - \alpha}{\epsilon^k} k! \right]. \quad \blacksquare$$

By applying this lemma in (33a–d) we deduce the following

**THEOREM 2.** (i) *If functional iteration applied to (3) is of order  $p \geq 2$*

† If the functions in (1), (2), and (3) have  $m$  derivatives, we may verify the equivalence of the statements:

- (i)  $\alpha$  is a root of  $f(x)$  of multiplicity  $m \geq 2$ ;
- (ii)  $f(\alpha) = f'(\alpha) = \dots = f^{(m-1)}(\alpha) = 0, f^{(m)}(\alpha) \neq 0$ ;
- (iii)  $g(\alpha) = \alpha, g'(\alpha) = 1, g''(\alpha) = \dots = g^{(m-1)}(\alpha) = 0, g^{(m)}(\alpha) \neq 0$ .

Statements (i) and (ii) are equivalent by definition. The equivalence of (ii) and (iii) follows from Leibnitz' rule,

$$(\phi f)^{(k)} = \phi^{(k)} f + k\phi^{(k-1)} f' + \dots + k\phi' f^{(k-1)} + \phi f^{(k)},$$

and the fact that  $\phi \neq 0$ , by induction on  $k = 1, 2, \dots, m$ .

for some root  $\alpha$  of (1), then the  $\delta^2$ -method (24)–(26) is of order  $2p - 1$  for this root.

(ii) If functional iteration in (3) is of first order (but not necessarily convergent) for a simple root  $\alpha$  of (1), then the  $\delta^2$ -method is of second order for this root.

(iii) If as in (ii), the root  $\alpha$  of (1) has multiplicity  $m \geq 2$ , then the  $\delta^2$ -method is first order with asymptotic convergence factor  $1 - 1/m$ .

*Proof.* Part (i) follows from (29), Lemma 1, and (33a). Part (ii) follows from Lemma 1 and (33c) since  $g'(\alpha) \neq 1$  is equivalent to  $f'(\alpha) \neq 0$  (i.e., that the root is simple). Finally, (iii) follows from Lemma 1 and (33d) since an  $m$ -fold† root of  $f(x) = 0$  at  $x = \alpha$  implies  $g'(\alpha) = 1, g''(\alpha) = \dots = g^{(m-1)}(\alpha) = 0$  and  $g^{(m)}(\alpha) \neq 0$ . (This proof has assumed that  $f(x)$ ,  $g(x)$  and  $G(x)$  have as many derivatives as required.) ■

From this theorem, it follows that in all cases Aitken's  $\delta^2$ -method converges if  $|\alpha - x_0|$  is sufficiently small. Furthermore, *it is always at least of second order for simple roots*. It is clear that this method can be quite effective and it, or generalizations of it described below, may be very profitably used in practice.

Iterations which converge even faster than the  $\delta^2$ -method are naturally suggested by the above “derivation” of (26). One such generalization is to consider the set of more than two errors associated with  $x_v, \hat{x}_{v+1}, \dots, \hat{x}_{v+\mu}$  as defined in (24) and (25), say

$$(36) \quad e_v, \hat{e}_{v+1}, \dots, \hat{e}_{v+\mu}, \quad \mu > 1;$$

and then determining  $x_{v+1}$  such that this set of errors is “extrapolated” to zero. The details of such a procedure require a knowledge of polynomial interpolation which is discussed in Chapter 6 (see Section 2 in particular). The main point in the correct application of this procedure is to consider the  $x_v$  as functions of the  $e_v$  (i.e., inverse interpolation) in which case the approximation  $x_{v+1}$  can be computed directly by evaluating at  $e = 0$  the polynomial of  $\mu$ th degree in  $e$  that takes on the values  $x_v$  at  $e_v$ , and  $\hat{x}_{v+k}$  at  $\hat{e}_{v+k}$  for  $1 \leq k \leq \mu$ . Other generalizations of these procedures can be obtained by successively increasing the value of  $\mu$ . These considerations are, in fact, the same as those in Subsection 2.3 where generalizations of false position were discussed. Another obvious type of modification is described by introducing  $G^{(0)}(x) \equiv g(x)$ ,  $G^{(1)}(x) = G(x)$  and then forming  $G^{(n)}(x)$  by recursive application of (27).

It should be noted that the correction in (26b),  $-e_v^2/(\hat{e}_{v+1} - e_v)$ , is the quotient of very small quantities. The denominator, being a difference

† See previous footnote.

of small quantities, may require multiple precision evaluation of  $g(g(x_v))$  and  $g(x_v)$ , in order not to lose too many significant figures, especially if  $g'(\alpha) \approx 1$  (i.e., if the root is multiple or nearly so). For these reasons it is important to determine an appropriate  $\delta$  which can be used in Theorem 1.3 in estimating the effect of errors in the  $\delta^2$ -method.

## PROBLEMS, SECTION 2

1. Solve the recursion  $m_{v+1} = m_v + m_{v-1}$ ,  $v = 1, 2, \dots$ , where  $m_0 = m_1 = 1$ . (Try a solution of the form  $m_v = r^v$  which leads to a quadratic with roots  $r_{\pm}$ . Then set  $m_v = ar_+^v + br_-^v$  and determine  $a$  and  $b$  from  $v = 0, 1$ .)
2. The second divided difference of a function  $f(x)$  is defined by:

$$f[x_1, x_2, x_3] \equiv \frac{\frac{f(x_1) - f(x_2)}{x_1 - x_2} - \frac{f(x_2) - f(x_3)}{x_2 - x_3}}{x_1 - x_3}.$$

The first divided difference is just the difference quotient. Use these definitions to verify the derivation of (16).

3. Verify (17).

[Hint: If  $f''(x)$  is continuous in an interval containing  $x_1$ ,  $x_2$ , and  $x_3$ , then the second result in (17) can be derived by the expansion, via Taylor's formula with remainder, of  $f(x_1)$  and  $f(x_3)$  about  $x_2$  plus the fact that a continuous function takes on all values between any two of its values. (Assume, with no loss in generality, that  $x_1 < x_2 < x_3$  in the definition above. That is, it is easy to verify  $f[x_1, x_2, x_3] = f[x_i, x_j, x_k]$  where  $(i, j, k)$  is any permutation of  $(1, 2, 3)$ . This is a special case of a property established in Chapter 6, Section 1, namely that the divided difference is a symmetric function of its arguments.)]

4. Let  $g(\alpha) = \alpha$ ,  $g'(\alpha) = 1$ ,  $g''(\alpha) = g'''(\alpha) = \dots = g^{m-1}(\alpha) = 0$ , and  $g^m(\alpha) \neq 0$ . Then if we assume  $g$  has derivatives of order  $2m$ ,  $g(\alpha + \epsilon) = \alpha + \epsilon + B\epsilon^2$  where  $m \geq 2$  and

$$B(\epsilon) = \frac{g^{(m)}(\alpha)}{m!} \epsilon^{m-2} + \frac{g^{(m+1)}(\alpha)}{(m+1)!} \epsilon^{m-1} + \dots$$

and similarly, with  $\delta \equiv \epsilon + B\epsilon^2$  then  $g(\alpha + \delta) = \alpha + \delta + B'\delta^2$ , where

$$B'(\delta) = \frac{g^m(\alpha)}{m!} \delta^{m-2} + \frac{g^{(m+1)}(\alpha)}{(m+1)!} \delta^{m-1} + \dots$$

Now observe that

$$\delta = \epsilon + B\epsilon^2 = \epsilon + \frac{g^{(m)}(\alpha)}{m!} \epsilon^m + \dots$$

and therefore

$$B'B = \left[ \frac{g^{(m)}(\alpha)}{m!} \right]^2 \epsilon^{2m-4} + \dots$$

$$B' - B = (m-2) \left[ \frac{g^{(m)}(\alpha)}{m!} \right]^2 \epsilon^{2m-3} + \dots$$

Hence, show that formula (33b) yields the results of (33d) for  $m = 1, 2, 3, \dots$

5. Verify that the functional iteration scheme is divergent for (a)  $g(x) \equiv x + x^3$  and (b)  $g(x) \equiv 2x + x^3$ . Nevertheless, as stated in part (ii) of Theorem 2, the Aitken  $\delta^2$ -method is convergent and

$$(a) \quad G(x) = x - \frac{x}{3 + 3x^2 + x^4} = \frac{2}{3}x + \mathcal{O}(x^3)$$

$$(b) \quad G(x) = 6x^3 + \mathcal{O}(x^5).$$

6. (Aitken's  $\delta^2$ -process). Let  $\{x_n\}$ ,  $n = 0, 1, 2, \dots$ , converge to  $\alpha$ ; so that, for some constant  $b$ ,

$$\begin{aligned} r_n &\equiv x_n - \alpha \neq 0, \quad n \geq N; \\ r_{n+1} &= (b + \epsilon_n)r_n, \quad |b| < 1, \epsilon_n = o(1). \dagger \end{aligned}$$

Show that (28) is meaningful for  $n \geq N$ , i.e.,

$$x_{n+2} - 2x_{n+1} + x_n \neq 0 \quad \text{for } n \geq N;$$

and that

$$\lim_{n \rightarrow \infty} \frac{x_n' - \alpha}{x_n - \alpha} = 0.$$

[Hint: Verify that

$$\begin{aligned} x_{n+2} - 2x_{n+1} + x_n &= r_{n+2} - 2r_{n+1} + r_n \\ &= r_n[(b - 1)^2 + o(1)]. \end{aligned}$$

Also show from (28) that

$$\begin{aligned} x_n' - \alpha &= r_n - r_n \frac{[b - 1 + o(1)]^2}{(b - 1)^2 + o(1)} \\ &= r_n o(1). \end{aligned}$$

7. Apply Aitken's  $\delta^2$ -process (28) to the sequence

$$x_n = \alpha + b\rho_1^n + c\rho_2^n, \quad n = 0, 1, 2, \dots,$$

where  $|\rho_2| < |\rho_1| < 1$ . Show that

$$x_n' = \alpha + \mathcal{O}(\rho_2^n) + \mathcal{O}(\rho_1^{2n}).$$

What improvement results by applying the  $\delta^2$ -process to the sequence  $\{x_n'\}$ ?

### 3. FUNCTIONAL ITERATION FOR A SYSTEM OF EQUATIONS

Let  $\mathbf{x}$  be an  $n$ -dimensional column vector with components  $x_1, x_2, \dots, x_n$  and  $\mathbf{g}(\mathbf{x})$  an  $n$ -dimensional vector valued function, i.e., a column vector with components  $g_1(\mathbf{x}), g_2(\mathbf{x}), \dots, g_n(\mathbf{x})$ . Then the system to be solved is

$$(1) \quad \mathbf{x} = \mathbf{g}(\mathbf{x}).$$

$\dagger$  We write  $\delta_n = o(1)$  iff there is some number  $N$  such that  $\delta_n$  is defined for all  $n \geq N$  and  $\lim_{n \rightarrow \infty} \delta_n = 0$ . In the text, we also use  $o(1)$  as a generic symbol to represent the members of any sequence which tends to zero.

The solution (or root) is some vector, say  $\alpha$ , with components  $\alpha_1, \alpha_2, \dots, \alpha_n$  which is, of course, some point in the  $n$ -dimensional space. Starting with a point  $\mathbf{x}^{(0)} = [x_1^{(0)}, x_2^{(0)}, \dots, x_n^{(0)}]^T$ , the exact analog of the functional iteration of Section 1 is

$$(2) \quad \mathbf{x}^{(\nu+1)} = \mathbf{g}(\mathbf{x}^{(\nu)}), \quad \nu = 0, 1, 2, \dots$$

The first result is analogous to Theorem 1.1. But where absolute values were used previously, we must now use some vector norm (see Chapter 1, Section 1). For example, we may choose any one of the norms

$$(3) \quad \begin{aligned} \|\mathbf{x}\|_\infty &\equiv \max_{1 \leq i \leq n} |x_i|, \\ \|\mathbf{x}\|_1 &\equiv \sum_{i=1}^n |x_i|, \\ \|\mathbf{x}\|_2 &\equiv \sqrt{\sum_{i=1}^n |x_i|^2}. \end{aligned}$$

**THEOREM 1.** *Let  $\mathbf{g}(\mathbf{x})$  satisfy*

$$(4a) \quad \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\| \leq \lambda \|\mathbf{x} - \mathbf{y}\|$$

*for all vectors  $\mathbf{x}, \mathbf{y}$  such that  $\|\mathbf{x} - \mathbf{x}^{(0)}\| \leq \rho, \|\mathbf{y} - \mathbf{x}^{(0)}\| \leq \rho$  with the Lipschitz constant,  $\lambda$ , satisfying*

$$(4b) \quad 0 \leq \lambda < 1.$$

*Let the initial iterate,  $\mathbf{x}^{(0)}$ , satisfy*

$$(5) \quad \|\mathbf{g}(\mathbf{x}^{(0)}) - \mathbf{x}^{(0)}\| \leq (1 - \lambda)\rho.$$

*Then:* (i) *all iterates, (2), satisfy*

$$\|\mathbf{x}^{(\nu)} - \mathbf{x}^{(0)}\| \leq \rho;$$

(ii) *the iterates converge to some vector, say*

$$\lim_{\nu \rightarrow \infty} \mathbf{x}^{(\nu)} = \alpha,$$

*which is a root of (1);*

(iii)  *$\alpha$  is the only root of (1) in  $\|\mathbf{x} - \mathbf{x}^{(0)}\| \leq \rho$ .*

*Proof.* Duplicate the proof of Theorem 1.1 with the replacement of absolute value signs by norm symbols. ■

As a consequence of this proof, it is also seen that the iterates converge geometrically, and at least as fast as  $\lambda^\nu \rightarrow 0$ . Of course, it is more difficult to verify (4), the Lipschitz continuity of a vector valued function, than it is in the case of a scalar function.

However, again as in Section 1, a more useful result can be obtained if we are willing to place more restrictions on  $\mathbf{g}(\mathbf{x})$  and assume the existence of a root. We immediately see that Theorem 1.2 and its proof hold if absolute value signs are replaced by norms. Furthermore, the corollary to Theorem 1.2 becomes

**THEOREM 2.** *Let (1) have a root  $\mathbf{x} = \boldsymbol{\alpha}$ . Let the components  $g_i(\mathbf{x})$  have continuous first partial derivatives and satisfy*

$$(6) \quad \left| \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right| \leq \frac{\lambda}{n}, \quad \lambda < 1;$$

for all  $\mathbf{x}$  in

$$(7) \quad \|\mathbf{x} - \boldsymbol{\alpha}\|_\infty \leq \rho.$$

*Then:* (i) *For any  $\mathbf{x}^{(0)}$  satisfying (7) all the iterates  $\mathbf{x}^{(v)}$  of (2) also satisfy (7).*

(ii) *For any  $\mathbf{x}^{(0)}$  satisfying (7) the iterates (2) converge to the root  $\boldsymbol{\alpha}$  of (1) which is unique in (7).*

*Proof.* For any two points  $\mathbf{x}, \mathbf{y}$  in (7) we have by Taylor's theorem:

$$(8) \quad g_i(\mathbf{x}) - g_i(\mathbf{y}) = \sum_{j=1}^n \frac{\partial g_i(\xi^{(i)})}{\partial x_j} (x_j - y_j), \quad i = 1, 2, \dots, n;$$

where  $\xi^{(i)}$  is a point on the open line segment joining  $\mathbf{x}$  and  $\mathbf{y}$ . Thus,  $\xi^{(i)}$  is in (7), and using (3) and (6) yields

$$\begin{aligned} |g_i(\mathbf{x}) - g_i(\mathbf{y})| &\leq \sum_{j=1}^n \left| \frac{\partial g_i(\xi^{(i)})}{\partial x_j} \right| \cdot |x_j - y_j|, \\ &\leq \|\mathbf{x} - \mathbf{y}\|_\infty \sum_{j=1}^n \left| \frac{\partial g_i(\xi^{(i)})}{\partial x_j} \right|, \\ &\leq \lambda \|\mathbf{x} - \mathbf{y}\|_\infty. \end{aligned}$$

Since the inequality holds for each  $i$ , we have

$$(9) \quad \|\mathbf{g}(\mathbf{x}) - \mathbf{g}(\mathbf{y})\|_\infty \leq \lambda \|\mathbf{x} - \mathbf{y}\|_\infty,$$

and thus we have proven that  $\mathbf{g}(\mathbf{x})$  is Lipschitz continuous in the domain (7), with respect to the indicated norm. Now note that for any  $\mathbf{x}^{(0)}$  in (7),

$$\begin{aligned} \|\mathbf{x}^{(1)} - \boldsymbol{\alpha}\|_\infty &= \|\mathbf{g}(\mathbf{x}^{(0)}) - \mathbf{g}(\boldsymbol{\alpha})\|_\infty \\ &\leq \lambda \|\mathbf{x}^{(0)} - \boldsymbol{\alpha}\|_\infty \\ &\leq \lambda \rho, \end{aligned}$$

and so  $\mathbf{x}^{(1)}$  is also in (7). By an obvious induction we have then

$$\begin{aligned}
 \| \mathbf{x}^{(v)} - \boldsymbol{\alpha} \|_{\infty} &= \| \mathbf{g}(\mathbf{x}^{(v-1)}) - \mathbf{g}(\boldsymbol{\alpha}) \|_{\infty} \\
 &\leq \lambda \| \mathbf{x}^{(v-1)} - \boldsymbol{\alpha} \|_{\infty} \\
 (10) \quad &\vdots \\
 &\leq \lambda^v \| \mathbf{x}^{(0)} - \boldsymbol{\alpha} \|_{\infty} \\
 &\leq \lambda^v \rho
 \end{aligned}$$

and hence all  $\mathbf{x}^{(v)}$  lie in (7). The convergence immediately follows from (10) since  $\lambda < 1$ . The uniqueness follows as before. ■

The crucial point in the preceding proof is the derivation of (9). It is clear from this derivation that (6) could be replaced by a number of conditions which are perhaps less restrictive and the theorem would still remain valid. One such condition is

$$(11) \quad \max_i \sum_{j=1}^n |g_{ij}(\mathbf{x})| \leq \lambda < 1, \quad \text{for all } \| \mathbf{x} - \boldsymbol{\alpha} \|_{\infty} < \rho,$$

where we have introduced the elements  $g_{ij}(\mathbf{x}) = \partial g_i(\mathbf{x}) / \partial x_j$ . If we define the matrix  $G(\mathbf{x}) \equiv (g_{ij}(\mathbf{x}))$  then (11) may be written as  $\| G(\mathbf{x}) \|_{\infty} \leq \lambda < 1$  in which case we mean the natural matrix norm induced by the maximum vector norm (see Chapter 1, Section 1).

If the function  $\mathbf{g}(\mathbf{x})$  is such that at a root

$$(12) \quad G(\boldsymbol{\alpha}) \equiv \left( \frac{\partial g_i(\boldsymbol{\alpha})}{\partial x_j} \right) = 0, \quad i, j = 1, 2, \dots, n$$

and these derivatives are continuous near the root, then (6) as well as (11) can be satisfied for some  $\rho > 0$ . If, in addition, the second derivatives

$$\frac{\partial^2 g_i(\mathbf{x})}{\partial x_j \partial x_k}$$

all exist in a neighborhood of the root, then again by Taylor's theorem

$$g_i(\mathbf{x}) - g_i(\boldsymbol{\alpha}) = \frac{1}{2} \sum_{j=1}^n \sum_{k=1}^n \frac{\partial^2 g_i(\xi^i)}{\partial x_j \partial x_k} (x_j - \alpha_j)(x_k - \alpha_k).$$

Now in the iteration, (2), we find

$$\| \mathbf{x}^{(v)} - \boldsymbol{\alpha} \|_{\infty} \leq M \| \mathbf{x}^{(v-1)} - \boldsymbol{\alpha} \|_{\infty}^2,$$

where  $M$  is chosen such that

$$\max_{i, j, k} \left| \frac{\partial^2 g_i(\mathbf{x})}{\partial x_j \partial x_k} \right| \leq \frac{2M}{n^2}.$$

Thus, quadratic convergence can occur in solving systems of equations by iteration.

### 3.1. Some Explicit Iteration Schemes for Systems

In the general case, the system to be solved is of the form

$$(13) \quad \mathbf{f}(\mathbf{x}) = \mathbf{0}$$

where  $\mathbf{f}(\mathbf{x}) = [f_1(\mathbf{x}), f_2(\mathbf{x}), \dots, f_n(\mathbf{x})]^T$  is an  $n$ -component column vector. Such a system can be written in the form (1) in a variety of ways; we examine here the choice

$$(14) \quad \mathbf{g}(\mathbf{x}) \equiv \mathbf{x} - A(\mathbf{x})\mathbf{f}(\mathbf{x}),$$

where  $A(\mathbf{x})$  is an  $n$ th order square matrix with components  $a_{ij}(\mathbf{x})$ . The equations (1) and (13) will have the same set of solutions if  $A(\mathbf{x})$  is non-singular [since in that case  $A(\mathbf{x})\mathbf{f}(\mathbf{x}) = \mathbf{0}$  implies  $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ ].

The simplest choice for  $A(\mathbf{x})$  is

$$(15) \quad A(\mathbf{x}) \equiv A,$$

a constant non-singular matrix. If we introduce the matrix

$$(16) \quad J(\mathbf{x}) \equiv \left( \frac{\partial f_i(\mathbf{x})}{\partial x_j} \right),$$

whose determinant is the Jacobian of the functions  $f_i(\mathbf{x})$ , then from (14)–(16) we have

$$(17) \quad G(\mathbf{x}) \equiv \left( \frac{\partial g_i(\mathbf{x})}{\partial x_j} \right) = I - AJ(\mathbf{x}).$$

Thus by Theorem 2, or its modification in which (11) replaces (6), the iterations determined by using

$$\mathbf{x}^{(v+1)} = \mathbf{x}^{(v)} - A\mathbf{f}(\mathbf{x}^{(v)})$$

will converge, for  $\mathbf{x}^{(0)}$  sufficiently close to  $\alpha$ , if the elements in the matrix (17) are sufficiently small, for example, as in the case that  $J(\alpha)$  is non-singular and  $A$  is approximately the inverse of  $J(\alpha)$ . This procedure is the analog of the chord method and it naturally suggests a modification which is again called Newton's method.

In Newton's method (15) is replaced by the choice

$$(18) \quad A(\mathbf{x}) \equiv J^{-1}(\mathbf{x}),$$

with the assumption of course that  $\det |J(\mathbf{x})| \neq 0$  for  $\mathbf{x}$  in  $\|\mathbf{x} - \alpha\| \leq \rho$ . In actually using the above procedure, an inverse need not be computed at each iteration; instead, a linear system of order  $n$  has to be solved. To

see this, and at the same time gain some insight into the method, we note that by using (18) in (14) the iterations for Newton's method are:

$$(19a) \quad \begin{aligned} \mathbf{x}^{(v+1)} &= \mathbf{g}(\mathbf{x}^{(v)}), \\ &= \mathbf{x}^{(v)} - J^{-1}(\mathbf{x}^{(v)})\mathbf{f}(\mathbf{x}^{(v)}). \end{aligned}$$

From this we obtain

$$(19b) \quad J(\mathbf{x}^{(v)})(\mathbf{x}^{(v)} - \mathbf{x}^{(v+1)}) = \mathbf{f}(\mathbf{x}^{(v)}),$$

which is the system to be solved for the vector  $(\mathbf{x}^{(v)} - \mathbf{x}^{(v+1)})$ .

To show that this method is of second order we must verify that (12) is satisfied when (18) is used in (14). The  $j$ th column of  $G(\mathbf{x})$  is then given by

$$\begin{aligned} \frac{\partial \mathbf{g}(\mathbf{x})}{\partial x_j} &= \frac{\partial \mathbf{x}}{\partial x_j} - \frac{\partial}{\partial x_j} [J^{-1}(\mathbf{x})\mathbf{f}(\mathbf{x})], \\ &= \frac{\partial \mathbf{x}}{\partial x_j} - J^{-1}(\mathbf{x}) \frac{\partial \mathbf{f}(\mathbf{x})}{\partial x_j} - \frac{\partial J^{-1}(\mathbf{x})}{\partial x_j} \mathbf{f}(\mathbf{x}). \end{aligned}$$

By setting  $\mathbf{x} = \boldsymbol{\alpha}$  in the above and recalling that  $\mathbf{f}(\boldsymbol{\alpha}) = \mathbf{0}$  and  $J = (\partial f_i / \partial x_j)$  we get

$$G(\boldsymbol{\alpha}) = I - J^{-1}(\boldsymbol{\alpha})J(\boldsymbol{\alpha}) - \mathbf{O} = \mathbf{O}.$$

To determine  $\partial J^{-1}(\mathbf{x}) / \partial x_j$ , note that

$$\frac{\partial (J^{-1}J)}{\partial x_j} = J^{-1} \frac{\partial J}{\partial x_j} + \frac{\partial J^{-1}}{\partial x_j} J = \frac{\partial I}{\partial x_j} = \mathbf{O}$$

and hence

$$\frac{\partial J^{-1}(\mathbf{x})}{\partial x_j} = -J^{-1}(\mathbf{x}) \frac{\partial J(\mathbf{x})}{\partial x_j} J^{-1}(\mathbf{x}).$$

Thus, we need only require that  $\mathbf{f}(\mathbf{x})$  have two derivatives and  $J(\mathbf{x})$  be non-singular at the root, and then the convergence of Newton's method is quadratic.

For a geometric interpretation of Newton's method we consider a system of two equations and drop subscripts by using

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \equiv \begin{pmatrix} x \\ y \end{pmatrix}, \quad \begin{pmatrix} f_1(x_1, x_2) \\ f_2(x_1, x_2) \end{pmatrix} \equiv \begin{pmatrix} f(x, y) \\ g(x, y) \end{pmatrix}.$$

Then

$$J(x) = \begin{pmatrix} f_x & f_y \\ g_x & g_y \end{pmatrix}$$

and the system (19) can be written as:

$$(20a) \quad (x_{v+1} - x_v)f_x(x_v, y_v) + (y_{v+1} - y_v)f_y(x_v, y_v) + f(x_v, y_v) = 0$$

$$(20b) \quad (x_{v+1} - x_v)g_x(x_v, y_v) + (y_{v+1} - y_v)g_y(x_v, y_v) + g(x_v, y_v) = 0$$

In the  $(x, y, z)$ -space the equations

$$(21a) \quad z = (x - x_v)f_x(x_v, y_v) + (y - y_v)f_y(x_v, y_v) + f(x_v, y_v),$$

$$(21b) \quad z = (x - x_v)g_x(x_v, y_v) + (y - y_v)g_y(x_v, y_v) + g(x_v, y_v),$$

each represent planes. The plane (21a) is tangent to the surface  $z = f(x, y)$  at the point  $(x_v, y_v, f(x_v, y_v))$ , and the plane (21b) is tangent to  $z = g(x, y)$  at the point  $(x_v, y_v, g(x_v, y_v))$ . Clearly, the point  $(x_{v+1}, y_{v+1})$  determined from (20) is the point of intersection of these two planes with the plane  $z = 0$ , i.e., the  $(x, y)$ -plane. Thus, in passing from one dimension (Section 2.2) to two dimensions, Newton's method is generalized by replacing tangent lines with tangent planes. In the more general case of  $n$  dimensions the obvious interpretation, using tangent hyperplanes, is valid. Each of the equations

$$z = \sum_{k=1}^n (x_k - x_k^{(v)}) \frac{\partial f_i(\mathbf{x}^{(v)})}{\partial x_k} + f_i(\mathbf{x}^{(v)}), \quad i = 1, 2, \dots, n,$$

represents a hyperplane in the  $(x_1, x_2, \dots, x_n, z)$  space of  $n + 1$  dimensions which is tangent at the point  $(x_1^{(v)}, x_2^{(v)}, \dots, x_n^{(v)})$  to the corresponding hypersurface

$$z = f_i(x_1, x_2, \dots, x_n).$$

The difficulties which may arise in the solution of systems using Newton's method can be interpreted by means of these geometric considerations.

### 3.2. Convergence of Newton's Method

If the initial iterate  $\mathbf{x}^{(0)}$  is sufficiently close to the root  $\alpha$  of  $\mathbf{f}(\mathbf{x}) = \mathbf{0}$ , then Theorem 2 can be used to prove that the Newton iterates,  $\mathbf{x}^{(v)}$ , defined in (19) converge to the root. In addition, if the Jacobian  $J(\mathbf{x})$  is non-singular at the root,  $\mathbf{x} = \alpha$ , and differentiable there, then the convergence is second order. However, we do not know from these results if a given initial iterate  $\mathbf{x}^{(0)}$  is close enough to the unknown root,  $\alpha$ . We shall develop a sufficient condition, under which Newton's scheme converges, with the property that this condition may be explicitly checked without a knowledge of  $\alpha$ . In fact, the theorem to be established also proves the existence of a unique root of  $\mathbf{f}(\mathbf{x})$  in an appropriate interval about the initial iterate,  $\mathbf{x}^{(0)}$ . Thus, we have an alternative to Theorem 1 which we state as

**THEOREM 3.** *Let the initial iterate  $\mathbf{x}^{(0)}$  be such that the Jacobian matrix  $J(\mathbf{x}^{(0)})$  defined in (16) has an inverse with norm bounded by*

$$(22a) \quad \|J^{-1}(\mathbf{x}^{(0)})\| \leq a.$$

Let the difference of the first two Newton iterates be bounded by

$$(22b) \quad \|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| = \|J^{-1}(\mathbf{x}^{(0)})\mathbf{f}(\mathbf{x}^{(0)})\| \leq b.$$

Let the components of  $\mathbf{f}(\mathbf{x})$  have continuous second derivatives which satisfy

$$(22c) \quad \sum_{k=1}^n \left| \frac{\partial^2 f_i(\mathbf{x})}{\partial x_j \partial x_k} \right| \leq \frac{c}{n},$$

for all  $\mathbf{x}$  in  $\|\mathbf{x} - \mathbf{x}^{(0)}\| \leq 2b$ ;  $i, j = 1, 2, \dots, n$ . If the constants  $a$ ,  $b$ , and  $c$  are such that

$$(22d) \quad abc \leq \frac{1}{2}$$

then: (i) the Newton iterates (19) are uniquely defined and lie in the “ $2b$ -sphere” about  $\mathbf{x}^{(0)}$ :

$$\|\mathbf{x}^{(v)} - \mathbf{x}^{(0)}\| \leq 2b;$$

(ii) the iterates converge to some vector, say  $\lim_{v \rightarrow \infty} \mathbf{x}^{(v)} = \boldsymbol{\alpha}$ , for which  $\mathbf{f}(\boldsymbol{\alpha}) = \mathbf{0}$  and

$$(23) \quad \|\mathbf{x}^{(v)} - \boldsymbol{\alpha}\| \leq \frac{2b}{2^v}.$$

[All vector norms in the statement and proof of this theorem are maximum norms, i.e.,  $\|\mathbf{x}\| = \max_i |x_i|$ , and matrix norms are the corresponding induced natural norm, i.e.,  $\|A\| = \max_i \left( \sum_{j=1}^n |a_{ij}| \right)$ .]

*Proof.* The proof proceeds by a somewhat lengthy induction. For convenience, we use the notation  $J_v \equiv J(\mathbf{x}^{(v)})$  for the Jacobian matrices (16) and show for all  $v = 0, 1, 2, \dots$  that with  $A_{v+1} \equiv I - J_v^{-1} J_{v+1}$ ,

$$(24a) \quad \|\mathbf{x}^{(v+1)} - \mathbf{x}^{(v)}\| \leq \frac{b}{2^v},$$

$$(24b) \quad \|\mathbf{x}^{(v+1)} - \mathbf{x}^{(0)}\| \leq 2b,$$

$$(24c) \quad \|A_{v+1}\| \equiv \|J_v^{-1}(J_v - J_{v+1})\| \leq \frac{1}{2},$$

$$(24d) \quad \|J_{v+1}^{-1}\| = \|(I - A_{v+1})^{-1} J_v^{-1}\| \leq 2^{v+1}a.$$

From the hypothesis (22b) it trivially follows that (24a, b) are satisfied for  $v = 0$ . Now when (24b) is established up to and including any value  $v$  then  $\mathbf{x}^{(v+1)}$  and  $\mathbf{x}^{(v)}$  are in the  $2b$ -sphere about  $\mathbf{x}^{(0)}$  in which we are assured that the second derivatives of the  $f_i(\mathbf{x})$  are continuous. Then we can apply Taylor's theorem to the components of  $J_{v+1}$  to obtain

$$\begin{aligned} \frac{\partial f_i(\mathbf{x}^{(v+1)})}{\partial x_j} &= \frac{\partial f_i(\mathbf{x}^{(v)})}{\partial x_j} + \sum_{k=1}^n (x_k^{(v+1)} - x_k^{(v)}) \frac{\partial^2 f_i[\mathbf{x}^{(v)} + \theta_i(\mathbf{x}^{(v+1)} - \mathbf{x}^{(v)})]}{\partial x_j \partial x_k}, \\ &\quad 0 < \theta_i < 1. \end{aligned}$$

Since  $\mathbf{x}^{(\nu+1)}$  and  $\mathbf{x}^{(\nu)}$  are in  $\|\mathbf{x} - \mathbf{x}^{(0)}\| \leq 2b$ , so is the point  $\mathbf{x}^{(\nu)} + \theta(\mathbf{x}^{(\nu+1)} - \mathbf{x}^{(\nu)})$ , and (22c) applies. This gives from the above

$$(25) \quad \|J_{\nu+1} - J_\nu\| \leq c\|\mathbf{x}^{(\nu+1)} - \mathbf{x}^{(\nu)}\|.$$

At the present stage in the proof this is valid only for  $\nu = 0$ . But then using this and (22a, b, d) in (24c) with  $\nu = 0$  yields

$$\begin{aligned} \|A_1\| &\leq \|J_0^{-1}\| \cdot \|J_1 - J_0\| \\ &\leq ac\|\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\| \\ &\leq abc \\ &\leq \frac{1}{2}. \end{aligned}$$

Now (24a, b, c) have been established for  $\nu = 0$ .

If for any  $\nu$  the matrix  $J_\nu$  is non-singular, then we have the identity

$$J_{\nu+1} = J_\nu(I - A_{\nu+1}),$$

where, as in (24c),  $A_{\nu+1} \equiv J_\nu^{-1}(J_\nu - J_{\nu+1})$ . But from the Corollary to Theorem 1.5 of Chapter 1 it follows that if  $\|A_{\nu+1}\| < 1$  then  $J_{\nu+1}$  is non-singular and

$$(26) \quad \|J_{\nu+1}^{-1}\| \leq \frac{\|J_\nu^{-1}\|}{1 - \|A_{\nu+1}\|}.$$

Since (24c) is valid for  $\nu = 0$  we can use this in (26) to get

$$\|J_1^{-1}\| \leq 2a.$$

Thus (24) has been verified for  $\nu = 0$ .

Let us now make the inductive assumption that (24) is valid for all  $\nu \leq k - 1$  and proceed to show that it is also valid for  $\nu = k$ . Since  $J_k$  is non-singular, the  $(k + 1)$ st Newton iterate,  $\mathbf{x}^{(k+1)}$ , is uniquely defined and we have from (19a):

$$\begin{aligned} (27) \quad \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| &= \|J_k^{-1}\mathbf{f}(\mathbf{x}^{(k)})\| \\ &\leq \|J_k^{-1}\| \cdot \|\mathbf{f}(\mathbf{x}^{(k)})\|. \end{aligned}$$

However, since (24b) is valid for  $\nu = k - 1$ , the point  $\mathbf{x}^{(k)}$  is in the  $2b$ -sphere about  $\mathbf{x}^{(0)}$ . Then by Taylor's theorem, with remainder term  $\mathbf{R}$ , and (19b) with  $\nu = k - 1$ :

$$\begin{aligned} \mathbf{f}(\mathbf{x}^{(k)}) &= \mathbf{f}(\mathbf{x}^{(k-1)}) + J_{k-1}[\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}] + \mathbf{R}(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}) \\ &= \mathbf{R}(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)}). \end{aligned}$$

Using (22c), we can bound the above remainder term to yield

$$\begin{aligned}
 \|f(\mathbf{x}^{(k)})\| &= \max_i |R_i(\mathbf{x}^{(k)}, \mathbf{x}^{(k-1)})| \\
 &= \max_i \left| \sum_{j=1}^n \sum_{l=1}^n \frac{(x_j^{(k)} - x_j^{(k-1)})(x_l^{(k)} - x_l^{(k-1)})}{2!} \right. \\
 &\quad \times \left. \frac{\partial^2 f_i}{\partial x_j \partial x_l} [\mathbf{x}^{(k-1)} + \phi_i(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})] \right|, \quad 0 < \phi_i < 1, \\
 (28) \quad &\leq \frac{c}{2} \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2.
 \end{aligned}$$

Again, we have used the fact that  $\mathbf{x}^{(k-1)} + \phi(\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)})$  is in the  $2b$ -sphere about  $\mathbf{x}^{(0)}$  since  $\mathbf{x}^{(k)}$  and  $\mathbf{x}^{(k-1)}$  are in it. Now using (28) in (27) and recalling that (24) is assumed valid for all  $\nu \leq k-1$  we get

$$\begin{aligned}
 (29) \quad \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| &\leq \frac{c}{2} \|J_k^{-1}\| \cdot \|\mathbf{x}^{(k)} - \mathbf{x}^{(k-1)}\|^2 \\
 &\leq \frac{c}{2} (2^k a) \left( \frac{b}{2^{k-1}} \right)^2 = \frac{ab^2 c}{2^{k-1}} \\
 &\leq \frac{b}{2^k}.
 \end{aligned}$$

Thus (24a) is established for  $\nu = k$ . Then since

$$\begin{aligned}
 \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(0)}\| &= \left\| \sum_{l=0}^k (\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}) \right\| \\
 &\leq \sum_{l=0}^k \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\| \\
 &\leq b \sum_{l=0}^k \frac{1}{2^l} \\
 &\leq 2b,
 \end{aligned}$$

we have also established (24b) for  $\nu = k$ . But then  $\mathbf{x}^{(k+1)}$  is in the  $2b$ -sphere about  $\mathbf{x}^{(0)}$  and so (25) is valid with  $\nu = k$ . This gives

$$\begin{aligned}
 \|A_{k+1}\| &= \|J_k^{-1}(J_k - J_{k+1})\| \\
 &\leq \|J_k^{-1}\| \cdot \|J_{k+1} - J_k\| \\
 &\leq c \|J_k^{-1}\| \cdot \|\mathbf{x}^{(k+1)} - \mathbf{x}^{(k)}\| \\
 &\leq abc \\
 &\leq \frac{1}{2}.
 \end{aligned}$$

Thus (24c) is valid with  $\nu = k$  and implies that  $J_{k+1}$  is non-singular. Then using (26) with  $\nu = k$  yields (24d), and the inductive proof of (24) is complete.

Part (i) of the theorem follows from (24b, d). The convergence of the  $\mathbf{x}^{(\nu)}$  follows from (24a) since they form a Cauchy sequence: i.e.,

$$(30) \quad \begin{aligned} \|\mathbf{x}^{(\nu+m)} - \mathbf{x}^{(\nu)}\| &= \left\| \sum_{l=\nu}^{\nu+m-1} (\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}) \right\| \\ &\leq \sum_{l=\nu}^{\nu+m-1} \|\mathbf{x}^{(l+1)} - \mathbf{x}^{(l)}\| \leq b \sum_{l=\nu}^{\nu+m-1} \frac{1}{2^l} \\ &\leq \frac{b}{2^{\nu-1}}. \end{aligned}$$

Calling the limit vector  $\alpha$ , we use (24a), (28) and the continuity of  $\mathbf{f}(\mathbf{x})$  to deduce that

$$\|\mathbf{f}(\mathbf{x}^{(k)})\| \leq \frac{2b^2c}{4^k},$$

and  $\lim_{k \rightarrow \infty} \mathbf{f}(\mathbf{x}^{(k)}) = \mathbf{f}(\alpha) = \mathbf{0}$ . Letting  $m \rightarrow \infty$ , (30) implies

$$\|\alpha - \mathbf{x}^{(\nu)}\| \leq \frac{2b}{2^\nu},$$

and so, part (ii) is established, concluding the proof of the theorem. ■

This theorem is valid if  $n = 1$ . The hypothesis permits the case that  $J(\alpha)$  is singular. Hence, it is reasonable that the conclusion (ii) shows at most linear convergence. But (ii), moreover, implies that the convergence factor is at most  $\frac{1}{2}$ . This seems to contradict the fact shown earlier for the scalar case ( $n = 1$ ), that the convergence factor is only  $1 - 1/p$  if  $f(\alpha)$  is a zero of order  $p > 1$ . The contradiction doesn't exist because, as we show in Problem 6, the requirement  $abc \leq \frac{1}{2}$  can only be satisfied if  $p \leq 2$ . [For example,  $f(x) \equiv x^p$  and  $x_0 \neq 0$  satisfy

$$a \cdot b \cdot c \geq \left| \frac{1}{f'(x_0)} \cdot \frac{f(x_0)}{f'(x_0)} \cdot f''(x_0) \right| \equiv 1 - 1/p.]$$

On the other hand, if  $h \equiv abc < \frac{1}{2}$ , then Kantorovich has shown in general that

$$\|\mathbf{x}^{(\nu)} - \alpha\| \leq \frac{(2h)^{2^\nu-1}}{2^{\nu-1}} b,$$

i.e., quadratic convergence. See Problems 3, 4, 5, and 6 for further results in the scalar case.

### 3.3. A Special Acceleration Procedure for Non-Linear Systems

If the non-linear system to be solved is written in the form

$$(31) \quad \mathbf{x} = \mathbf{g}(\mathbf{x}),$$

then the obviously implied iterations

$$(32) \quad \mathbf{x}^{(0)} = \text{arbitrary}; \quad \mathbf{x}^{(\nu+1)} = \mathbf{g}(\mathbf{x}^{(\nu)}), \quad \nu = 0, 1, \dots,$$

may or may not converge. However, as with linear systems, we can alter the procedure (32) in a manner which will generally improve the rate of convergence or may even yield a convergent scheme when the basic one diverges. The acceleration procedure is defined by:

$$(33a) \quad \begin{aligned} \mathbf{x}^{(0)} &= \text{arbitrary} \\ \mathbf{x}^{(\nu+1)} &= \Theta \mathbf{g}(\mathbf{x}^{(\nu)}) + (I - \Theta) \mathbf{x}^{(\nu)}, \quad \nu = 0, 1, \dots, \end{aligned}$$

where  $\Theta$  is a diagonal matrix given by

$$(33b) \quad \Theta \equiv (\theta_i \delta_{ij}), \quad \det [\Theta] = \theta_1 \theta_2 \dots \theta_n \neq 0.$$

Of course, if  $\Theta = I$ , then the basic scheme (32) results. The scalar form of the  $i$ th equation in (33a) is clearly

$$x_i^{(\nu+1)} = \theta_i g_i(\mathbf{x}^{(\nu)}) + (1 - \theta_i) x_i^{(\nu)}, \quad 1 \leq i \leq n,$$

and so the iterations are easily evaluated in an explicit manner.

Let us assume that (31) has a solution, say  $\mathbf{x} = \alpha$ , and that in some  $\rho$ -sphere about this solution,  $\|\mathbf{x} - \alpha\| \leq \rho$ , the vector function  $\mathbf{g}$  has continuous first partial derivatives,  $g_{ij}(\mathbf{x}) \equiv \partial g_i(\mathbf{x}) / \partial x_j$ , which satisfy the conditions

$$(34) \quad |1 - g_{ii}(\mathbf{x})| > \sum_{j \neq i} |g_{ij}(\mathbf{x})|, \quad 1 \leq i \leq n.$$

Under these conditions it can be shown that the iterations (33) will converge, for some choice of the  $\theta_i$ , to a solution of (31) for any initial guess  $\mathbf{x}^{(0)}$  in  $\|\mathbf{x}^{(0)} - \alpha\| \leq \rho$ . In fact, under slightly different assumptions we could even demonstrate the existence of a solution if  $\|\mathbf{g}(\mathbf{x}^{(0)}) - \mathbf{x}^{(0)}\|$  is sufficiently small. However, we shall not present such specific theorems but instead shall indicate the relevant arguments and concern ourselves with the determination of the appropriate  $\theta_i$  to be used in (33). These considerations, in turn, suggest a modification in which the  $\theta_i$  are changed at each iteration.

If the error vector after the  $\nu$ th iteration is denoted by

$$\mathbf{e}^{(\nu)} = \mathbf{x}^{(\nu)} - \alpha,$$

then from (33a):

$$\mathbf{e}^{(\nu+1)} = \Theta[\mathbf{g}(\mathbf{x}^{(\nu)}) - \mathbf{g}(\alpha)] + (I - \Theta)\mathbf{e}^{(\nu)}.$$

However, by Taylor's theorem we then have

$$(35) \quad \mathbf{e}^{(\nu+1)} = (I - \Theta + \Theta G_\nu) \mathbf{e}^{(\nu)} \equiv M_\nu \mathbf{e}^{(\nu)}$$

where  $G_\nu = (g_{ij})$  and the  $i$ th row of  $G_\nu$  is evaluated at some point  $\xi^{(i, \nu)} = \mathbf{x}^{(\nu)} - \phi_i(\mathbf{x}^{(\nu)} - \alpha)$ ,  $0 < \phi_i < 1$ , for  $i = 1, 2, \dots, n$ . Clearly, the iterations will converge if the coefficient matrices in (35) satisfy  $\|M_\nu\| \leq q < 1$  for all  $\nu$ . Now if we use the maximum vector norm and corresponding induced matrix norm, these inequalities are satisfied if

$$(36) \quad R_i(\theta_i) \equiv |1 - \theta_i[1 - g_{ii}(\xi^{(i, \nu)})]| + |\theta_i| \sum_{j \neq i} |g_{ij}(\xi^{(i, \nu)})| \leq q < 1, \\ 1 \leq i \leq n; \quad \nu = 0, 1, 2, \dots.$$

Thus, we are led to consider inequalities of the form

$$R(\theta) \equiv |1 - \theta a| + |\theta| b < 1, \quad b \geq 0.$$

It is easily shown that  $R(\theta) < 1$  if  $|a| > b$  and  $\theta$  is in the interval:

$$(37a) \quad 0 < \theta < \frac{2}{a+b} \quad \text{if } a > b;$$

$$(37b) \quad \frac{2}{a-b} < \theta < 0 \quad \text{if } -a > b.$$

Furthermore, in each of these intervals the minimum value of  $R(\theta)$  is attained at

$$(37c) \quad \theta^* = \frac{1}{a}, \quad \text{and} \quad R^* = R(\theta^*) = \left| \frac{b}{a} \right|.$$

These results are easily deduced by considering the graphs of  $|\theta|b$  and  $|1 - \theta a|$  as functions of  $\theta$ . It is also clear from such graphs that underestimates of  $|\theta^*|$  produce a smaller  $R$  than do overestimates (see Figure 1 for the case  $1 > a > b > 0$ ).

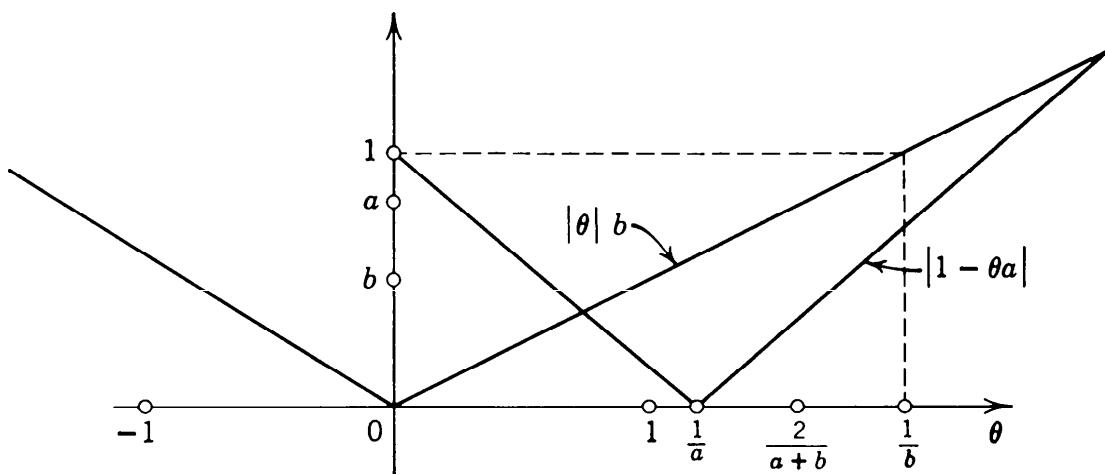


Figure 1

Employing (37) in (36) we find that: *The scheme (33) converges if (34) is satisfied and the acceleration parameters  $\theta_i$  lie in the intervals*

$$(38a) \quad 0 < \theta_i < \frac{2}{1 - g_{ii}(\xi) + r_i(\xi)}, \quad \text{if } 1 - g_{ii}(\xi) > r_i(\xi);$$

$$(38b) \quad \frac{2}{1 - g_{ii}(\xi) - r_i(\xi)} < \theta_i < 0, \quad \text{if } 1 - g_{ii}(\xi) < -r_i(\xi);$$

where  $r_i(\xi) \equiv \sum_{j \neq i} |g_{ij}(\xi)|$ . [It should be noted that by the assumed continuity of the  $g_{ij}(\mathbf{x})$  and (34), if  $r_i(\mathbf{x}) \neq 0$  then only one of the ranges (38) can apply for each  $i = 1, 2, \dots, n$ .]

From a graph of any  $R_i(\theta_i)$ , it is clear that the value of  $\theta_i$  should not lie near the end points of the intervals in (38). In fact, from (37c) it is suggested that  $\theta_i = 1/[1 - g_{ii}(\xi)]$  is the “best” value for  $\theta_i$ . However, since this value depends upon  $\xi$ , a safe choice,  $\theta_i^*$ , which can be rigorously justified is that for which (38) holds

$$(39) \quad |\theta_i^*| = \min_{\xi} \frac{1}{|1 - g_{ii}(\xi)|}, \quad |\xi - \alpha| \leq \rho.$$

These considerations suggest a modification of the scheme (33) in which an approximation to the best  $\theta_i$  is used at each step of the iteration. In fact, if  $\mathbf{x}^{(v)}$  is close to a solution then the values

$$(40) \quad \theta_i^{(v)} = \frac{1}{1 - g_{ii}(\mathbf{x}^{(v)})}, \quad i = 1, 2, \dots, n,$$

can be used in (33), to replace the constants  $\theta_i$ , and this practical scheme is of the form:

$$(41) \quad \mathbf{x}^{(v+1)} = \Theta^{(v)} \mathbf{g}(\mathbf{x}^{(v)}) + (I - \Theta^{(v)}) \mathbf{x}^{(v)}, \quad v = 0, 1, \dots$$

In carrying out this procedure, we need only evaluate the  $n$  partial derivatives,  $g_{ii} = \partial g_i / \partial x_i$ , at each step to predict the appropriate  $\theta_i^{(v)}$ . If these derivatives are not easily obtained or evaluated, one may frequently use difference approximations which just require one extra evaluation of each of the functions  $g_i(\mathbf{x})$ , i.e., since  $g_i(\mathbf{x}^{(v)})$  is known we use

$$g_{ii}(\mathbf{x}^{(v)}) \approx \frac{g_i(x_1^{(v)}, \dots, x_{i-1}^{(v)}, x_i^{(v)} + h, x_{i+1}^{(v)}, \dots, x_n^{(v)}) - g_i(\mathbf{x}^{(v)})}{h}$$

with some suitably small value of  $h$ .

Although the conditions (34) seem severe and perhaps unusual they are frequently satisfied in practice. In fact, many difference methods for solving non-linear boundary value problems in ordinary and partial differential equations result in such systems. For many of these systems, convergence can even be obtained for some choice  $\theta_1 = \theta_2 = \dots = \theta_n = \theta$  (for example, see a related scheme in Chapter 8, Subsection 7.2).

## PROBLEMS, SECTION 3

1. State and prove a generalization of Theorem 1.3 for systems of equations.
2. State and prove a version of Theorem 2 which employs a different norm (say  $\|\cdot\|_2$  or  $\|\cdot\|_1$ ).
3. For  $n = 1$ , use the hypothesis of Theorem 3, with  $h \equiv abc \leq \frac{1}{2}$  and show directly that there exists a root  $\alpha$  with  $|\alpha - x_0| \leq 2b$ .

[Hint: Use Taylor's theorem,

$$f(x) = f(x_0) + (x - x_0)f'(x_0) + \frac{(x - x_0)^2}{2}f''(\xi)$$

to show that

$$\frac{f(x_0 - 2b)}{f'(x_0)} \leq 0 \leq \frac{f(x_0 + 2b)}{f'(x_0)}.$$

4. Under the same assumptions as in Problem 3, show that the root  $\alpha$  with  $|\alpha - x_0| \leq 2b$  is unique.

[Hint: If not unique, there exists  $\eta$  with  $|\eta - x_0| < 2b$  and  $f'(\eta) = 0$ . But from Taylor's formula

$$f'(\eta) = f'(x_0) \left[ 1 + \frac{(\eta - x_0)f''(\zeta)}{f'(x_0)} \right],$$

show that  $f'(\eta) \neq 0$ .]

5. Under assumptions of Problem 3, with  $h < \frac{1}{2}$ , show that  $f'(\alpha) \neq 0$  (use hint of Problem 4).

6. Under the assumptions of Problem 3, if  $f'(\alpha) = 0$ , show that  $|\alpha - x_0| = 2b$  (by hint of Problem 4). Furthermore, if  $f'(\alpha) = 0$ , show that  $f''(\alpha) \neq 0$ .

[Hint:  $f''(\alpha) = \lim_{\eta \rightarrow \alpha} f'(\eta)/(\eta - \alpha)$ , but by hint of Problem 4 and  $\eta < \alpha$ ,

$$|f'(\eta)| > |f'(x_0)| \left( 1 - \frac{|\eta - x_0|}{2b} \right) = |f'(x_0)| \frac{|\eta - \alpha|}{2b}.$$

## 4. SPECIAL METHODS FOR POLYNOMIALS

All of the previous schemes for single equations can be employed to compute the roots of polynomials. Complex roots can be obtained by simply using complex arithmetic and complex initial estimates. Or, by reducing the evaluation of a polynomial at a complex point to its real and imaginary parts, the iterative methods for real systems (of order two in this case) could also be used to obtain the complex roots of polynomials. However, it is possible to devise special iterative methods which are frequently more advantageous than the general methods. We shall consider some of these polynomial methods in this section.

It is of interest to note, first, that a very simple a posteriori test of the accuracy of an approximate root of a polynomial is frequently quite effective. Let the  $n$ th degree polynomial  $P_n(x)$  be

$$(1) \quad P_n(x) \equiv a_0x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

with roots  $r_1, r_2, \dots, r_n$ . Then since

$$(2) \quad P_n(x) = a_0(x - r_1) \cdots (x - r_n)$$

we have the well-known result that

$$(3) \quad \frac{a_n}{a_0} = (-1)^n r_1 r_2 \cdots r_n.$$

Now let  $\sigma$  be an approximate root of  $P_n(x)$  which satisfies the test

$$(4) \quad |P_n(\sigma)| \leq \epsilon.$$

Then from (2) and (3) it follows that (we assume  $a_n \neq 0$ )

$$\left| \frac{P_n(\sigma)}{a_n} \right| = \left| 1 - \frac{\sigma}{r_1} \right| \cdot \left| 1 - \frac{\sigma}{r_2} \right| \cdots \left| 1 - \frac{\sigma}{r_n} \right| \leq \frac{\epsilon}{|a_n|}.$$

Taking the  $n$ th root we now conclude, since

$$\left| \frac{P_n(\sigma)}{a_n} \right| \geq \min_j \left| 1 - \frac{\sigma}{r_j} \right|^n,$$

that

$$(5) \quad \min_j \left| 1 - \frac{\sigma}{r_j} \right| \leq \left( \frac{\epsilon}{|a_n|} \right)^{1/n}.$$

Thus we obtain an *exact bound on the relative error* of  $\sigma$  as an approximation to some root of  $P_n(x)$ . In many of the methods to be studied,  $P_n(\sigma)$  is already computed and so no extra calculations are required to employ this test. Note that the roots and approximations in this test may be real or complex.

#### 4.1. Evaluation of Polynomials and Their Derivatives

An interesting special feature of polynomials is the ease with which they may be evaluated. Let us write (1) as

$$(6) \quad \begin{aligned} P_n(x) &= a_0 x^n + a_1 x^{n-1} + \cdots + a_{n-1} x + a_n \\ &= \{ \cdots [(a_0 x + a_1) x + a_2] x + \cdots \} x + a_n. \end{aligned}$$

The usual way to evaluate  $P_n(x)$  is by means of this “nesting” procedure. More explicitly, to calculate  $P_n(\xi)$  we form:

$$(7) \quad \begin{aligned} b_0 &= a_0, \\ b_1 &= b_0 \xi + a_1, \\ b_2 &= b_1 \xi + a_2, \\ &\vdots \\ b_\nu &= b_{\nu-1} \xi + a_\nu, \quad \nu = 1, 2, \dots, n; \end{aligned}$$

and note that  $b_n = P_n(\xi)$ . These operations are just those employed in the elementary process of *synthetic division*. In fact, if we write

$$P_n(x) = (x - \xi)Q_{n-1}(x) + R_0,$$

then clearly,  $R_0 = P_n(\xi) = b_n$  and it is easily verified (by multiplying out and equating coefficients of like powers of  $x$ ) that, using the quantities in (7):

$$(8) \quad Q_{n-1}(x) = b_0x^{n-1} + b_1x^{n-2} + \cdots + b_{n-1}.$$

Dividing again by  $(x - \xi)$  we get, say,

$$Q_{n-1}(x) = (x - \xi)Q_{n-2}(x) + R_1$$

and hence

$$P_n(x) = (x - \xi)^2Q_{n-2}(x) + (x - \xi)R_1 + R_0.$$

Differentiating the last expression, we find that  $R_1 = P'_n(\xi)$ . Thus by performing synthetic division of  $Q_{n-1}(x)$  by  $x - \xi$ , we could determine  $Q_{n-2}(x)$  and  $P'_n(\xi)$ . Clearly this procedure can be continued to yield finally

$$(9) \quad P_n(x) = R_n(x - \xi)^n + \cdots + R_1(x - \xi) + R_0.$$

The successive calculations to determine the  $R_v$  and coefficients of the intermediate polynomials  $Q_v(x)$  can be indicated by the array in Table 1.

**Table 1**

	0	0		0
$a_0$	$b_0$	$c_0$	$\dots$	$R_n$
$a_1$	$b_1$	$c_1$		$R_{n-1}$
$\vdots$	$\vdots$	$\vdots$		$\ddots$
$a_{n-2}$	$b_{n-2}$	$c_{n-2}$	$\ddots$	
$a_{n-1}$	$b_{n-1}$	$R_1$		
$a_n$	$R_0$			

Any entry of Table 1 not in the first row or column (which are given initially), is computed by multiplying the entry above by  $\xi$  and adding the entry to the left. It follows from (9) by differentiation that

$$R_v = \frac{1}{v!} \left. \frac{d^v P_n(x)}{dx^v} \right|_{x=\xi}; \quad v = 0, 1, \dots, n.$$

From (9) it also follows easily that *the polynomial with coefficients  $R_v$ , say*

$$R(y) \equiv R_n y^n + R_{n-1} y^{n-1} + \cdots + R_1 y + R_0,$$

*has as roots all those of  $P(x)$  reduced by the amount  $\xi$ .*

Finally we note that the evaluation of the entire Table 1 requires  $\frac{1}{2}(n^2 + n)$  multiplications and as many additions since the evaluation of the first column requires only  $n$  of each operation. But in computing the entries of the table, significant figures may be lost in any of the additions; see eqs. (7). Often this necessitates the use of multiple precision arithmetic.

To employ Newton's method on polynomial equations, only the first two columns in Table 1 need be computed. This is easily accomplished by means of the recursions (7) and a similar set with  $a_v, b_v$  replaced by  $b_v, c_v$  for  $v = 0, 1, \dots, n - 1$ . An interesting application of Newton's method to polynomials is found in Problem 1.

## 4.2. Sturm Sequences

It would be very desirable to obtain successive upper and lower bounds on the *real roots* of a polynomial equation or indeed of any equation, as then the error in approximating the root is easily estimated. This could be done if the number of real roots of the equation in any interval could be determined, and this can in fact be done by means of the so-called Sturm sequences which we proceed to define. Let the equation to be solved be  $f_0(x) = 0$  where  $f_0(x)$  is differentiable in  $[a, b]$ . Then the continuous functions

$$f_0(x), f_1(x), \dots, f_m(x)$$

form a Sturm sequence on  $[a, b]$  if they satisfy there:

- (i)  $f_0(x)$  has at most simple roots in  $[a, b]$ ;
- (ii)  $f_m(x)$  does not vanish in  $[a, b]$ ;
- (iii) if  $f_v(\alpha) = 0$ , then  $f_{v-1}(\alpha)f_{v+1}(\alpha) < 0$  for any root  $\alpha \in [a, b]$ ;
- (iv) if  $f_0(\alpha) = 0$ , then  $f_0'(\alpha)f_1(\alpha) > 0$  for any root  $\alpha \in [a, b]$ .

For every such sequence there follows Sturm's

**THEOREM 1.** *The number of zeros of  $f_0(x)$  in  $(a, b)$  is equal to the difference between the number of sign variations in  $\{f_0(a), f_1(a), \dots, f_m(a)\}$  and in  $\{f_0(b), f_1(b), \dots, f_m(b)\}$  provided that  $f_0(a)f_0(b) \neq 0$  and  $\{f_0(x), f_1(x), \dots, f_m(x)\}$  form a Sturm sequence on  $[a, b]$ .*

*Proof.* The number of sign variations can change as  $x$  goes from  $a$  to  $b$  only by means of some function changing sign in the interval. By (ii) it cannot be  $f_m(x)$ . Assume that at some  $\hat{x} \in (a, b)$ ,  $f_v(\hat{x}) = 0$  for some  $v$  in  $0 < v < m$ . In a neighborhood of  $\hat{x}$ , the signs must be either

$x$	$f_{v-1}(x)$	$f_v(x)$	$f_{v+1}(x)$	or	$x$	$f_{v-1}(x)$	$f_v(x)$	$f_{v+1}(x)$
$\hat{x} - \epsilon$	+	$\pm$	-		$\hat{x} - \epsilon$	-	$\pm$	+
$\hat{x}$	+	0	-		$\hat{x}$	-	0	+
$\hat{x} + \epsilon$	+	$\pm$	-		$\hat{x} + \epsilon$	-	$\pm$	+

The signs in the row  $x = \hat{x}$  follow from (iii). The signs in the first and last columns then follow by continuity for sufficiently small  $\epsilon$ . The sign possibilities in the middle columns are general. But we see from these tables that as  $x$  passes through  $\hat{x}$ , there are no changes in the number of sign variations in the Sturm sequence. We now examine the signs near a zero  $x = \hat{x}$  of  $f_0(x)$ :

$x$	$f_0(x)$	$f_1(x)$	or	$x$	$f_0(x)$	$f_1(x)$
$\hat{x} - \epsilon$	+	-		$\hat{x} - \epsilon$	-	+
$\hat{x}$	0	-		$\hat{x}$	0	+
$\hat{x} + \epsilon$	-	-		$\hat{x} + \epsilon$	+	+

The  $f_0(x)$  columns represent the two possible cases for a simple zero. The sign of  $f_1(\hat{x})$  then follows from (iv) and the continuity implies the other signs in the last columns. Clearly, there is now a decrease of one change in the number of sign variations as  $x$  increases through a zero of  $f_0(x)$ . When these results are combined the theorem follows. ■

It is easy to construct a Sturm sequence when  $f_0(x) \equiv P_n(x)$  is a polynomial, say of degree  $n$ . We define

$$f_1(x) \equiv f_0'(x)$$

so that (iv) is satisfied at simple roots. Divide  $f_0(x)$  by  $f_1(x)$  and call the remainder  $-f_2(x)$ . Then divide  $f_1(x)$  by  $f_2(x)$  and call the remainder  $-f_3(x)$ . Continue this procedure until it terminates [which it must since each polynomial  $f_v(x)$  is of lower degree than its predecessor,  $f_{v-1}(x)$ ]. We thus have the array:

$$\begin{aligned}
 f_0(x) &= q_1(x)f_1(x) - f_2(x), \\
 f_1(x) &= q_2(x)f_2(x) - f_3(x), \\
 &\vdots \\
 f_{m-2}(x) &= q_{m-1}(x)f_{m-1}(x) - f_m(x), \\
 f_{m-1}(x) &= q_m(x)f_m(x).
 \end{aligned}
 \tag{10}$$

This procedure is well known as the Euclidean algorithm for determining the highest common factor,  $f_m(x)$ , of  $f_0(x)$  and  $f_1(x)$ . [It is easily seen that

$f_m(x)$  is a factor of all the  $f_v(x)$ ,  $v = 0, 1, \dots, m - 1$ ; conversely any common factor of  $f_0(x)$  and  $f_1(x)$  must be a factor of  $f_v(x)$ ,  $v = 2, 3, \dots, m$ .] Thus all multiple roots of  $f_0(x)$  are also roots of  $f_m(x)$  with multiplicity reduced by one. If  $f_m(x)$  is not a constant (i.e.,  $f_0(x)$  has multiple roots) then we may divide all the  $f_v(x)$  by  $f_m(x)$  and (denoting these quotients by their numerators) obtain the sequence (10) in which  $f_0(x)$  has only simple roots and  $f_m(x)$  is a constant. It now follows that this reduced sequence  $\{f_0(x), f_1(x), \dots, f_m(x)\}$  is a Sturm sequence. Only (iii) requires proof: If  $f_v(\hat{x}) = 0$  then, by (10), at this point  $f_{v-1}(\hat{x}) = -f_{v+1}(\hat{x})$ ; but if  $f_{v-1}(\hat{x}) = 0$ , then also  $f_0(\hat{x}) = f_1(\hat{x}) = 0$ , a contradiction. Simple formulas for computing the coefficients of the  $f_v(x)$  can be obtained from (10); we present this as Problem 2.

The usual way to employ a Sturm sequence is with successive bisections of initially chosen intervals. In this way, with each new evaluation of the sequence, the error in determining a real root is at least halved. Thus this procedure converges with an asymptotic convergence factor of at least  $\frac{1}{2}$ . The value of Sturm sequences is clearly not in rapid convergence properties but rather in the ability to obtain good estimates of all real roots. When the desired root or roots have been located it is more efficient to employ a more rapidly converging iteration, say false position. Or, in fact, when a root is known to lie in the given interval,  $(a, b)$ , because  $f(a) \cdot f(b) < 0$ , we may simply calculate  $f((a + b)/2)$ . Now,  $(a + b)/2$  may be a root. If  $(a + b)/2$  is not a root, then we would know from the sign of  $f((a + b)/2)$  in which of the two sub-intervals,  $(a, (a + b)/2)$  or  $((a + b)/2, b)$ ,  $f(x)$  does have a root. In this way, we may continue to bisect successive sub-intervals in which we know  $f(x)$  to have a root. This procedure is known as the *bisection method*. It has the convergence factor  $\frac{1}{2}$ . Each step of the bisection method requires fewer calculations than does the evaluation of the Sturm sequence. At each step of the bisection method we have upper and lower bounds for a real root.

### 4.3. Bernoulli's Method

Consider the polynomial equation

$$(11) \quad P(x) \equiv x^n + a_1x^{n-1} + \dots + a_{n-1}x + a_n = 0$$

and assume that its roots,  $r_j$ , are distinct and ordered by

$$(12) \quad |r_1| > |r_2| > \dots > |r_n|.$$

This equation and its roots bear an important relationship to the difference equation or recursion (see Chapter 8, Section 4):

$$(13a) \quad g_v = -(a_1g_{v-1} + \dots + a_ng_{v-n}), \quad v = n, n + 1, \dots$$

$$(13b) \quad g_0 = c_0, \dots, g_{n-1} = c_{n-1}.$$

Given the  $n$  starting values,  $c_j$ , the  $g_v$  are easily evaluated (with  $n$  multiplications and  $n - 1$  additions per step). By seeking a formal solution of (13a) of the form  $g_v = r^v$ , we find that any root of (11) yields a solution. Furthermore, since the difference equations are linear it easily follows that any linear combination of the powers of the roots of (11) is also a formal solution of (13a): thus we may write:

$$(14) \quad g_v = b_1 r_1^v + b_2 r_2^v + \cdots + b_n r_n^v; \quad v = 0, 1, 2, \dots$$

The polynomial (11) is called the *characteristic equation* of the *difference equation* (13a). The conditions (13b) yield a linear system of  $n$  equations for the determination of the  $b_j$ . The determinant of the coefficient matrix of this system is a Vandermonde determinant which, by (12), does not vanish [see (2.4) in Chapter 5]. It can also be shown that (14) is the most general solution of (13a), and hence with the  $b_j$  as determined above, (14) is the unique solution of (13a, b) (see Problem 3).

For  $v \gg 1$  we have, recalling (12),  $g_v \approx b_1 r_1^v$ , and hence we are led to consider the sequence

$$(15) \quad \sigma_v \equiv \frac{g_{v+1}}{g_v} = r_1 \frac{b_1 + b_2 \left( \frac{r_2}{r_1} \right)^{v+1} + \cdots + b_n \left( \frac{r_n}{r_1} \right)^{v+1}}{b_1 + b_2 \left( \frac{r_2}{r_1} \right)^v + \cdots + b_n \left( \frac{r_n}{r_1} \right)^v}$$

$$= r_1 + \mathcal{O}\left(\left|\frac{r_2}{r_1}\right|^v\right).$$

Here we have assumed that  $b_1 \neq 0$  and, in this case, clearly  $\lim_{v \rightarrow \infty} \sigma_v = r_1$ . The rapidity of the approach is determined by the ratio  $|r_2/r_1|$ . If this ratio is not near unity,  $r_1$  is easily obtained and can be eliminated from the original polynomial by synthetic division as in Subsection 4.1 and then a new recursion is evaluated to determine  $r_2$ , etc. (In practice, elimination of the roots in decreasing order of magnitude may produce considerably larger errors than elimination in increasing order of magnitude. See Problem 6.)

If  $b_1 = 0$  for an unfortunate choice of starting values (13b), it would seem that the ratios then converge to  $r_2$ . This is theoretically correct but for actual computations the roundoff introduced in the successive evaluations of (13a) has the effect of altering the exact  $b_j$  which should occur in (14). Thus after a few recursions there will be some perhaps small but non-zero component  $b_1$  present and in subsequent steps the dominant root  $r_1$  may still be determined. In like manner, if some error or blunder is committed in the course of these Bernoulli iterations the subsequent steps,

performed correctly, will obliterate the error and yield the correct result.

The expansion in (15) shows that the convergence of  $\sigma_v$  to  $r_1$  is geometric with ratio  $|r_2/r_1|$ . To test the convergence of the sequence  $\{\sigma_v\}$ , a number of devices can be employed. Perhaps the most frequently used procedure in practice is to compute the differences  $|\sigma_{v+1} - \sigma_v|$  and to stop when these are less than some predetermined small quantity. Another possibility which yields more precise information at the cost of some extra computation is to use the test (4). Of course, if sufficiently many steps have been performed,  $\sigma = \sigma_v$  is closest to  $r_1$ , the dominant root, and so  $j = 1$  in (5). Test (4) should not be made after every iteration but say every several steps to reduce the computational effort.

The conditions (12) are most likely not satisfied for a polynomial since in general some conjugate complex roots are to be expected. Suppose then that the dominant real roots have been eliminated and (11) is the reduced equation with  $r_1$  and  $r_2$  conjugate complex roots,  $r_2 = \bar{r}_1$ , satisfying

$$(16) \quad |r_1| = |r_2| > |r_3| > \dots > |r_n|.$$

The solution of (13) is again of the form (14) where the  $b_j$  may be complex and while (15) is valid the  $\sigma_v$  do not converge to  $r_1$  since  $|r_2/r_1| = 1$ . For large  $v$  we now expect that

$$g_v \approx b_1 r_1^v + b_2 r_2^v.$$

Here we note that  $r_2 = \bar{r}_1$  and  $b_2 = \bar{b}_1$  since the  $g_v$  must be real (assuming that the  $a_j$  and  $c_j$  are real). A simple calculation now reveals that

$$A_v \equiv g_{v+1}g_{v-1} - g_v^2 \approx |b_1|^2(r_1 - r_2)^2(r_1 r_2)^{v-1},$$

$$B_v \equiv g_{v+2}g_{v-1} - g_{v+1}g_v \approx |b_1|^2(r_1 - r_2)^2(r_1 r_2)^{v-1}(r_1 + r_2).$$

Thus we expect that with  $s_v \equiv B_v/A_v$  and  $p_v \equiv A_{v+1}/A_v$ :

$$\lim_{v \rightarrow \infty} s_v = r_1 + r_2 \equiv s, \quad \lim_{v \rightarrow \infty} p_v = r_1 r_2 \equiv p,$$

and the roots  $r_1$  and  $r_2$  are those of the quadratic equation

$$\xi^2 - s\xi + p = 0.$$

Recalling (16) we can now estimate the error in this procedure for complex roots. Let  $|r_1| = |r_2| \equiv r$ ,  $|r_3|/r \equiv \delta$  and we find as above:

$$g_v = b_1 r_1^v + b_2 r_2^v + \mathcal{O}(r^v \delta^v),$$

$$A_v = |b_1|^2(r_1 - r_2)^2(r_1 r_2)^{v-1} + \mathcal{O}(r^{2v} \delta^v),$$

$$B_v = |b_1|^2(r_1 - r_2)^2(r_1 r_2)^{v-1}(r_1 + r_2) + \mathcal{O}(r^{2v} \delta^v).$$

Thus by the usual expansions

$$(17) \quad s_v = s + \mathcal{O}(\delta^v), \quad p_v = p + \mathcal{O}(\delta^v).$$

The equation actually solved is the quadratic

$$(18) \quad \xi^2 - s_v \xi + p_v = 0.$$

It is easily shown, since  $s^2 - 4p \neq 0$ , that the roots of this equation differ from those for  $v = \infty$  by terms  $\mathcal{O}(\delta^v)$ .

If the dominant roots are equal but of opposite sign, then the above procedure for complex roots is still applicable where now  $r_2 = -r_1$ . If the dominant root is real but multiple, then the original procedure is applicable but converges more slowly (see Problem 5).

#### 4.4. Bairstow's Method

A much better scheme (than Bernoulli's) for determining quadratic factors of a polynomial,  $P_n(x)$ , is based on a generalization of synthetic division and Newton's method for systems of equations. This procedure also avoids the use of complex arithmetic. In brief, we seek real numbers, say  $s$  and  $p$ , such that the quadratic

$$(19) \quad x^2 + sx + p$$

is an exact factor of  $P_n(x)$ . The division of  $P_n(x)$  by this factor may be indicated as:

$$(20) \quad P_n(x) = (x^2 + sx + p)Q_{n-2}(x) + [xR_1(s, p) + R_0(s, p)].$$

Here,  $R_1$  and  $R_2$  are the coefficients in the remainder which is at most linear in  $x$ . As indicated, these coefficients are functions of  $s$  and  $p$ , the parameters in the quadratic (19). In order that the remainder be zero,  $s$  and  $p$  must satisfy

$$(21) \quad R_1(s, p) = 0, \quad R_0(s, p) = 0.$$

This is a system of two equations in two unknowns which we solve by Newton's method. For this purpose we must evaluate the four derivatives

$$(22) \quad \begin{aligned} & \frac{\partial R_1}{\partial s}, \quad \frac{\partial R_0}{\partial s} \\ & \frac{\partial R_1}{\partial p}, \quad \frac{\partial R_0}{\partial p}. \end{aligned}$$

We determine the quantities in (22) indirectly. Let  $Q_{n-2}(x)$  in (20) be divided by the quadratic (19) to yield

$$(23) \quad Q_{n-2}(x) = (x^2 + sx + p)Q_{n-4}(x) + [xR_3(s, p) + R_2(s, p)].$$

We note that the specific values of the  $R_j(s, p)$  in (20) and (23) for any fixed  $(s, p)$  are easily obtained by carrying out the two indicated divisions.

Using (23) in (20) we have the identity:

$$(24) \quad P_n(x) = (x^2 + sx + p)^2 Q_{n-4}(x) \\ + (x^2 + sx + p)[xR_3(s, p) + R_2(s, p)] \\ + [xR_1(s, p) + R_0(s, p)].$$

Since  $P_n(x)$  is independent of  $s$  and  $p$ , we can differentiate (24) with respect to  $s$  and  $p$  and evaluate the result at a root  $x = z$  of  $z^2 + sz + p = 0$  to get:

$$z(zR_3 + R_2) + \left( z \frac{\partial R_1}{\partial s} + \frac{\partial R_0}{\partial s} \right) = 0, \\ (zR_3 + R_2) + \left( z \frac{\partial R_1}{\partial p} + \frac{\partial R_0}{\partial p} \right) = 0.$$

Since  $z^2 = -(sz + p)$  these equations can be written as

$$z \left( \frac{\partial R_1}{\partial s} + R_2 - sR_3 \right) + \left( \frac{\partial R_0}{\partial s} - pR_3 \right) = 0, \\ z \left( \frac{\partial R_1}{\partial p} + R_3 \right) + \left( \frac{\partial R_0}{\partial p} + R_2 \right) = 0.$$

If the quadratic is not a perfect square (in which case the roots would be real and equal), the above must hold for two distinct roots,  $z$ , and hence each coefficient in parentheses must vanish. Thus we deduce that:

$$(25) \quad \begin{aligned} \frac{\partial R_1}{\partial s} &= sR_3 - R_2, & \frac{\partial R_0}{\partial s} &= pR_3; \\ \frac{\partial R_1}{\partial p} &= -R_3, & \frac{\partial R_0}{\partial p} &= -R_2. \end{aligned}$$

The iteration scheme proceeds from an initial estimate  $(s_0, p_0)$  which is such that  $p_0 \neq s_0^2/4$  and defines recursively the sequence  $(s_v, p_v)$  by Newton's method of solving (21); i.e.,

$$(26) \quad \begin{aligned} (s_{v+1} - s_v) \frac{\partial R_1(s_v, p_v)}{\partial s} + (p_{v+1} - p_v) \frac{\partial R_1(s_v, p_v)}{\partial p} &= -R_1(s_v, p_v), \\ (s_{v+1} - s_v) \frac{\partial R_0(s_v, p_v)}{\partial s} + (p_{v+1} - p_v) \frac{\partial R_0(s_v, p_v)}{\partial p} &= -R_0(s_v, p_v). \end{aligned}$$

The coefficients and inhomogeneous terms in (26) are obtained by carrying out the two divisions indicated in (20) and (23) with the quadratic factor  $x^2 + s_v x + p_v$  and then using (25). The divisions can be reduced to the

evaluation of simple recursions by equating the coefficients on both sides of the equality signs in (20) and (23); we leave the derivation of this generalized synthetic division as an exercise. During the course of the iterations it should be checked that  $p_v \neq s_v^2/4$ , so that (25) is valid. To test the convergence we may employ (4) and (5), after noting that if  $z$  is a root of  $x^2 + sx + p = 0$  then, by (20),

$$P_n(z) = zR_1(s, p) + R_0(s, p).$$

Usually the coefficients  $(s_v, p_v)$  converge quickly, and so a direct test on the roots need only be applied when the iterations are about to be terminated.

## PROBLEMS, SECTION 4

- 1.** The square roots of a positive number  $\beta$  are the zeros of the polynomial  $f(x) \equiv x^2 - \beta$ .

- (a) Apply Newton's method to obtain the sequence of approximations

$$x_{v+1} = \frac{1}{2} \left( x_v + \frac{\beta}{x_v} \right).$$

This procedure is known as the Newton-Raphson method for computing square roots; it is frequently employed on high-speed digital computers.

- (b) If  $x_0 > \sqrt{\beta}$  show that:  $x_{v+1} < x_v$ ,  $v = 0, 1, \dots$  (assuming exact calculations).

- (c) Derive and study the analogous procedure for the  $n$ th root of any positive number where  $n$  is an integer.

- 2.** Derive a recursion formula for finding the coefficients of the Sturm sequence (10). Assume  $f_0(x) = \sum_{j=0}^n a_{j,0} x^{n-j}$ .

- 3.** Show that every solution of (13a, b) is unique and hence can be represented in the form (14).

- 4.** If the coefficients  $a_j$  in the polynomial  $P_n(x) \equiv x^n + a_1 x^{n-1} + \dots + a_n$  are altered by an amount at most  $\epsilon$ , then the polynomial  $P_{n,\epsilon}(x) \equiv P_n(x) + \epsilon g_{n-1}(x)$  is obtained where  $g_{n-1}(x) \equiv b_1 x^{n-1} + b_2 x^{n-2} + \dots + b_{n-1}$  and  $|b_i| \leq 1$ .

- Show that to each simple real root  $r_i$  of  $P_n(x)$ , corresponds a simple real root  $r_{i,\epsilon}$  of  $P_{n,\epsilon}(x)$  such that  $r_{i,\epsilon} - r_i = O(\epsilon)$  as  $\epsilon \rightarrow 0$ .

[Hint: Plot  $P_{n,\epsilon}(x)$  in a neighborhood of  $r_i$  for sufficiently small  $\epsilon$ .]

- 5.** Show how to modify Bernoulli's method for the case of a dominant real multiple root. Estimate the convergence rate.

- 6.** If  $P_n(x) \equiv x^n + a_1 x^{n-1} + \dots + a_{n-1} x + a_n$  with  $a_n \neq 0$ , then let  $Q_n(z) \equiv a_n z^n + a_{n-1} z^{n-1} + \dots + a_1 z + 1$ . Show that the roots  $\{z_k\}$  of  $Q_n(z) = 0$  and roots  $\{x_k\}$  of  $P_n(x) = 0$  are related by  $z_k = 1/x_k$ ;  $k = 1, 2, \dots, n$ . Hence, show how the Bernoulli method may be used to find the zeros of  $P_n(x)$  in ascending order of magnitude.

# 4

## Computation of Eigenvalues and Eigenvectors

### 0. INTRODUCTION

The *eigenvalue-eigenvector problem* for a square matrix  $A \equiv (a_{ij})$  of order  $n$  is that of determining a scalar  $\lambda$  and vector  $\mathbf{x}$  such that

$$(1) \quad A\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0}.$$

The problem is clearly non-linear since both  $\lambda$  and  $\mathbf{x}$  are unknowns. In fact, as is well known, the *eigenvalues*,  $\lambda$ , are the  $n$  roots of the *characteristic equation*

$$(2) \quad \det(\lambda I - A) \equiv p_A(\lambda) = 0.$$

Thus the eigenvalues can, in principle, be found as the zeros of  $p_A(\lambda)$  without recourse to any of the eigenvectors. Given some eigenvalue,  $\lambda$ , then the corresponding *eigenvector* is a non-trivial solution of the homogeneous linear system (1). On the other hand, if some particular eigenvector  $\mathbf{x}$  is known, then the eigenvalue to which  $\mathbf{x}$  belongs can be obtained by taking the inner product of (1) with  $\mathbf{x}$  to get

$$(3) \quad \lambda = \frac{\mathbf{x}^* A \mathbf{x}}{\mathbf{x}^* \mathbf{x}}.$$

Alternatively, we could use any non-zero component  $x_i$  to get, from the  $i$ th row of (1),

$$\lambda = \frac{1}{x_i} \sum_{j=1}^n a_{ij} x_j.$$

If some of the  $n$  eigenvalues of  $A$  are not distinct [i.e., if  $p_A(\lambda)$  has multiple roots] then there *may* be fewer than  $n$  eigenvectors. For example,

$$A \equiv \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}$$

has the repeated eigenvalue  $\lambda = 1$ , and only one eigenvector,

$$\mathbf{x} = \begin{pmatrix} 1 \\ 0 \end{pmatrix}.$$

We consider, in Section 1, the well-posedness of the eigenvalue problem and a posteriori error bounds.

In Section 2, we study the power method and its ramifications. These yield a sequence of scalars and vectors that converge (when the procedure works) to some particular eigenvalue and its corresponding eigenvector. In order to compute other eigenpairs, the iteration scheme must be modified. These simple methods yield successively only a few of the eigenvalues and vectors with acceptable accuracy, but for many applications this will be all that is needed.

The methods studied in Section 3 use a finite or infinite sequence of matrix transformations to find a similar matrix  $B = P^{-1}AP$ , such that the evaluation of  $\det(\lambda I - B)$  is simpler than the calculation of  $\det(\lambda I - A)$ . We then find the eigenvalues as the zeros of  $p_B(\lambda)$ , by means of an iterative scheme, which does not explicitly use the coefficients of  $p_B$ . In the methods based on the use of an infinite sequence of matrix transformations, the eigenvalues themselves are usually explicitly exhibited in  $B$ . A special calculation is then required to obtain the eigenvectors. In summary, these methods may yield all of the eigenvalues without determining any eigenvectors.

[We emphasize the practical importance of not finding explicitly the coefficients of  $p_B(\lambda)$  or  $p_A(\lambda)$  in order to evaluate the polynomials. All experienced practitioners are aware of the large errors that may result from the use of the approximate coefficients of  $p_B(\lambda)$  or  $p_A(\lambda)$  for the calculation of the zeros of the characteristic polynomials. We do not study the methods based on finding the coefficients of  $p_A(\lambda)$ .]

## 1. WELL-POSEDNESS, ERROR ESTIMATES

A simple criterion for *localizing eigenvalues* is given in

**THEOREM 1 (GERSCHGORIN).** *Let  $A \equiv (a_{ij})$  have eigenvalues  $\lambda$  and define the absolute row and column sums by*

$$r_i \equiv \sum_{\substack{j=1 \\ j \neq i}}^n |a_{ij}|, \quad c_j \equiv \sum_{\substack{i=1 \\ i \neq j}}^n |a_{ij}|.$$

*Then,*

(a) each eigenvalue lies in the union of the row circles  $R_i$ ,  $i = 1, 2, \dots, n$ , where

$$R_i \equiv \{z \mid |z - a_{ii}| \leq r_i\};$$

(b) each eigenvalue lies in the union of the column circles  $C_j$ ,  $j = 1, 2, \dots, n$ , where

$$C_j \equiv \{z \mid |z - a_{jj}| \leq c_j\};$$

(c) each component (maximal connected union of circles) of  $\bigcup_i R_i$  or  $\bigcup_j C_j$  contains exactly as many eigenvalues as circles. (The eigenvalues and the circles are counted with their multiplicities.)

*Proof.* (a) If  $\lambda$  is an eigenvalue of  $A$ , then there exists an eigenvector  $\mathbf{x} \neq \mathbf{0}$  such that

$$A\mathbf{x} = \lambda\mathbf{x},$$

or

$$(\lambda - a_{ii})x_i = \sum_{\substack{j=1 \\ j \neq i}}^n a_{ij}x_j, \quad i = 1, 2, \dots, n.$$

Pick an  $i$  such that  $|x_i| = \|\mathbf{x}\|_\infty \neq 0$ . Then

$$|\lambda - a_{ii}| \leq \sum_{j \neq i} |a_{ij}| \left| \frac{x_j}{x_i} \right| \leq r_i.$$

(b) Since the eigenvalues of  $A$  and  $A^T$  are identical, (b) follows from (a).

(c) Here we apply a simpler form of the basic lemma of the theory of functions of a complex variable quoted in Chapter 1, Section 3, namely:

**LEMMA 1.** *The  $n$  zeros of a polynomial*

$$p(x) \equiv x^n + a_1x^{n-1} + \cdots + a_{n-1}x + a_n$$

*are continuous functions of the coefficients  $\{a_j\}$ .*

Consider the one-parameter family of matrices  $A(t) = D + tB$ , where  $D \equiv (a_{ii}\delta_{ij})$  is a diagonal matrix and  $B = A - D$ . Consider, corresponding to  $A(t)$ , the circles  $R_i(t)$  and  $C_j(t)$  with respective radii  $tr_i$  and  $tc_j$  about respective centers  $a_{ii}$  and  $a_{jj}$ . Now,  $A(0) = D$  and clearly (c) holds for this diagonal matrix. The eigenvalues  $\{\lambda(t)\}$  of  $A(t)$  are the zeros of  $p_{A(t)}(\lambda) \equiv \det[\lambda I - A(t)]$ . As  $t$  increases continuously to  $t = 1$ , the integer number of circles and, by Lemma 1, of eigenvalues in each component of  $\bigcup_i R_i(t)$  [or of  $\bigcup_j C_j(t)$ ] varies continuously, and hence is constant, except for a finite number of values,  $t = t^l$ , at which the number

of circles in a component increases. At such values  $t = t^l$ , (c) holds because of Lemma 1. In other words, when the number of circles in a component increases, the number of eigenvalues in the component increases by the same amount. ■

Gershgorin's theorem has many applications; the first we make is in the treatment of the well-posedness of the eigenvalue problem. In this connection, it is necessary to introduce the notions of left and right eigenvectors.

The eigenvector  $\mathbf{x}$  that satisfies (0.1),

$$A\mathbf{x} = \lambda\mathbf{x}, \quad \mathbf{x} \neq \mathbf{0},$$

is more properly called a *right eigenvector* of  $A$ . Correspondingly a *left eigenvector*,  $\mathbf{y}$ , of  $A$  is a vector that satisfies

$$(1) \quad \mathbf{y}^* A = \mu \mathbf{y}^*, \quad \mathbf{y} \neq \mathbf{0}.$$

In other words,  $\mathbf{y}$  is a left eigenvector of  $A$  iff  $\mathbf{y}$  is a right eigenvector of  $A^*$ , i.e., by *starring* both sides (by taking the complex conjugate transpose of both sides),

$$A^*\mathbf{y} = \bar{\mu}\mathbf{y}.$$

( $\mathbf{y}^*$  is also called a *row eigenvector* of  $A$ .) The sets of “left”  $\{\mu\}$  and “right”  $\{\lambda\}$  eigenvalues are identical, since

$$\det(\lambda I - A) = \overline{\det(\bar{\lambda}I - \bar{A})} = \overline{\det(\bar{\lambda}I - A^*)},$$

where the last equality follows from  $\det B = \det B^T$ .

Now, when the matrix  $A$  is Hermitian, the left and right eigenvectors are also identical; otherwise, when  $A$  is not Hermitian, the left and right eigenvectors are distinct, in general.

**LEMMA 2.** *Any left eigenvector,  $\mathbf{y}$ , and any right eigenvector,  $\mathbf{x}$ , of the matrix  $A$ , corresponding to any pair of distinct eigenvalues, are orthogonal, i.e.,  $\mathbf{y}^*\mathbf{x} = 0$ .*

*Proof.* Given as Problem 1. ■

We say that the left and right eigenvectors are *biorthogonal*.

We may now ask, “How well-posed is the eigenvalue problem (0.1) or (1)?” The answer is given in a special case by

**THEOREM 2.** *Let  $A$  be of order  $n$  and have  $n$  linearly independent eigenvectors. For any fixed matrix  $C$ , with  $\|C\| = \|A\|$ , define the perturbed matrix*

$$(2) \quad A(\epsilon) \equiv A + \epsilon C.$$

Then, if  $\lambda$  is any eigenvalue of  $A$ , there is an eigenvalue  $\lambda(\epsilon)$  of  $A(\epsilon)$  such that

$$(3) \quad |\lambda(\epsilon) - \lambda| = O(\epsilon).$$

Moreover, if  $\lambda$  is simple

$$(4) \quad \lim_{|\epsilon| \rightarrow 0} \frac{\lambda(\epsilon) - \lambda}{\epsilon} = \frac{\mathbf{y}^* C \mathbf{x}}{\mathbf{y}^* \mathbf{x}},$$

where  $\mathbf{x}$  and  $\mathbf{y}$  are, respectively, right and left eigenvectors of  $A$  corresponding to  $\lambda$ .

*Proof.* Let  $P$  be the matrix whose columns are the right eigenvectors of  $A$ , i.e.,

$$(5) \quad AP = P\Lambda$$

where  $\Lambda \equiv (\lambda_{ij})$  is a diagonal matrix, such that  $\{\lambda_{ii}\}$  are the eigenvalues of  $A$  in some order. Therefore, from (5) and (2),

$$(6) \quad \begin{aligned} P^{-1}A(\epsilon)P &= \Lambda + \epsilon P^{-1}CP \equiv \Lambda + \epsilon E, \\ E &\equiv (e_{ij}) \equiv P^{-1}CP. \end{aligned}$$

The estimate (3) now follows from Gerschgorin's theorem part (a), since the eigenvalues of  $A(\epsilon)$  are unchanged by the similarity transformation.

We can now improve upon (3), if the eigenvalue  $\lambda = \lambda_{ii}$  is simple. That is, we observe that the circle  $R_i(\epsilon)$  of the matrix  $P^{-1}A(\epsilon)P$  will not intersect any other circles  $R_j(\epsilon)$ , if  $|\epsilon|$  is small enough, since  $\lambda_{ii}$  is distinct from all other  $\lambda_{jj}$ . Therefore, there is a unique simple eigenvalue  $\lambda(\epsilon)$  in  $R_i(\epsilon)$  and therefore also a unique corresponding eigenvector  $\mathbf{x}(\epsilon)$  of  $A(\epsilon)$ .

Now, the eigenvalues  $\lambda$ ,  $\lambda(\epsilon)$  and eigenvectors  $\mathbf{x}$ ,  $\mathbf{x}(\epsilon)$  satisfy

$$(7) \quad \begin{aligned} (A + \epsilon C)\mathbf{x}(\epsilon) &= \lambda(\epsilon)\mathbf{x}(\epsilon) \\ A\mathbf{x} &= \lambda\mathbf{x}. \end{aligned}$$

Therefore by subtraction,

$$(8) \quad A[\mathbf{x}(\epsilon) - \mathbf{x}] + \epsilon C\mathbf{x}(\epsilon) = [\lambda(\epsilon) - \lambda]\mathbf{x}(\epsilon) + \lambda[\mathbf{x}(\epsilon) - \mathbf{x}].$$

If we left-multiply both sides of (8) by the row vector  $\mathbf{y}^*$ , that satisfies  $\mathbf{y}^* A = \lambda \mathbf{y}^*$ , then

$$(9) \quad \epsilon \mathbf{y}^* C \mathbf{x}(\epsilon) = [\lambda(\epsilon) - \lambda] \mathbf{y}^* \mathbf{x}(\epsilon).$$

The conclusion (4) follows if we show that

$$\mathbf{y}^* \mathbf{x} \neq 0$$

and that we may select  $\mathbf{x}(\epsilon)$  and  $\mathbf{x}$  so that, as  $\epsilon \rightarrow 0$ ,

$$\mathbf{x}(\epsilon) \rightarrow \mathbf{x}.$$

The fact that  $\mathbf{y} \neq \mathbf{0}$  is orthogonal to each of the  $n - 1$  other right eigenvectors of  $A$ , establishes the non-orthogonality of  $\mathbf{y}$  and  $\mathbf{x}$ . From the relation (3), equation (7), and the assumed simplicity of  $\lambda$ , we know that the matrices  $A - \lambda I$  and  $A + \epsilon C - \lambda(\epsilon)I$  have rank  $n - 1$  and that we can delete the same row and the same column from each to form non-singular submatrices, if  $|\epsilon|$  is sufficiently small. Hence, if the deleted column is the  $j$ th, we may set  $x_j(\epsilon) = 1$  and then, from Problem 3,  $\mathbf{x}(\epsilon)$  will converge as  $|\epsilon| \rightarrow 0$ , to an eigenvector  $\mathbf{x}$  satisfying (7). ■

**COROLLARY (BAUER-FIKE).** *Under the same hypothesis and definitions as are used to establish (3), each eigenvalue  $\lambda(\epsilon)$  of  $A + \epsilon C$  satisfies*

$$(3') \quad \min_{\lambda} |\lambda(\epsilon) - \lambda| \leq |\epsilon| \|C\|_p \|P^{-1}\|_p \|P\|_p,$$

for any  $p$ -norm,  $\|\mathbf{x}\|_p = \left( \sum_i |x_i|^p \right)^{1/p}$ , with  $1 \leq p \leq \infty$ .

*Proof.* Given as Problem 5. ■

Observe that the algebraic Lemma of Section 3 in Chapter 1 suggests only that (3) holds if  $\lambda$  is a simple zero of  $p_A(\lambda)$ ; but if  $\lambda$  has multiplicity  $m$ , then  $|\lambda(\epsilon) - \lambda| = \mathcal{O}(\epsilon^{1/m})$ . On the other hand, although (9) holds for the general case, (4) may not hold because  $\mathbf{y}^* \mathbf{x}$  may vanish as shown in Problem 2.

We see that when (4) holds, the *well-posedness* of the eigenvalue problem for determining  $\lambda$  depends on the magnitude of

$$\left| \frac{\mathbf{y}^* C \mathbf{x}}{\mathbf{y}^* \mathbf{x}} \right|.$$

If we normalize the vectors, so that  $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2 = 1$ , and use the Schwarz inequality (see Problem 6), there results

$$(10) \quad \left| \frac{\mathbf{y}^* C \mathbf{x}}{\mathbf{y}^* \mathbf{x}} \right| \leq \frac{\|C\|_2}{|\mathbf{y}^* \mathbf{x}|}.$$

In the special case that  $C = \|A\|_2 U$ , where  $U$  is a unitary matrix (i.e.,  $U^* = U^{-1}$ ) such that  $U\mathbf{x} = \mathbf{y}$  (see Problems 9 and 12), we find that  $\|C\|_2 = \|A\|_2$  and furthermore

$$(10') \quad \left| \frac{\mathbf{y}^* C \mathbf{x}}{\mathbf{y}^* \mathbf{x}} \right| = \|A\|_2 \left| \frac{\mathbf{y}^* \mathbf{y}}{\mathbf{y}^* \mathbf{x}} \right| = \frac{\|A\|_2}{|\mathbf{y}^* \mathbf{x}|} = \frac{\|C\|_2}{|\mathbf{y}^* \mathbf{x}|}.$$

Now, it is possible that the problem of finding  $\lambda$ , an eigenvalue of  $A$ , is not well-posed, although the problem of finding the same  $\lambda$  as an eigenvalue of  $B = P^{-1}AP$  is well-posed (this fact is illustrated in Problem 13).

The reader may on first impulse think that this statement is contradicted by Theorem 3 given in Problem 11. But a closer examination of Theorem 3 indicates that although the perturbed matrices  $A + \epsilon C$  and  $B + \epsilon D$  are in the one to one correspondence  $D = P^{-1}CP$ , the magnitudes of the corresponding matrices,  $\|D\|$  and  $\|C\|$ , may differ considerably unless  $\|P\| \cdot \|P^{-1}\|$  is not very large, since (see Problem 7)

$$(11) \quad \frac{\|C\|}{\|P^{-1}\| \|P\|} \leq \|P^{-1}CP\| \leq \|C\| \|P^{-1}\| \|P\|.$$

On the other hand, in the test for well-posedness of  $\lambda$ , it is implied that we consider only perturbations such that  $\|C\| = \|A\|$  or such that  $\|D\| = \|B\|$ . Hence no contradiction is involved.

We may then justifiably say that when (4) holds for all  $C$  and  $|\mathbf{y}^* \mathbf{x}|$  is not small (for say  $\|\mathbf{y}\|_2 = \|\mathbf{x}\|_2 = 1$ ), the eigenvalue problem for  $\lambda$  of  $A$  is *well-posed*, since from (4), (10), and (10'), with  $\|C\|_2 = \|A\|_2$ ,

$$\max_{\{C\}} \lim_{\epsilon \rightarrow 0} \left| \frac{\lambda(\epsilon) - \lambda}{\epsilon} \right| = \frac{\|A\|_2}{|\mathbf{y}^* \mathbf{x}|}.$$

We have the immediate consequence of (10).

**THEOREM 4.** *If  $A$  is Hermitian and  $\lambda$  any eigenvalue of  $A$ , then the eigenvalue problem for  $\lambda$  is well-posed.*

*Proof.* If  $\lambda$  is not simple, use (7) and Problem 4 to find  $\mathbf{x}(\epsilon)$  and  $\mathbf{x}$  with  $\|\mathbf{x}(\epsilon) - \mathbf{x}\| \rightarrow 0$ . Now set  $\mathbf{y} = \mathbf{x}$  in (9) and note that  $\mathbf{y}^* \mathbf{x}(\epsilon) \rightarrow 1$ . ■

Fortunately, the most common matrix transformation methods use orthogonal or unitary similarity transformations of  $A$  to produce a simpler matrix  $B$  which, by accounting for roundoff errors, can be written as

$$B(\epsilon) = P^{-1}(A + \epsilon C)P.$$

The matrix  $P$  is unitary and, depending on the kind of arithmetic used,

$$(12) \quad \|C\|_2 \leq 1; \quad |\epsilon| = \mathcal{O}(n^p 10^{-t} a),$$

where  $a = \max_{i,j} |a_{ij}|$ ,  $p \leq 2$ , and  $t$  represents the number of figures used in single precision arithmetic. A priori error estimates of the form (12) have been obtained by Givens and Wilkinson. We shall not pursue this topic but rather give an account of some a posteriori error estimates.

### 1.1. A Posteriori Error Estimates

In the case of a Hermitian matrix,  $A$ , we can give a simple error estimate for a computed eigenvalue,  $\lambda$ , in terms of the *residual vector*,  $\eta$ .

**THEOREM 5.** *Let  $A$  be a Hermitian matrix of order  $n$  and have eigenvalues  $\{\lambda_i\}$ . If*

$$(13) \quad Ax - \lambda x \equiv \eta, \quad x \neq 0,$$

*then*

$$\min_i |\lambda - \lambda_i| \leq \frac{\|\eta\|_2}{\|x\|_2}.$$

*Proof.* Since  $A$  is Hermitian there exists, in  $C_n$ , an orthonormal basis of eigenvectors  $\{u_i\}$  such that

$$Au_i = \lambda_i u_i, \quad i = 1, 2, \dots, n.$$

Therefore, we can express  $x$  as a linear combination

$$(14) \quad x = \sum_{i=1}^n a_i u_i,$$

with  $a_i = u_i^* x$ . From (13) and (14),

$$\eta = \sum_{i=1}^n a_i (\lambda_i - \lambda) u_i.$$

Hence

$$\eta^* \eta = \sum_{i=1}^n |a_i|^2 (\lambda_i - \lambda)^2.$$

Therefore

$$(15) \quad \frac{\eta^* \eta}{x^* x} = \sum_{i=1}^n b_i (\lambda_i - \lambda)^2,$$

with

$$b_i = \frac{|a_i|^2}{\sum_{j=1}^n |a_j|^2}.$$

Note that

$$b_i \geq 0, \quad \sum_{i=1}^n b_i = 1,$$

whence

$$\min_i (\lambda_i - \lambda)^2 \leq \sum_{i=1}^n b_i (\lambda_i - \lambda)^2.$$

The conclusion

$$\min_i |\lambda - \lambda_i| \leq \frac{\|\eta\|_2}{\|x\|_2}$$

is thus established. ■

An obvious way to improve upon  $\lambda$  when given an approximate eigenvector,  $x$ , is suggested by Theorem 5, namely

**THEOREM 6.** *Given  $A$ , a Hermitian matrix, and  $\mathbf{x}$ , the residual vector  $\boldsymbol{\eta}$  defined by (13) is minimal for*

$$(16) \quad \lambda = \lambda_R \equiv \frac{\mathbf{x}^* A \mathbf{x}}{\mathbf{x}^* \mathbf{x}}.$$

*Proof.* The quadratic expression in  $\lambda$

$$\|\boldsymbol{\eta}\|_2^2 = (\mathbf{x}^* A - \lambda \mathbf{x}^*)(A \mathbf{x} - \lambda \mathbf{x}) = \mathbf{x}^* A^2 \mathbf{x} - 2\lambda \mathbf{x}^* A \mathbf{x} + \lambda^2 \mathbf{x}^* \mathbf{x}$$

assumes its minimum at  $\lambda = \lambda_R$ . That is,

$$\min_{\lambda} \boldsymbol{\eta}^* \boldsymbol{\eta} = \mathbf{x}^* A^2 \mathbf{x} - \lambda_R^2 \mathbf{x}^* \mathbf{x}. \quad \blacksquare$$

The quantity  $\lambda_R$  defined by (16) is called the *Rayleigh quotient* and will be referred to later on in the discussion of iterative methods.

For the Hermitian matrix  $A$ , with eigenvalues  $\{\lambda_i\}$  and corresponding eigenvectors  $\{\mathbf{u}_i\}$ , let  $U_r$  denote the linear space spanned by the eigenvectors  $\mathbf{u}_i$ ,  $i = 1, 2, \dots, r$ . If we know something about the spacing of the eigenvalues, we may be able to estimate the error,  $\|\mathbf{x} - U_r\|_2$ , of an approximate eigenvector  $\mathbf{x}$ . (We use here the notation for *distance between a vector  $\mathbf{x}$  and a set  $S$* :

$$\|\mathbf{x} - S\| = \text{g.l.b.}_{\mathbf{y} \in S} \|\mathbf{x} - \mathbf{y}\|.)$$

That is,

**THEOREM 7.** *If  $|\lambda_i - \lambda| \leq \|\boldsymbol{\eta}\|_2$  for  $i = 1, 2, \dots, r$ , with  $\lambda$ ,  $\boldsymbol{\eta}$ , and  $\mathbf{x}$  that satisfy (13), (14);  $|\lambda_i - \lambda| \geq d > 0$  for  $i = r + 1, r + 2, \dots, n$ , then*

$$(17) \quad \|\mathbf{x} - U_r\|_2 \leq \|\mathbf{x} - \sum_{i=1}^r a_i \mathbf{u}_i\|_2 \leq \frac{\|\boldsymbol{\eta}\|_2}{d}.$$

*Proof.* From the above definition and (14),

$$(18) \quad \|\mathbf{x} - U_r\|_2^2 \leq \|\mathbf{x} - \sum_{i=1}^r a_i \mathbf{u}_i\|_2^2 = \sum_{i=r+1}^n |a_i|^2.$$

Now by (13) and (14)

$$\|\boldsymbol{\eta}\|_2^2 = \sum_{i=1}^n |a_i|^2 (\lambda_i - \lambda)^2 \geq \sum_{i=r+1}^n |a_i|^2 (\lambda_i - \lambda)^2.$$

Hence by the hypothesis on the spacing of eigenvalues,

$$d^2 \sum_{i=r+1}^n |a_i|^2 \leq \|\boldsymbol{\eta}\|_2^2,$$

or

$$\sum_{i=r+1}^n |a_i|^2 \leq \frac{\|\boldsymbol{\eta}\|_2^2}{d^2}.$$

Now, from (18), the estimate (17) follows. ■

Theorems 5 and 7 give us a posteriori estimates for the error in an approximate eigenvalue and eigenvector of a Hermitian matrix. These estimates do not require detailed information about all of the eigenvalues and eigenvectors of the matrix. Unfortunately, for a non-Hermitian matrix, the situation is more complicated and we state

**THEOREM 8 (FRANKLIN).** *Let  $A$  be of order  $n$ , and have a set of  $n$  linearly independent eigenvectors  $\{\mathbf{u}_i\}$ , eigenvalues  $\{\lambda_i\}$ , which satisfy  $AU = U\Lambda$ , with  $U \equiv (\mathbf{u}_1, \mathbf{u}_2, \dots, \mathbf{u}_n)$ ,  $\Lambda \equiv (\lambda_i \delta_{ij})$ .*

If for some  $\epsilon > 0$ ,

$$(19) \quad \|Ax - \lambda x\|_2 \leq \epsilon \|Ax\|_2,$$

then

$$(20) \quad \min_{\lambda_j \neq 0} \left| 1 - \frac{\lambda}{\lambda_j} \right| \leq \epsilon \|U\|_2 \|U^{-1}\|_2.$$

*Proof.* Define  $\mathbf{b} \equiv U^{-1}\mathbf{x}$ , so that

$$\mathbf{x} = U\mathbf{b},$$

(21)

$$Ax - \lambda x = U(\Lambda - \lambda I)\mathbf{b}.$$

Now  $\mathbf{y} = U^{-1}(U\mathbf{y})$ , implies  $\|\mathbf{y}\| \leq \|U^{-1}\| \|U\mathbf{y}\|$ . Therefore

$$(22) \quad \|U\mathbf{y}\| \geq \|U^{-1}\|^{-1} \|\mathbf{y}\|.$$

With  $\mathbf{y} \equiv (\Lambda - \lambda I)\mathbf{b}$ , (22) and (21) imply

$$(23) \quad \|Ax - \lambda x\| \geq \|U^{-1}\|^{-1} \|(\Lambda - \lambda I)\mathbf{b}\|.$$

But from (19) and (23)

$$\epsilon \|U\Lambda\mathbf{b}\|_2 = \epsilon \|Ax\|_2 \geq \|Ax - \lambda x\|_2 \geq \|U^{-1}\|^{-1} \|(\Lambda - \lambda I)\mathbf{b}\|_2.$$

Hence

$$\epsilon \|U\|_2 \|\Lambda\mathbf{b}\|_2 \geq \|U^{-1}\|^{-1} \|(\Lambda - \lambda I)\mathbf{b}\|_2.$$

Therefore, since  $Ax \neq \mathbf{0}$  implies  $\Lambda\mathbf{b} \neq \mathbf{0}$ ,

$$(24) \quad \frac{\|(\Lambda - \lambda I)\mathbf{b}\|_2}{\|\Lambda\mathbf{b}\|_2} \leq \epsilon \|U^{-1}\|_2 \|U\|_2.$$

Now, let  $\mathbf{b}$  have components  $(b_i)$ ,  $i = 1, 2, \dots, n$ . Clearly with the norm  $\|\cdot\|_2$

$$\begin{aligned}\|(\Lambda - \lambda I)\mathbf{b}\|_2^2 &= \sum_{i=1}^n |\lambda_i - \lambda|^2 |b_i|^2 \geq \sum_{\lambda_i \neq 0} |\lambda_i - \lambda|^2 |b_i|^2; \\ \|\Lambda\mathbf{b}\|_2^2 &= \sum_{j=1}^n |\lambda_j|^2 |b_j|^2 = \sum_{\lambda_j \neq 0} |\lambda_j|^2 |b_j|^2.\end{aligned}$$

Therefore,

$$\frac{\|(\Lambda - \lambda I)\mathbf{b}\|_2^2}{\|\Lambda\mathbf{b}\|_2^2} \geq \sum_{\lambda_i \neq 0} \left|1 - \frac{\lambda}{\lambda_i}\right|^2 \frac{|\lambda_i b_i|^2}{\alpha^2}$$

where  $\alpha^2 \equiv \sum_{\lambda_j \neq 0} |\lambda_j|^2 |b_j|^2$ . Hence

$$(25) \quad \frac{\|(\Lambda - \lambda I)\mathbf{b}\|_2^2}{\|\Lambda\mathbf{b}\|_2^2} \geq \min_{\lambda_i \neq 0} \left|1 - \frac{\lambda}{\lambda_i}\right|^2$$

since  $|\lambda_i b_i|^2 / \alpha^2 \geq 0$  and  $\sum_{\lambda_i \neq 0} |\lambda_i b_i|^2 / \alpha^2 = 1$ .

By combining the inequalities (24) and (25), the estimate (20) results. ■

We have also the

**THEOREM 9.** *The hypothesis of Theorem 8, with (19) replaced by*

$$(19') \quad \|A\mathbf{x} - \lambda\mathbf{x}\|_2 \leq \epsilon \|\mathbf{x}\|_2,$$

*implies*

$$(20') \quad \min_i |\lambda - \lambda_i| \leq \epsilon \|U^{-1}\|_2 \|U\|_2.$$

*Proof.* Left as Problem 8. ■

Unfortunately, Theorems 7 and 8 require information about the matrix of eigenvectors, which is not generally available in problems where we are only interested in obtaining a few of the eigenvalues and eigenvectors. In the special case of a Hermitian matrix,  $A$ , the matrix of eigenvectors,  $U$ , is unitary (see Problem 9), whence  $\|U\|_2 = \|U^{-1}\|_2 = 1$ . In this case, the estimate (20') of Theorem 9 reduces to the estimate given in Theorem 5. If, on the other hand, for the case that the eigenvectors  $\{\mathbf{u}_i\}$  of  $A$  form a basis of  $C_n$ , we have a set of vectors  $\{\mathbf{v}_i\}$  that approximate the eigenvectors, then let  $P \equiv (\mathbf{v}_1, \mathbf{v}_2, \dots, \mathbf{v}_n)$ .

Clearly, if we define  $R$  and  $\epsilon$  by

$$P^{-1}AP \equiv \Lambda + \epsilon R$$

where  $\Lambda \equiv (\lambda_i \delta_{ij})$ ,  $\|R\| = 1$ , then  $\epsilon$  is small when  $P$  is a good approximation of  $U$ . The Gershgorin circles may now be small enough to give close estimates for the eigenvalues.

## PROBLEMS, SECTION 1

1. If  $\mathbf{y}$  is a left eigenvector and  $\mathbf{x}$  is a right eigenvector of  $A$ , corresponding to distinct eigenvalues, show that  $\mathbf{y}^* \mathbf{x} = 0$ .

2. (a) Show that the left and right eigenvectors corresponding to  $\lambda = 1$ , are orthogonal for

$$A \equiv \begin{pmatrix} 1 & 1 \\ 0 & 1 \end{pmatrix}.$$

(b) Find the eigenvalues  $\lambda(\epsilon)$  for  $A(\epsilon) \equiv A + \epsilon C$ ,

$$C \equiv \begin{pmatrix} 0 & 0 \\ 1 & 0 \end{pmatrix},$$

and verify that

$$|\lambda(\epsilon) - \lambda| = O(\sqrt{\epsilon}).$$

(c) Find the eigenvectors  $\mathbf{x}(\epsilon)$  of  $A(\epsilon)$  and verify by substitution that (9) holds and that both the eigenvectors  $\mathbf{x}(\epsilon)$  converge to the eigenvector of  $A$ .

3. Let  $B$  and  $B(\epsilon)$  be of order  $n$ , have rank  $n - 1$ , and

$$B(\epsilon) \rightarrow B \quad \text{as} \quad \epsilon \rightarrow 0.$$

Show that for  $|\epsilon|$  sufficiently small there exists a vector  $\mathbf{x}(\epsilon)$  in the null space of  $B(\epsilon)$ , i.e.,

$$B(\epsilon)\mathbf{x}(\epsilon) = \mathbf{0},$$

such that

$$\|\mathbf{x}(\epsilon)\|_\infty \geq 1$$

and

$$\mathbf{x}(\epsilon) \rightarrow \mathbf{x}(0) \quad \text{as} \quad \epsilon \rightarrow 0,$$

where

$$B\mathbf{x}(0) = \mathbf{0}.$$

[Hint: Since  $B(\epsilon) \rightarrow B$ , we see that for all sufficiently small  $|\epsilon|$ , the  $(n - 1)$ st order square submatrices  $B_{ij}$  of  $B$  and  $B_{ij}(\epsilon)$  of  $B(\epsilon)$  found by deleting the  $i$ th row and  $j$ th column from each of  $B$  and  $B(\epsilon)$  for some pair of indices  $(i, j)$ , are non-singular and  $B_{ij}(\epsilon) \rightarrow B_{ij}$ . Hence, the Gaussian elimination method may be used to triangularize  $B_{ij}$ . The same pivotal elements may be used to triangularize  $B_{ij}(\epsilon)$  for  $|\epsilon|$  sufficiently small. Set  $x_j(\epsilon) = 1$  and solve  $B(\epsilon)\mathbf{x}(\epsilon) = \mathbf{0}$  by using the above triangularization.]

4. Let  $B$  and  $B(\epsilon)$  be of order  $n$ ,  $B$  have rank  $n - r$ ,  $B(\epsilon)$  have rank  $< n$ , and

$$B(\epsilon) \rightarrow B \quad \text{as} \quad \epsilon \rightarrow 0.$$

Let the null space of  $B$  be  $S \equiv \{\mathbf{x} \mid B\mathbf{x} = \mathbf{0}\}$ .

Show that for  $|\epsilon|$  sufficiently small, there exists a vector  $\mathbf{x}(\epsilon)$  in the null space of  $B(\epsilon)$ , such that

$$\|\mathbf{x}(\epsilon)\|_\infty \geq 1,$$

$$\min_{\mathbf{x} \in S} \|\mathbf{x}(\epsilon) - \mathbf{x}\| \rightarrow 0 \quad \text{as} \quad \epsilon \rightarrow 0.$$

5. Carry out the proof of (3') (see corollary to Theorem 2).

[Hint: Let  $\mathbf{x}(\epsilon) \neq \mathbf{0}$  be any eigenvector satisfying

$$(A + \epsilon C)\mathbf{x}(\epsilon) = \lambda(\epsilon)\mathbf{x}(\epsilon).$$

With the matrices  $P$  and  $\Lambda$  used in the proof of Theorem 1,

$$[\lambda(\epsilon)I - \Lambda]P^{-1}\mathbf{x}(\epsilon) = \epsilon P^{-1}CP[P^{-1}\mathbf{x}(\epsilon)].$$

If  $\lambda(\epsilon)I - \Lambda$  is singular, (3') holds. If  $\lambda(\epsilon)I - \Lambda$  is not singular,

$$P^{-1}\mathbf{x}(\epsilon) = \epsilon[\lambda(\epsilon)I - \Lambda]^{-1}P^{-1}CP[P^{-1}\mathbf{x}(\epsilon)].$$

Take any norm  $\|\cdot\|_p$  of both sides to get (3').]

**6.** Apply Schwarz' inequality (i.e.,  $|\mathbf{y}^*\mathbf{x}| \leq \|\mathbf{y}\|_2\|\mathbf{x}\|_2$ ) to show that

$$|\mathbf{y}^*C\mathbf{x}| \leq \|C\|_2\|\mathbf{y}\|_2\|\mathbf{x}\|_2.$$

**7.** Show that

$$\frac{\|C\|}{\|P^{-1}\| \|P\|} \leq \|P^{-1}CP\| \leq \|C\| \|P^{-1}\| \|P\|.$$

[Hint:  $C = P(P^{-1}CP)P^{-1}$ .]

**8.** Prove Theorem 9.

[Hint: Estimate

$$\frac{\|(\Lambda - \lambda I)\mathbf{b}\|_2^2}{\sum \|\mathbf{b}\|_2^2} = \frac{\sum_{i=1}^n |\lambda_i - \lambda|^2 |b_i|^2}{\sum_{j=1}^n |b_j|^2} \geq \min_i |\lambda_i - \lambda|^2.$$

**9.** If  $U$  is unitary, show that  $\|U\|_2 = \|U^{-1}\|_2 = 1$ .

**10.** If  $P_m(\lambda) \equiv c_0\lambda^m + c_1\lambda^{m-1} + \dots + c_m$ , and  $A$  has the eigenvalues  $\{\lambda_i\}$  and the eigenvectors  $\{\mathbf{u}_i\}$ , then  $P_m(A)$  has the eigenvalues  $\{P_m(\lambda_i)\}$  with the same eigenvectors.

If  $A$  is Hermitian, show that the Rayleigh quotient, defined by (16) for any  $\mathbf{x} \neq \mathbf{0}$ , satisfies

$$\lambda_1 \leq \lambda_R \leq \lambda_n,$$

where  $\lambda_1$  and  $\lambda_n$  are the smallest and largest eigenvalues of  $A$ .

**11.** Prove

**THEOREM 3.** Let  $\mathbf{x}$  be a right eigenvector and  $\mathbf{y}$  a left eigenvector of  $A$  for the eigenvalue  $\lambda$ . For the similarity transformation given by any non-singular  $P$ , set

$$B \equiv P^{-1}AP, \quad B(\epsilon) \equiv P^{-1}A(\epsilon)P.$$

Then  $\mathbf{y}^*C\mathbf{x}/\mathbf{y}^*\mathbf{x}$  is invariant under  $P$ .

[Hint:  $\mathbf{u} \equiv P^{-1}\mathbf{x}$  is the right eigenvector of  $B$  corresponding to  $\mathbf{x}$ ;  $\mathbf{v} \equiv P^*\mathbf{y}$  is the left eigenvector of  $B$  corresponding to  $\mathbf{y}$ ;  $\epsilon P^{-1}CP$  is the perturbation corresponding to  $\epsilon C$ . Hence, with  $D \equiv P^{-1}CP$ ,

$$\frac{\mathbf{v}^*Du}{\mathbf{v}^*\mathbf{u}} = \frac{\mathbf{v}^*P^{-1}CP\mathbf{u}}{\mathbf{v}^*\mathbf{u}} = \frac{\mathbf{y}^*C\mathbf{x}}{\mathbf{y}^*\mathbf{x}}. \quad \blacksquare$$

**12.** Given  $\mathbf{x}$  and  $\mathbf{y}$  with  $\|\mathbf{x}\|_2 = \|\mathbf{y}\|_2 = 1$ , construct a unitary matrix  $C$  such that  $C\mathbf{x} = \mathbf{y}$ .

**13.** Construct the matrix  $A$  which has the eigenvalues  $\lambda_j$  and the corresponding eigenvectors  $\mathbf{x}_j$  where  $x_{1k} = 1$  for  $1 \leq k \leq n$ ,  $x_{kk} = \delta$  for

$2 \leq k \leq n$ , and  $x_{ij} = 0$  otherwise. Show that the left unit eigenvectors  $\mathbf{y}_j$  are determined by the biorthogonality property. Hence verify that

$$|y_{11}| = \mathcal{O}\left(\frac{|\delta|}{\sqrt{n-1}}\right), \quad \text{and} \quad |\mathbf{y}^* \mathbf{x}|^{-1} = \mathcal{O}\left(\frac{\sqrt{n-1}}{|\delta|}\right).$$

(Therefore, for small  $\delta$ , the eigenvalue problem for finding  $\lambda_1$  is not well-posed. But the eigenvalue problem for  $B \equiv P^{-1}AP$  is exceedingly well-posed if  $P \equiv (\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n)$ .)

[Hint: Start with the diagonal matrix  $B \equiv (\lambda_i \delta_{ij})$  and  $P$ . Construct  $P^{-1}$  and then  $A = PBP^{-1}$ . Sketch a picture of  $\mathbf{x}_j$  and  $\mathbf{y}_j$  in the three dimensional case.]

## 2. THE POWER METHOD

The power method, in its basic form, is conceptually the simplest iterative procedure for approximating the largest or *principal eigenvalue* and eigenvector of a matrix. Let us assume, throughout this subsection, that the  $n$ th order matrix  $A$  has real elements  $(a_{ij})$ ,  $n$  linearly independent eigenvectors  $\{\mathbf{u}_j\}$ ,  $j = 1, 2, \dots, n$ , and a unique eigenvalue of maximum magnitude, i.e., the eigenvalues satisfy

$$|\lambda_1| > |\lambda_2| \geq |\lambda_3| \geq \dots \geq |\lambda_n|.$$

Since the  $\lambda_j$  are the roots of a characteristic polynomial with real coefficients, the complex eigenvalues occur in complex conjugate pairs. Hence  $\lambda_1$  is real. Since  $\mathbf{u}_1$  satisfies  $A\mathbf{u}_1 = \lambda_1\mathbf{u}_1$ , the components  $(u_{i1})$  of  $\mathbf{u}_1$  may be taken to be real.

Let  $\mathbf{x}_0$  be an arbitrarily chosen real  $n$ -dimensional vector and form the sequence of vectors

$$(1) \quad \mathbf{x}_v = A\mathbf{x}_{v-1} = A^v\mathbf{x}_0; \quad v = 1, 2, \dots$$

Since  $A$  has a complete set of eigenvectors  $\{\mathbf{u}_j\}$ , say with components  $(u_{ij})$ , there exist  $n$  scalars  $a_j$  such that

$$(2) \quad \mathbf{x}_0 = \sum_{j=1}^n a_j \mathbf{u}_j.$$

Then the sequence (1) has the representation

$$(3) \quad \begin{aligned} \mathbf{x}_v &= \sum_{j=1}^n \lambda_j^v a_j \mathbf{u}_j \\ &= \lambda_1^v \left[ a_1 \mathbf{u}_1 + \sum_{j=2}^n \left( \frac{\lambda_j}{\lambda_1} \right)^v a_j \mathbf{u}_j \right]; \quad v = 1, 2, \dots \end{aligned}$$

Now clearly, since  $|\lambda_j/\lambda_1| < 1$  for all  $j \geq 2$ , the directions of the vectors  $\mathbf{x}_v$  converge to that of  $\mathbf{u}_1$  as  $v \rightarrow \infty$ , provided only that  $a_1 \neq 0$ . Of course,

$\lambda_1^\nu$ , in general, either converges to zero or becomes unbounded, and so the sequence (1) may not be practical for computations. However, a simple scaling of the iterates  $\mathbf{x}_\nu$ , to be introduced later, will remedy this defect.

Since  $\mathbf{x}_\nu$ , for large  $\nu$ , may be a close approximation to the eigenvector belonging to  $\lambda_1$ , we can employ the methods indicated in the introduction to approximate the eigenvalue. Thus let us form the ratio of, say, the  $k$ th components of  $\mathbf{x}_\nu$  and  $A\mathbf{x}_\nu = \mathbf{x}_{\nu+1}$ . Call the ratio  $\sigma_{\nu+1}$  and get from (3),

$$(4) \quad \sigma_{\nu+1} = \frac{x_{k,\nu+1}}{x_{k\nu}} = \lambda_1 \frac{a_1 u_{k1} + \sum_{j=2}^n \left(\frac{\lambda_j}{\lambda_1}\right)^{\nu+1} a_j u_{kj}}{a_1 u_{k1} + \sum_{j=2}^n \left(\frac{\lambda_j}{\lambda_1}\right)^\nu a_j u_{kj}}.$$

If  $a_1 \neq 0$  and  $k$  is chosen such that  $u_{k1} \neq 0$ , then for  $\nu$  so large that  $|\lambda_2/\lambda_1|^\nu \ll 1$ , (4) yields

$$(5) \quad \sigma_{\nu+1} = \lambda_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^\nu\right).$$

Thus, in approximating the eigenvalues, the growth or decay of  $\lambda_1^\nu$  causes no theoretical difficulties. The convergence of  $\sigma_\nu$  to  $\lambda_1$  is seen to be at least of first order with ratio at most  $|\lambda_2/\lambda_1|$ . As in our previous studies of iteration schemes (e.g., Chapter 2, Section 4; Chapter 3, Section 1), we may define the rate of convergence as

$$(6) \quad R = \ln \left| \frac{\lambda_1}{\lambda_2} \right|.$$

Difficulties in convergence may be expected if the first two eigenvalues (in magnitude) are “close.”

Another way to approximate  $\lambda_1$  is by means of the Rayleigh quotient indicated in (0.3). Thus we define

$$(7) \quad \sigma'_{\nu+1} = \frac{\mathbf{x}_\nu^* A \mathbf{x}_\nu}{\mathbf{x}_\nu^* \mathbf{x}_\nu} = \frac{\mathbf{x}_\nu^* \mathbf{x}_{\nu+1}}{\mathbf{x}_\nu^* \mathbf{x}_\nu}.$$

Then from (3) and (7) we have

$$\sigma'_{\nu+1} = \lambda_1 \frac{\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\bar{\lambda}_i}{\lambda_1}\right)^\nu \left(\frac{\lambda_j}{\lambda_1}\right)^{\nu+1} \bar{a}_i a_j \mathbf{u}_i^* \mathbf{u}_j}{\sum_{i=1}^n \sum_{j=1}^n \left(\frac{\bar{\lambda}_i}{\lambda_1}\right)^\nu \left(\frac{\lambda_j}{\lambda_1}\right)^\nu \bar{a}_i a_j \mathbf{u}_i^* \mathbf{u}_j}.$$

A calculation reveals that  $\sigma'_\nu$  converges to  $\lambda_1$  just as does  $\sigma_\nu$  [i.e., as in equation (5)]. However, if the vectors  $\mathbf{u}_i$  are mutually orthogonal or say

for convenience *orthonormal*, i.e.,  $\mathbf{u}_i^* \mathbf{u}_j = \delta_{ij}$ , then by Problem 1,  $A$  is symmetric and has real eigenvalues and eigenvectors, so that

$$(8) \quad \sigma'_{v+1} = \lambda_1 \frac{|a_1|^2 + \sum_{j=2}^n \left(\frac{\lambda_j}{\lambda_1}\right)^{2v+1} |a_j|^2}{|a_1|^2 + \sum_{j=2}^n \left|\frac{\lambda_j}{\lambda_1}\right|^{2v} |a_j|^2}.$$

Hence,

$$(9) \quad \sigma'_{v+1} = \lambda_1 + \mathcal{O}\left(\left|\frac{\lambda_2}{\lambda_1}\right|^{2v}\right),$$

when  $A$  is symmetric, and the rate of convergence in this case is twice that of the scheme used in (4). Of course, this gain in using (7) is only achieved when the matrix  $A$  is symmetric. An interesting motivation for using the approximation (7) in general is furnished in Problem 2; this result should be compared with that of Theorem 1.6.

In order to terminate the iterative computations (1) and (4) or (7), a variety of different tests can be suggested. For instance, if the quotients in (4) agree for several values of  $k$  (i.e., ratios of several components) then a fairly good approximation to  $\lambda_1$  has been obtained. Usually the obvious test of little or no change in the eigenvalue iterates  $\sigma_v$  or  $\sigma'_v$  for several successive values of  $v$  may be successfully employed. However, a quantitative test, based on Theorems 1.8 or 1.9, requires little additional computing and yields very precise information when  $A$  is symmetric. That is, pick an arbitrary  $\epsilon > 0$  and iterate until

$$(10) \quad \|A\mathbf{x}_v - \sigma\mathbf{x}_v\|_2 \leq \epsilon \|A\mathbf{x}_v\|_2 \equiv \epsilon \|\mathbf{x}_{v+1}\|_2.$$

Here  $\sigma = \sigma_{v+1}$  or  $\sigma'_{v+1}$ . From Theorem 1.8, we then find

$$(11) \quad \min_j \left|1 - \frac{\sigma}{\lambda_j}\right| \leq \epsilon \|U\|_2 \cdot \|U^{-1}\|_2.$$

For sufficiently large  $v$  we can be assured that the minimum in (11) is attained for  $j = 1$  [assuming as usual that  $a_1 \neq 0$  in (2)].

Thus, a bound on the relative error in approximating  $\lambda_1$  is attained. However, the quantities  $\|U\|_2$  and  $\|U^{-1}\|_2$  cannot be estimated in the general case. But if  $A$  is symmetric, the matrix  $U$  is unitary and

$$\|U\|_2 = \|U^{-1}\|_2 = 1.$$

Thus for the symmetric case the condition (10) implies the precise bound

$$(12) \quad \left|1 - \frac{\sigma}{\lambda_1}\right| \leq \epsilon.$$

The present iterative scheme usually requires that the iterates  $\mathbf{x}_v$  be scaled at each step of the process. Thus in place of the sequence (1), we actually calculate, with arbitrary  $\mathbf{y}_0$ ,

$$(13) \quad \xi_v = A\mathbf{y}_{v-1}, \quad \mathbf{y}_v = \frac{1}{s_v} \xi_v; \quad v = 1, 2, \dots$$

The sequence of scale factors,  $s_v$ , can be chosen in a variety of ways. The two most commonly used choices are

$$(14a) \quad s_v = \max_j |\xi_{j,v}| = \|\xi_v\|_\infty,$$

and

$$(14b) \quad s'_v = \|\xi_v\|_2.$$

The choice (14b) requires the taking of a square root at each step, but the Rayleigh quotient estimate (7) of the eigenvalue becomes, since  $\|\mathbf{y}_v\|_2 = 1$ ,

$$\begin{aligned} (\sigma'_{v+1})^2 &= \frac{\mathbf{y}_v^* A \mathbf{y}_v}{\mathbf{y}_v^* \mathbf{y}_v} \\ &= \mathbf{y}_v^* \xi_{v+1}. \end{aligned}$$

The ratio estimate (4) is now computed as

$$\sigma_{v+1} = \frac{\xi_{k,v+1}}{y_{kv}}.$$

The convergence test (10) now takes the form

$$\|\xi_{v+1} - \sigma \mathbf{y}_v\|_2 \leq \epsilon \|\xi_{v+1}\|_2.$$

If the normalization (14b) is employed the computations for this test can be simplified to

$$(s'_{v+1})^2 + \sigma^2 - 2\sigma \xi_{v+1}^* \mathbf{y}_v \leq \epsilon^2 (s'_{v+1})^2.$$

In any event, the convergence rates of (5) or (9) still apply in the appropriate cases as do the error estimates (11) or (12).

The power method as presented here is frequently adequate for approximating a simple principal eigenvalue and eigenvector. In the event that the principal eigenvalues are  $\lambda_1 = \bar{\lambda}_2$ , i.e., complex conjugate, but simple, then the numbers  $\lambda_1, \lambda_2$  will be approximated by the roots of a quadratic equation found by examining three successive iterates  $\mathbf{x}_n, \mathbf{x}_{n+1}$ , and  $\mathbf{x}_{n+2}$  (see Problem 6). Similarly, if the principal eigenvalue has a known multiplicity, the scheme for approximating  $\lambda_1$  can be suitably modified (see Problem 7).

Of course, if the matrix  $A$  has no zero components, the operational count for each iteration (1) is  $n^2$ , in general. We now turn to modifications of the power method which improve its rate of convergence.

## 2.1. Acceleration of the Power Method

Convergence to the principal eigenvalue by the power method has been shown to be geometric with convergence factor  $|\lambda_2/\lambda_1|$  (or by using the Rayleigh quotient for a symmetric matrix, the eigenvalue convergence factor is  $|\lambda_2/\lambda_1|^2$ ). This ratio is frequently too near unity for practical computations. Thus we seek modifications, analogous to those employed in Chapter 2, Section 5, and Chapter 3, Subsection 3.3, to accelerate the convergence.

Assume  $A$  to be *symmetric* with *unique principal eigenvalue*  $\lambda_1$  and consider the power method for the modified real symmetric matrix

$$(15) \quad B \equiv A + \alpha I.$$

The eigenvalues  $\mu_i$  of  $B$  are

$$\mu_i = \lambda_i + \alpha, \quad i = 1, 2, \dots, n.$$

If  $\mu_1$  is the unique principal eigenvalue of  $B$ , the rate of convergence is now determined by

$$\max_{j \neq 1} \left| \frac{\lambda_j + \alpha}{\lambda_1 + \alpha} \right|.$$

We minimize this ratio with respect to  $\alpha$  and find, if the  $\lambda_i$  are now ordered by

$$\lambda_1 > \lambda_2 \geq \dots \geq \lambda_n \quad (\text{or } \lambda_1 < \lambda_2 \leq \dots \leq \lambda_n),$$

that the *optimal value* of  $\alpha$  is

$$(16) \quad \alpha = -\frac{\lambda_2 + \lambda_n}{2}.$$

The proof is a simple modification of that of Theorem 5.1 in Chapter 2.

In order to apply this improvement, estimates of  $\lambda_2$  and  $\lambda_n$  are required. Such estimates may require auxiliary computations which is one of the disadvantages of the proposed acceleration procedure. For example, if a crude estimate  $\sigma$  of  $\lambda_1$  is known (obtained perhaps by the ordinary power method) then the principal eigenvalue of  $C \equiv A - \sigma I$  will be  $\lambda_n - \sigma$ . Thus the power method applied to  $C$  will yield an estimate of  $\lambda_n$ . Good estimates of  $\lambda_2$  are more difficult to compute. However, if  $\mathbf{u}$  is an approximation to  $\mathbf{u}_1$ , then using any  $\mathbf{x}_0$  such that  $\mathbf{x}_0^* \mathbf{u} = 0$  as the initial vector in (1) for a few iterations may yield a reasonable estimate of  $\lambda_2$  (assuming, of course, that  $|\lambda_2| > |\lambda_n|$ ). The reason is that, if  $\mathbf{x}_0$  is almost orthogonal to  $\mathbf{u}_1$ , then  $a_1$  in (3) will be quite small. Whence, for appropriately “small” values of  $\nu$ , we may have  $|\lambda_1^\nu a_1| \ll |\lambda_2^\nu a_2|$  and the ratio  $x_{k,\nu+1}/x_{k,\nu}$  will be a better approximation to  $\lambda_2$  than to  $\lambda_1$ .

On the other hand, if we are given  $\alpha$  and  $\beta$  such that, for example,

$$(17) \quad \alpha \leq \lambda_n \leq \cdots \leq \lambda_2 \leq \beta < \lambda_1,$$

then more efficient acceleration schemes for finding  $\lambda_1$  can be devised. Let  $P_m(\lambda)$  be a polynomial of degree  $m$ , say

$$(18) \quad P_m(\lambda) \equiv c_0\lambda^m + c_1\lambda^{m-1} + \cdots + c_m.$$

By  $P_m(A)$  we indicate the matrix which is the corresponding polynomial in the matrix  $A$ . In Problem 10, it is shown that the eigenvalues of  $P_m(A)$  are  $\{P_m(\lambda_i)\}$ . We now consider the power method applied to the matrix  $P_m(A)$  with the initial vector  $\mathbf{x}_0$  of (2). In place of (3), we now get after  $\tau$  iterations

$$(19) \quad \begin{aligned} \mathbf{x}_\tau &= P_m(A)\mathbf{x}_{\tau-1} \\ &= P_m^\tau(A)\mathbf{x}_0, \\ \mathbf{x}_\tau &= P_m^\tau(\lambda_1) \left[ a_1\mathbf{u}_1 + \sum_{j=2}^n \left( \frac{P_m(\lambda_j)}{P_m(\lambda_1)} \right)^\tau a_j\mathbf{u}_j \right]. \end{aligned}$$

To evaluate  $\mathbf{x}_\tau$ , we do not actually form the matrix  $P_m(A)$  but recursively compute the vectors  $A\mathbf{x}_{\tau-1}, \dots, A^m\mathbf{x}_{\tau-1}$ . Thus the number of operations performed in one iteration of (19) is equivalent to that for  $m$  iterations of (1). We then can compare the convergence rate by examining

$$\max_{j \neq 1} \left| \frac{\lambda_j}{\lambda_1} \right|^m \quad \text{and} \quad \max_{j \neq 1} \left| \frac{P_m(\lambda_j)}{P_m(\lambda_1)} \right|.$$

In fact, the “best” polynomial (18) of degree  $\leq m$  to employ is that for which  $\max_z |P_m(z)/P_m(\lambda_1)|$  is a minimum on  $\alpha \leq z \leq \beta$ . This problem has previously been met in Chapter 2, Section 5, in a similar context. The determination of this best polynomial is described in Chapter 5, Section 4, where we study the Chebyshev polynomials.

When the iterate  $\mathbf{x}_\tau$  has been computed the approximations  $\sigma_{\tau+1}$  or  $\sigma'_{\tau+1}$  are formed as before by using  $A\mathbf{x}_\tau$  and  $\mathbf{x}_\tau$ . The convergence test (10) may still be employed.

The convergence of the eigenvalue and eigenvector iterates can also be improved by the  $\delta^2$ -process described in Chapter 3, Subsection 2.4 (see Chapter 3, Problems 2.6, 2.7), when  $A$  has a unique principal eigenvalue and a complete set of eigenvectors (see Problem 5).

## 2.2. Intermediate Eigenvalues and Eigenvectors (Orthogonalization, Deflation, Inverse Iteration)

In the previous section, we have indicated procedures whereby the power method could be modified to obtain estimates of  $\lambda_2$  and  $\lambda_n$ . The careful

application of these methods can be made to yield accurate approximations to several eigenvalues and their eigenvectors. In principle, all the values (of a real symmetric matrix) could be determined by these procedures, but in practice much accuracy may be lost in the later stages of the process.

Assume that  $A$  is symmetric, with principal eigenvalue  $\lambda_1$ , and that the ordering is

$$\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_{n-1} > \lambda_n.$$

The matrix  $A - \sigma I$  has eigenvalues  $\lambda_j - \sigma$  and, for real  $\sigma$ , has principal eigenvalue either  $\lambda_1 - \sigma$  or  $\lambda_n - \sigma$  (since the above ordering is preserved). If  $\sigma \geq \lambda_1$  then  $\lambda_n - \sigma$  is the principal eigenvalue of  $A - \sigma I$  and by the ordinary power method  $\lambda_n$  and  $\mathbf{u}_n$  can be accurately approximated. (Note here that Theorem 1.1 provides bounds for  $\{\lambda_i\}$ , whence such a  $\sigma$  may be obtained.)

However, this device cannot readily be used to yield the intermediate eigenvalues.

The *orthogonalization method* is suitable for determining intermediate eigenvalues and eigenvectors and will be described next. Once  $\mathbf{u}_1$  has been accurately determined, we may form a vector  $\mathbf{x}_0$  which is orthogonal to  $\mathbf{u}_1$ . Such a vector has the eigenvector expansion (2) in which  $a_1 = 0$ . For example, if  $\mathbf{z}$  is any vector then

$$(20) \quad \mathbf{x}_0 = \mathbf{z} - \frac{\mathbf{u}_1^* \mathbf{z}}{\mathbf{u}_1^* \mathbf{u}_1} \mathbf{u}_1$$

has the property,  $\mathbf{x}_0^* \mathbf{u}_1 = 0$ . Now the sequence  $\mathbf{x}_v$  is formed and used to compute  $\lambda_2$  (assuming  $|\lambda_2| > |\lambda_j|$ ,  $j = 3, 4, \dots, n$ ) and  $\mathbf{u}_2$ . However, after several iterations, roundoff errors will usually introduce a small but non-zero component of  $\mathbf{u}_1$  in the  $\mathbf{x}_v$  and subsequent iterations will magnify this component. This *contamination by roundoff* may be reduced by removing the  $\mathbf{u}_1$  component periodically; i.e., say after every  $r$  steps  $\mathbf{x}_0$  is recomputed by using  $\mathbf{x}_v$  in place of  $\mathbf{z}$  in (20). When  $\lambda_2$  and  $\mathbf{u}_2$  have been determined in this manner the procedure can in principle be continued to determine  $\lambda_3$  and  $\mathbf{u}_3$ .

In general, if  $\lambda_i$  and  $\mathbf{u}_i$  are known for  $i = 1, 2, \dots, s$  then we form, for arbitrary  $\mathbf{z}$ ,

$$(21) \quad \mathbf{x}_0 = \mathbf{z} - \sum_{i=1}^s \frac{\mathbf{u}_i^* \mathbf{z}}{\mathbf{u}_i^* \mathbf{u}_i} \mathbf{u}_i.$$

Since the  $\mathbf{u}_i$  are orthogonal, with  $\mathbf{z} = \sum_{j=1}^n a_j \mathbf{u}_j$ , we find that

$$\mathbf{x}_0 = a_{s+1} \mathbf{u}_{s+1} + \cdots + a_n \mathbf{u}_n,$$

and the power method applied to this  $\mathbf{x}_0$  will yield  $\lambda_{s+1}$  and  $\mathbf{u}_{s+1}$ . The roundoff contamination is more pronounced for larger values of  $s$  and hence (21) must be frequently reapplied with  $\mathbf{z}$  replaced by the current iterate,  $\mathbf{x}_v$ .

If the matrix  $A$  were not symmetric, but did have a complete set of eigenvectors, then the corresponding biorthogonalization process could be carried out. For example, the unique principal left and right eigenvectors  $\mathbf{v}_1$  and  $\mathbf{u}_1$  could be found:  $\mathbf{u}_1$  by the iteration scheme (1); and  $\mathbf{v}_1$  by the scheme, with  $\mathbf{y}_0$  arbitrary,

$$(1^*) \quad \mathbf{y}_v^* = \mathbf{y}_{v-1}^* A, \quad v = 1, 2, \dots$$

The unique next largest eigenvalue  $\lambda_2$  could be found by selecting, for  $\mathbf{z}$  arbitrary,

$$(22a) \quad \mathbf{x}_0 = \mathbf{z} - \frac{\mathbf{v}_1^* \mathbf{z}}{\mathbf{v}_1^* \mathbf{v}_1} \mathbf{v}_1,$$

or

$$(22b) \quad \mathbf{y}_0 = \mathbf{z} - \frac{\mathbf{u}_1^* \mathbf{z}}{\mathbf{u}_1^* \mathbf{u}_1} \mathbf{u}_1$$

and then applying the power methods (1) or (1<sup>\*</sup>) respectively. The vectors  $\{\mathbf{x}_n\}$  and  $\{\mathbf{y}_n\}$  are orthogonal to  $\mathbf{v}_1$  and  $\mathbf{u}_1$  respectively (see Problem 8). Of course, in practice, the effect of rounding errors would have to be removed by periodically re-biorthogonalizing the vectors  $\{\mathbf{x}_n\}$  and  $\{\mathbf{y}_n\}$ . To Problem 9, we leave the development of the analog of (21) for the determination of  $\lambda_{s+1}$ ,  $\mathbf{u}_{s+1}$ , and  $\mathbf{v}_{s+1}$ .

A method for *roughly* approximating  $\lambda_2$  and  $\mathbf{u}_2$  for the symmetric matrix  $A$  is motivated as follows. With (1), (2), and (3), we define

$$(23) \quad \mathbf{z}_v \equiv \mathbf{x}_v - \lambda_1 v \mathbf{a}_1 \mathbf{u}_1 = \sum_{j=2}^n \lambda_j v \mathbf{a}_j \mathbf{u}_j, \quad v = 1, 2, \dots$$

By taking ratios of, say, the  $k$ th components of  $\mathbf{z}_v$  and  $\mathbf{z}_{v+1}$  we have, assuming  $|\lambda_2| > |\lambda_3| \geq \dots \geq |\lambda_n|$ ,

$$\sigma_{v+1} \equiv \frac{z_{k, v+1}}{z_{kv}} = \lambda_2 + \mathcal{O}\left(\left|\frac{\lambda_3}{\lambda_2}\right|^v\right).$$

Similarly, by forming the Rayleigh quotient we have

$$\sigma'_{v+1} = \frac{\mathbf{z}_v^* \mathbf{z}_{v+1}}{\mathbf{z}_v^* \mathbf{z}_v} = \lambda_2 + \mathcal{O}\left(\left|\frac{\lambda_3}{\lambda_2}\right|^{2v}\right).$$

Also as  $v \rightarrow \infty$  the direction of  $\mathbf{z}_v$  converges to that of  $\mathbf{u}_2$ .

Unfortunately, the process defined by (23) with

$$\lambda_1^v a_1 = \frac{\mathbf{u}_1^* \mathbf{x}_v}{\mathbf{u}_1^* \mathbf{u}_1},$$

and

$$\mathbf{u}_1 = \lim_{n \rightarrow \infty} \mathbf{x}_n,$$

has the feature that many significant figures are lost in (23) as  $v$  gets large. But these values  $\sigma_{v+1}$  or  $\sigma'_{v+1}$  are good approximations of  $\lambda_2$  only for large  $v$ . Hence we must choose  $v$  judiciously in order to get a reasonable estimate of  $\lambda_2$  and  $\mathbf{u}_2$ . This procedure can be readily adapted for the sequence of normalized iterates  $\{y_v\}$  defined in (13). It should be noted that the vectors  $\{\mathbf{x}_v\}$  (or  $\{\mathbf{y}_v\}$ ) which are required have already been computed in determining  $\lambda_1$  and  $\mathbf{u}_1$ . Thus, with little extra computation, we have found a rough approximation for  $\lambda_2$  and  $\mathbf{u}_2$ .

These two procedures removed components of the known eigenvectors from the iterated vectors. However, it is also possible to alter the real symmetric matrix  $A$  so that the known eigenvectors then correspond to zero eigenvalues. Iteration on an arbitrary vector with this altered matrix then automatically eliminates the known components. Thus, suppose  $\mathbf{u}_1$  and  $\lambda_1$  are known and that  $\mathbf{u}_1$  is normalized by  $\|\mathbf{u}_1\|_2^2 = \mathbf{u}_1^* \mathbf{u}_1 = 1$ . Then we may form the matrix  $A_1$  as follows

$$(24) \quad A_1 = A - \lambda_1 \mathbf{u}_1 \mathbf{u}_1^*.$$

Since  $A$  is symmetric, so is  $A_1$ . We also note that  $A_1 \mathbf{u}_1 = \mathbf{0}$ . For any other eigenvector,  $\mathbf{u}_j$ , belonging to an eigenvalue,  $\lambda_j$ ,  $j > 1$ , it follows from the orthogonality of the eigenvectors that  $A_1 \mathbf{u}_j = \lambda_j \mathbf{u}_j$ . Thus  $A_1$  has all the eigenvectors of  $A$  and all its eigenvalues except  $\lambda_1$  which is replaced by zero. A simple calculation and proof by induction reveals that (see Problem 10)

$$(25) \quad \begin{aligned} \mathbf{z}_v &\equiv A_1^v \mathbf{z} = A^v \mathbf{z} - \lambda_1^v \mathbf{u}_1 \mathbf{u}_1^* \mathbf{z} \\ &= A^v \mathbf{z} - \lambda_1^v (\mathbf{u}_1^* \mathbf{z}) \mathbf{u}_1. \end{aligned}$$

A comparison of this result and the sequence  $\{\mathbf{x}_v\}$  with  $\mathbf{x}_0$  given by (20) shows that, *for exact computation* with normalized eigenvectors, *the present method is exactly equivalent to the orthogonalization method*. The cancellation errors of (23) do not occur now but instead an error grows due to the fact that the  $\lambda_1$  and  $\mathbf{u}_1$  employed are not an exact eigenvalue and eigenvector respectively. Thus the computed  $A_1$  does not satisfy (25) exactly. However, iterations with  $A_1$  are usually more accurate than the more economical computations in (23).

If the matrix  $A$  were *sparse* (many zero elements), then we would not recommend using (24) since it would, in general, produce a full matrix  $A_1$ . When  $\lambda_2$  and  $\mathbf{u}_2$  have been determined from  $A_1$ , the procedure can be repeated by forming  $A_2 = A_1 - \lambda_2 \mathbf{u}_2 \mathbf{u}_2^*$ , etc., to determine the remaining eigenvalues and eigenvectors.

In the above modifications of the matrix  $A$ , the resulting matrices  $A_k$  are still of order  $n$ . It is, in principle, possible to successively alter the matrix  $A$ , so that the resulting matrices  $B_k$  are of order  $n - k$ ,  $k = 1, 2, \dots, n - 1$ , and have the same remaining eigenvalues. Such procedures are called *deflation methods*, by analogy with the process of dividing out the roots of a polynomial as they are successively found. For example, the simplest such scheme is based on the method used in Theorem 1.1 of Chapter 1 to show that every matrix is similar to a triangular matrix. (The deflated matrices are the  $A_k$  defined there.)

We now describe another scheme, which has the additional feature that when the matrix  $A$  is Hermitian, the deflated matrices are also Hermitian.

Let  $A\mathbf{u} = \lambda\mathbf{u}$ ,  $\mathbf{u}^*\mathbf{u} = 1$ ,  $u_1 \geq 0$ ; set

$$(26) \quad P = I - 2\boldsymbol{\omega}\boldsymbol{\omega}^*,$$

where  $\boldsymbol{\omega}$  is defined, with  $\mathbf{e}_1 \equiv (\delta_{i1})$ , by the properties

$$\boldsymbol{\omega} \equiv (\omega_i),$$

$$(27) \quad \boldsymbol{\omega}^*\boldsymbol{\omega} = 1, \quad \omega_1 \geq 0,$$

$$P\mathbf{e}_1 = \mathbf{u}.$$

In Problems 11 and 12, we show that  $P$  is unitary and that the components ( $\omega_i$ ) of  $\boldsymbol{\omega}$  are defined by

$$(28) \quad \omega_1 = \left(\frac{1 - u_1}{2}\right)^{\frac{1}{2}}, \quad \omega_k = -\frac{u_k}{2\omega_1}, \quad k = 2, 3, \dots, n.$$

Now it is easy to see that

$$AP\mathbf{e}_1 = \lambda P\mathbf{e}_1,$$

whence

$$(29) \quad P^{-1}AP\mathbf{e}_1 = \lambda\mathbf{e}_1.$$

Equation (29) shows that  $\mathbf{e}_1$  is an eigenvector of

$$(30) \quad B_1 \equiv P^{-1}AP = PAP.$$

Therefore, the first column of  $B_1$  is  $\lambda\mathbf{e}_1$ . In other words,  $A$  has been deflated.<sup>†</sup> This process can be continued with the matrix  $A_1$  of order  $n - 1$

<sup>†</sup> In practice the evaluation of  $B_k$  could be performed economically by adapting the procedure described in equations (3.16<sub>k</sub>).

consisting of the last  $n - 1$  rows and columns of  $B_1$ . Let the matrix,  $Q$ , of order  $n - 1$  be of the form (26), and satisfy (27) and (28) relative to the matrix  $A_1$  with an eigenvalue  $\mu$  and eigenvector  $\mathbf{v}$  (of order  $n - 1$ ). Then set

$$P_1 \equiv \begin{pmatrix} 1 & \mathbf{o}^* \\ \mathbf{o} & Q \end{pmatrix},$$

where  $\mathbf{o}$  is of order  $n - 1$ . It is easy to verify that  $P$  is unitary, in fact,  $P_1^{-1} = P_1^* = P_1$ . Hence the matrix

$$B_2 = P_1 P A P P_1$$

has the form

$$B_2 = \begin{bmatrix} \lambda & a_2 & a_3 & \cdots & a_n \\ 0 & \mu & b_3 & \cdots & b_n \\ 0 & 0 & & & \\ \vdots & \vdots & & A_2 & \\ 0 & 0 & & & \end{bmatrix}$$

where  $A_2$  is of order  $n - 2$ . This process may be continued to provide a proof of

**THEOREM 1 (SCHUR).** *The matrix  $A$ , of order  $n$ , is unitarily similar to a triangular matrix.*

*Proof.* Left as Problem 13. ■

**COROLLARY.** *The Hermitian matrix  $A$  is unitarily similar to a diagonal matrix.* ■

Finally, we describe another iteration scheme for determining the intermediate eigenvalues and eigenvectors. This procedure is called *inverse iteration* and is based upon solving

$$(31) \quad (A - \sigma I)\mathbf{x}_n = \mathbf{x}_{n-1}, \quad n = 1, 2, \dots,$$

with  $\mathbf{x}_0$  arbitrary and  $\sigma$  a constant.

Clearly, (31) is equivalent to the power method for the matrix  $(A - \sigma I)^{-1}$ . We may use Gaussian elimination and (31) to calculate  $\mathbf{x}_n$ .

Of course, the procedure will produce the principal eigenvalue of  $(A - \sigma I)^{-1}$ , i.e.,  $1/(\lambda_k - \sigma)$ , where

$$|\lambda_k - \sigma| = \min_i |\lambda_i - \sigma|,$$

provided that  $\sigma$  is closer to the simple eigenvalue  $\lambda_k$  of  $A$  than to any other eigenvalue of  $A$ . Each iteration step, after the first triangularization

of  $A - \sigma I$ , requires about  $n^2$  operations. The first iteration step requires about  $n^3/3$  operations [see Chapter 2, equation (1.9)]. The inverse iteration method is very useful for improving upon the accuracy of an approximation to any eigenvalue.

Other iteration schemes based on matrix transformations have been devised to approximate the normal forms given in Theorem 1 and the corollary. We treat such schemes in the next section.

## PROBLEMS, SECTION 2

1. Prove that a real square matrix  $A$  of order  $n$  is symmetric iff it has  $n$  orthogonal eigenvectors.

2. Let  $A$ ,  $y$ , and  $\sigma$  be real. Show that the scalar  $\sigma$  such that  $\sigma y$  is the best mean square approximation to the vector  $Ay$  is given by

$$\sigma = \frac{1}{2} \frac{\mathbf{y}^*(A + A^*)\mathbf{y}}{\mathbf{y}^*\mathbf{y}} = \frac{\mathbf{y}^*Ay}{\mathbf{y}^*\mathbf{y}}.$$

[Hint:  $(\mathbf{y}^*A^*\mathbf{y}) = (\mathbf{y}^*Ay)^*$ ;  $\|\eta\|_2^2 = \eta^*\eta$ .]

3. Show that if  $A$  is Hermitian, then with  $A = S + iK$ , where  $S$  is real symmetric,  $K$  real skew-symmetric, the eigenvector  $\mathbf{u} = \mathbf{x} + i\mathbf{y}$  and eigenvalue  $\lambda$  satisfy

$$\begin{pmatrix} S & -K \\ K & S \end{pmatrix} \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix} = \lambda \begin{pmatrix} \mathbf{x} \\ \mathbf{y} \end{pmatrix}.$$

Verify that if  $\lambda$  is a simple eigenvalue of  $A$ , then  $\lambda$  is a double eigenvalue of the compound matrix.

4. The power method for computing left eigenvectors is based on the sequence

$$\mathbf{z}_v^* = \mathbf{z}_{v-1}^* A = \mathbf{z}_0^* A^v, \quad v = 1, 2, \dots$$

Then, with the use of the sequence (1), we may approximate  $\lambda_1$  by

$$\sigma_{v+1}'' = \frac{\mathbf{z}_v^* \mathbf{x}_{v+1}}{\mathbf{z}_v^* \mathbf{x}_v}.$$

Show that, if the matrix  $A$  has a complete set of eigenvectors and a unique principal eigenvalue, then  $\sigma_v''$  converges with the ratio  $|\lambda_2/\lambda_1|^2$ . (Note, however, that twice as many computations are required to evaluate each iteration step here.)

5. Use Problems 2.6 and 2.7 of Chapter 3 to find the improvement in eigenvalues and eigenvectors obtained by applying the  $\delta^2$ -process to  $\sigma_{v+1}$  of (4) or  $\sigma_{v+1}'$  of (7), when  $A$  is real symmetric and has a unique principal eigenvalue.

6. Show how to find the coefficients  $s$ ,  $p$  of the quadratic equation  $\lambda^2 - s\lambda + p = 0$ , that is satisfied by the unique complex conjugate pair of simple principal eigenvalues  $\lambda_1, \lambda_2$  of the real matrix  $A$ .

[Hint: Given  $\lambda_1 = \bar{\lambda}_2$ ;  $|\lambda_1| = |\lambda_2| > |\lambda_i|$ ,  $i = 3, 4, \dots, n$ ; assume that for the corresponding eigenvectors  $\mathbf{u}_1$  and  $\mathbf{u}_2 = \bar{\mathbf{u}}_1$ , the respective first components  $u_{11}$  and  $u_{12}$  are maximal. Then apply the technique of Bernoulli's method, Chapter 3, equations (4.16)–(4.18).]

7. When the maximal eigenvalue has multiplicity  $m$ , how can the power method be used? (See Chapter 3, Problem 4.5.)

8. Verify that the sequences  $\{\mathbf{x}_n\}$  and  $\{\mathbf{y}_n\}$  defined by (1), (1\*), and (22) are orthogonal to  $\mathbf{v}_1$  and  $\mathbf{u}_1$  respectively. That is, show

$$\mathbf{v}_1^* \mathbf{x}_n = \mathbf{u}_1^* \mathbf{y}_n = 0.$$

[Hint: Use induction with respect to  $n$ .]

9. Develop the biorthogonal analog of (21), for the case that  $A$  has a complete set of eigenvectors. Describe the power method for the determination of the unique intermediate eigenvalues when  $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n|$ .

[Hint: Generalize (22) and use Problem 8.]

10. Prove (25), i.e., if  $A$  is real symmetric,  $A\mathbf{u} = \lambda\mathbf{u}$ , and  $\|\mathbf{u}\|_2 = 1$ , then for all  $\mathbf{z}$  and  $v = 1, 2, \dots$ ,

$$(A - \lambda\mathbf{u}\mathbf{u}^*)^v \mathbf{z} = A^v \mathbf{z} - \lambda^v (\mathbf{u}^* \mathbf{z}) \mathbf{u}.$$

11. Prove that the Hermitian matrix

$$P = I - 2\omega\omega^*$$

is unitary, in fact  $P^{-1} = P^* = P$ , iff

$$\omega^*\omega = 1.$$

12. With

$$\mathbf{u} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \quad u_1 \geq 0, \mathbf{u}^* \mathbf{u} = 1,$$

show that if the matrix  $P$  of (26) satisfies (27), then

$$\omega = \begin{pmatrix} \omega_1 \\ \omega_2 \\ \vdots \\ \omega_n \end{pmatrix}, \quad \text{and} \quad \omega_1 = \left( \frac{1 - u_1}{2} \right)^{\frac{1}{2}}$$

$$\omega_k = -\frac{u_k}{2\omega_1}, \quad k = 2, 3, \dots, n.$$

13. (a) Give a complete proof by induction on  $n$  of Theorem 1 (Schur), along the line indicated in text.

(b) Give another proof of Theorem 1 by making use of Theorem 1.1 of Chapter 1.

[Hint: Construct  $B = P^{-1}AP$  where  $B$  is triangular. Since  $P$  is non-singular, construct a matrix  $Q$  whose columns are an orthonormal set of vectors formed successively from the columns of  $P$  so that  $P = QR$ , where  $R$  is upper triangular and non-singular. Therefore  $RBR^{-1} = Q^{-1}AQ$ . Show that the product of two upper triangular matrices and the inverse of an upper triangular matrix are upper triangular matrices.]

### 3. METHODS BASED ON MATRIX TRANSFORMATIONS

The methods of Section 2 are suitable for the determination of a few eigenvalues and eigenvectors of the matrix  $A$ . However, when we are

interested in finding all of the eigenvalues and eigenvectors then the methods based on transforming the matrix  $A$  seem more efficient. Attempts to implement Schur's theorem, by approximating a unitary matrix  $H$  which triangularizes by similarity the given general matrix  $A$ , have been recently made. We shall describe one of these later. In the case of a Hermitian matrix  $A$ , however, the classical *Jacobi's method* does work well to produce a unitarily similar diagonal matrix. The methods of *Givens* and *Householder* produce either a similar tridiagonal matrix for a Hermitian  $A$  or a similar matrix of *Hessenberg*<sup>†</sup> form for a general matrix  $A$ , with the use of a unitary similarity transformation.

We will first describe Jacobi's, Givens', and Householder's methods for treating the Hermitian matrix  $A$ . In this case, since the computational problem in practice is usually reduced to that of finding the eigenvalues of a real symmetric matrix (see Problem 2.3), we will assume that  $A$  is real symmetric. All of these methods, however, require only simple modifications to be directly applicable to the Hermitian case. Jacobi's method reduces the real symmetric matrix  $A$  by an *infinite* sequence of simple orthogonal similarity transformations (two dimensional rotations) to diagonal form. The following lemmas provide the basis of the procedure.

**LEMMA 1.** *Let  $B \equiv P^{-1}AP$ . If  $P$  is orthogonal (unitary), then*

$$(1) \quad \begin{aligned} \text{trace}(B) &= \text{trace}(A), \\ \text{trace}(B^*B) &= \text{trace}(A^*A). \end{aligned}$$

*Proof.* Since  $B$  and  $A$  are *similar* matrices, the eigenvalues of  $B$  are the same as the eigenvalues of  $A$ . By definition,

$$\text{trace}(A) = \sum_{i=1}^n a_{ii}.$$

The eigenvalues of  $A$  are the roots of

$$p_A(\lambda) \equiv \det(\lambda I - A) = 0.$$

By partially expanding the determinant, the coefficient of  $\lambda^n$  in  $p_A(\lambda)$  is seen to be unity while the coefficient of  $\lambda^{n-1}$  is  $-\text{trace}(A)$ . Hence, we find that

$$\text{trace}(A) = \sum_{i=1}^n \lambda_i = \text{trace}(B).$$

The orthogonality of  $P$ , i.e.,  $P^* = P^{-1}$ , implies

$$B^*B = (P^*A^*P)(P^*AP) = P^*A^*AP.$$

Again, because the eigenvalues are unchanged under a similarity transformation, the trace of  $A^*A$  is unchanged. ■

<sup>†</sup> See the definition of (upper) Hessenberg form in Problem 1.7 of Chapter 2.

**LEMMA 2.** *Let*

$$A_{ij} \equiv \begin{pmatrix} a_{ii} & a_{ij} \\ a_{ji} & a_{jj} \end{pmatrix}, \quad a_{ij} \neq 0,$$

*be formed from elements of a real symmetric matrix A and let*

$$(2a) \quad R \equiv \begin{pmatrix} r_{ii} & r_{ij} \\ r_{ji} & r_{jj} \end{pmatrix},$$

*with  $r_{ii} = r_{jj} = \cos \phi$ ,  $r_{ij} = -r_{ji} = \sin \phi$ , where*

$$(2b) \quad \tan 2\phi = \frac{2a_{ij}}{a_{jj} - a_{ii}}, \quad -\frac{\pi}{4} \leq \phi \leq \frac{\pi}{4}.$$

*Then*

$$R^* = R^{-1}$$

*and*

$$B_{ij} \equiv R^* A_{ij} R$$

*is a diagonal matrix.*

*Proof.* Let

$$B_{ij} \equiv \begin{pmatrix} b_{ii} & b_{ij} \\ b_{ji} & b_{jj} \end{pmatrix}.$$

Equation (2), by Problem 1, guarantees that  $b_{ij} = b_{ji} = 0$ . ■

By Lemma 1, since  $A_{ij}$  and  $B_{ij}$  are orthogonally similar,

$$(3) \quad b_{ii}^2 + b_{jj}^2 = a_{ii}^2 + 2a_{ij}^2 + a_{jj}^2.$$

We say that the matrix  $R$  *reduced to zero* the element  $a_{ij}$ . We now construct an orthogonal matrix of order  $n$ , to reduce to zero the element  $a_{ij}$  of  $A$ . Let

$$(4) \quad P \equiv (p_{st})$$

where

$$p_{ii} = r_{ii}, \quad p_{ij} = r_{ij}, \quad p_{ji} = r_{ji}, \quad p_{jj} = r_{jj},$$

and

$$p_{st} = \delta_{st} \quad \text{otherwise.}$$

With the elements  $r_{ii}$ ,  $r_{ij}$ ,  $r_{ji}$ ,  $r_{jj}$  defined in (2),  $P^*P = PP^* = I$ .

We call such a matrix  $P$  a *two dimensional rotation*. Now set,

$$B \equiv P^* AP \equiv (b_{ij}).$$

Clearly,  $b_{ij} = b_{ji} = 0$ . Hence the matrix  $P$  reduces to zero the element  $a_{ij}$  of  $A$ . We now can show that  $P$  reduces the sum of the squares of the off-diagonal elements of  $A$ . From the definition of trace ( $A^*A$ ), it is easy to verify that

$$\text{trace}(A^*A) = \sum_{i,j=1}^n a_{ij}^2.$$

Now, by Lemmas 1 and 2 and equation (3), we find, since  $b_{kk} = a_{kk}$  for  $k \neq i, j$ ,

$$(5) \quad \begin{aligned} \sum_{r \neq s} b_{rs}^2 &= \text{trace}(B^*B) - \sum_{s=1}^n b_{ss}^2 \\ &= \text{trace}(A^*A) - \sum_{s=1}^n a_{ss}^2 - 2a_{ij}^2 \\ &= \sum_{r \neq s} a_{rs}^2 - 2a_{ij}^2. \end{aligned}$$

If we pick  $(i, j)$  so that  $a_{ij}^2 = \max_{k \neq s} \{a_{ks}^2\}$ , then

$$(6) \quad a_{ij}^2 \geq \frac{\sum_{k \neq s} a_{ks}^2}{(n^2 - n)} = \frac{\text{trace}(A^*A) - \sum_{i=1}^n a_{ii}^2}{n^2 - n}.$$

In fact, (6) is satisfied by any  $a_{ij}^2$  that is not less than the average of the squares of the off-diagonal elements of  $A$ .

By substituting the inequality (6) in (5), we have

$$(7) \quad \sum_{r \neq s} b_{rs}^2 \leq \left(1 - \frac{2}{n^2 - n}\right) \sum_{r \neq s} a_{rs}^2.$$

Therefore, each two dimensional rotation defined in (2) and (4), and such that (6) holds, reduces the sum of the squares of the off-diagonal elements of a symmetric matrix  $A$  by a factor not greater than

$$1 - \frac{2}{n^2 - n} < 1 \quad \text{for } n \geq 2.$$

In addition, we observe from (3) and the phrase preceding (5) that

$$(8) \quad \sum_{s=1}^n b_{ss}^2 = \sum_{s=1}^n a_{ss}^2 + 2a_{ij}^2.$$

We therefore have the basis for proving

**THEOREM 1 (JACOBI).** *Let  $A$  be real symmetric, with eigenvalues  $\{\lambda_i\}$ ; let the matrices  $\{P_m\}$  be two dimensional rotation matrices of the form (4) defined successively so that an above-average element of*

$$B_0 \equiv A$$

*is reduced to zero by  $P_1$ , and thereafter  $P_m$  reduces to zero an above-average element of  $B_{m-1}$  in the similarity transformation defining  $B_m$ ,*

$$B_m \equiv P_m^* B_{m-1} P_m, \quad m = 1, 2, \dots$$

Let

$$(9) \quad Q_m \equiv P_1 P_2 \cdots P_m.$$

Then, as  $m \rightarrow \infty$ ,

$$Q_m^* A Q_m \rightarrow \Lambda,$$

where  $\Lambda \equiv (\lambda_i \delta_{ij})$ , for some ordering of the  $\{\lambda_i\}$ .

*Proof.* For  $m$  sufficiently large, by (7) and Gerschgorin's theorem

$$B_m \equiv Q_m^* A Q_m$$

is approximately the diagonal matrix  $\Lambda$ , for some ordering of  $\{\lambda_i\}$ . Now, the angle of rotation  $\phi$ , determined for  $P_{m+1}$  from  $B_m$ , is close to zero, unless the two diagonal elements of  $B_m$  used to determine  $\phi$  correspond to a pair of identical eigenvalues. Hence it is easy to verify that

$$Q_m^* A Q_m \rightarrow \Lambda. \quad \blacksquare$$

*Jacobi's method* is the scheme in which  $P_{k+1}$  is determined so as to reduce a *maximal* off-diagonal element of  $B_k$  to zero,  $k = 0, 1, 2, \dots$ . In practice, by listing the magnitude and column index of the maximal off-diagonal element in each row of  $B_k$ , the Jacobi scheme is easily carried out. That is, the only elements that change in going from  $B_k$  to  $B_{k+1}$  are the elements in the rows and columns of index  $i$  or  $j$ . Hence, the list of maximal elements in each row of  $B_{k+1}$  is easily constructed from the list for  $B_k$  by making at most  $2n - 1$  comparisons. A common variation of the Jacobi method consists in examining the off-diagonal elements of  $B_m$  in a systematic *cyclic order* given by the indices  $(1, 2), (1, 3), \dots, (1, n), (2, 3), (2, 4), \dots, (2, n), (3, 4), \dots, (n - 1, n)$ . The indices  $(i, j)$  used to determine  $P_{m+1}$  correspond to the first element  $b_{ij}^{(m)}$  of  $B_m$  that satisfies

$$|b_{ij}^{(m)}| \geq t_m,$$

where  $\{t_m\}$  is a prescribed decreasing sequence of positive numbers called *thresholds*. Such an iteration procedure is called a *threshold scheme*. A bolder approach, namely to rotate the off-diagonal elements in sequence, irrespective of size, is called the *cyclic Jacobi scheme*. Surprisingly enough, with only a minor change in the definition of the angle  $\phi$ , when  $\phi \cong \pm \pi/4$ , the *cyclic Jacobi scheme* has been shown (Forsythe-Henrici) to be convergent also. In fact, if a comparison of the residual off-diagonal sum of squares,  $\sum_{r \neq s} (b_{rs}^{(m)})^2$ , is made after each complete cycle of  $q \equiv (n^2 - n)/2$  rotations, then it has been shown that the cyclic Jacobi scheme converges and, in fact, that it *converges quadratically for  $m$  large enough*, i.e.,

$$\sum_{r \neq s} (b_{rs}^{(m+q)})^2 \leq K \left[ \sum_{r \neq s} (b_{rs}^{(m)})^2 \right]^2$$

for  $m$  large enough.

In these Jacobi schemes, the eigenvalues of  $A$  are approximated by the diagonal elements of  $B_m$ , for sufficiently large  $m$ . Furthermore, the corresponding eigenvectors of  $B_m$  are then approximately the unit vectors  $\{\mathbf{e}_i\}$ . Hence, the eigenvectors of  $A$  are approximated by the columns of

$$Q_m = P_1 P_2 \cdots P_m.$$

Since only two columns of  $Q_m$  are changed in going from  $Q_m$  to  $Q_{m+1}$ , only  $4n$  multiplications are involved in this step. On the other hand, the elements of the matrix  $B_m$  that are unaffected by the rotation  $P_{m+1}$  are those that have indices  $(r, s)$  with  $r \neq i, j$  and  $s \neq i, j$ . Therefore, because of symmetry, approximately  $4n$  multiplications are needed to carry out the rotation of  $B_m$  into  $B_{m+1}$  (if we neglect to count the square root operations necessary to determine  $\cos \phi$  and  $\sin \phi$ ). Hence, about  $8n$  multiplications are required to determine  $B_{m+1}$  and  $Q_{m+1}$  from  $B_m$  and  $Q_m$ .

We now consider the *Givens transformation*. Here a finite sequence of

$$M \equiv \frac{(n-2)(n-1)}{2}$$

rotations are employed to reduce the real symmetric matrix  $A$  to *tri-diagonal form*.

That is, we successively construct a sequence of matrices  $\{P_k\}$ ,  $k = 1, 2, \dots, M$ , and define

$$(10) \quad \begin{aligned} B_0 &\equiv A, \\ B_k &\equiv P_k^* B_{k-1} P_k, \quad 1 \leq k \leq M. \end{aligned}$$

The matrices  $\{P_m\}$  are two dimensional rotations of the form (4) constructed so that the first  $k$  not-codiagonal elements of  $B_k$  are zero. That is, we say  $\{a_{ii}\}$  and  $\{a_{i,i+1}\}$  are the *codiagonal* elements of  $A$ . For a symmetric matrix, we refer to the not-codiagonal indices listed cyclically in the order

$$(11) \quad (1, 3), (1, 4), \dots, (1, n), (2, 4), (2, 5), \dots, (2, n), \dots, (n-2, n).$$

The first  $k$  not-codiagonal elements of  $B_k$  are the elements of  $B_k$  whose indices are among the first  $k$  in the list (11). The facts are summarized in

**THEOREM 2 (GIVENS).** *Let  $A$  be real symmetric;  $B_0 \equiv A$ . Let  $(i-1, j)$  be the  $k$ th pair of indices in the cyclic sequence (11), and  $B_{k-1}$  have elements  $(b_{rs}^{(k-1)})$ . Let  $P_k \equiv I$ , if  $b_{i-1,j}^{(k-1)} = 0$ ; otherwise, set  $P_k \equiv (p_{rs}^{(k)})$  with*

$$(12) \quad \begin{aligned} p_{ii}^{(k)} &= p_{jj}^{(k)} = \frac{b_{i-1,i}^{(k-1)}}{\sqrt{(b_{i-1,i}^{(k-1)})^2 + (b_{i-1,j}^{(k-1)})^2}}, \\ p_{ij}^{(k)} &= -p_{ji}^{(k)} = -\frac{b_{i-1,j}^{(k-1)}}{\sqrt{(b_{i-1,i}^{(k-1)})^2 + (b_{i-1,j}^{(k-1)})^2}}, \\ p_{rs}^{(k)} &= \delta_{rs} \quad \text{for other } (r, s). \end{aligned}$$

Let the matrices  $\{B_k\}$  and  $\{P_k\}$  be defined by (10) and (12) for  $k = 1, 2, \dots, M$ ,

$$M \equiv \frac{(n-2)(n-1)}{2}.$$

Then,  $B_k$  is real symmetric; the first  $k$  not-codiagonal elements of  $B_k$  are zero,  $k = 1, 2, \dots, M$ ;  $B_M$  is tridiagonal.

*Proof.* It is a simple matter to verify that if the first  $k-1$  not-codiagonal elements of  $B_{k-1}$  are zero, then the corresponding elements of  $B_k$  also vanish. [This property of preserving zeros is *not valid* in general for the rotations of the type used in Jacobi's method! In Jacobi's method the matrix  $P_k$ , a two dimensional rotation in the  $(i, j)$  coordinates, annihilated the  $(i, j)$  element; but in Givens' method the matrix  $P_k$  annihilates the  $(i-1, j)$  element]. Furthermore, by Problem 2, the definition (12) of  $P_k$  ensures that  $b_{i-1, j}^{(k)} = b_{j, i-1}^{(k)} = 0$ . Hence, by using mathematical induction the proof may be completed. ■

Aside from the calculation of the non-trivial elements of  $P_k$ , the calculation of the non-zero elements of  $B_k$  in (10) involves approximately  $4(n-i)$  multiplications. Now, in order to reduce to zero the elements in row  $i-1$ , this process must be carried out for  $j = i+1, i+2, \dots, n$ . That is,  $4(n-i)^2$  multiplications are required to put zeros in all of the not-codiagonal elements in row  $i-1$ . Therefore, for the complete reduction to tridiagonal form, we have the

**COROLLARY.** *The Givens method requires*

$$\frac{4}{3}n^3 + \mathcal{O}(n^2) \text{ operations}$$

*to transform the real symmetric matrix  $A$  to tridiagonal form.* ■

We shall complete the description of Givens' method for determining the eigenvalues and eigenvectors after we study *Householder's method* for reducing the matrix  $A$  to tridiagonal form. Householder's scheme uses a sequence of  $n-2$  orthogonal similarity transformations of the form

$$(13) \quad P = I - 2\omega\omega^*; \quad \omega^*\omega = 1,$$

with suitably chosen vectors  $\omega$ . In Problem 3,  $P^*P = PP^* = I$  is verified.

We now describe how the matrices  $P_k$  are defined. Let the rows  $i = 1, 2, \dots, k-1 \leq n-3$  of the symmetric matrix  $B_{k-1}$  have the reduced form

$$b_{rs} = 0 \quad \text{for } 1 \leq r \leq k-1 \text{ and } r+2 \leq s.$$

If

$$\sum_{t=2}^{n-k} b_{k,k+t}^2 = 0,$$

set

$$P_k \equiv I;$$

if

$$\sum_{t=2}^{n-k} b_{k,k+t}^2 \neq 0,$$

set

$$(14)_k$$

$$P_k = I - 2\omega\omega^*,$$

with

$$\omega = \beta v, \quad v = (v_i),$$

where

$$v_i = 0, \quad i = 1, 2, \dots, k,$$

$$v_{k+1} = 2y^2 S,$$

$$v_i = b_{ki}, \quad i = k+2, k+3, \dots, n;$$

$$S^2 = \sum_{t=1}^{n-k} b_{k,k+t}^2,$$

and if  $b_{k,k+1} \neq 0$ ,

$$S = \text{sign}(b_{k,k+1})\sqrt{S^2},$$

$$y = \frac{(b_{k,k+1} + S)}{2K}, \quad \beta = \frac{1}{2yS},$$

$$2K^2 = S^2 + b_{k,k+1}S.$$

We then have

**THEOREM 3 (HOUSEHOLDER).** *Let  $A$  be real symmetric. Set  $B_0 \equiv A$ , define*

$$(15)_k \quad B_k = P_k^* B_{k-1} P_k, \quad k = 1, 2, \dots, n-2,$$

*by means of  $(14)_k$ . Then  $\omega^*\omega = 1$ ;  $B_k$  is real symmetric; all of the not-codiagonal elements of  $B_k$ , in the rows  $i = 1, 2, \dots, k$ , are zero;  $B_{n-2}$  is tridiagonal.*

*Proof.* We leave the verification to Problem 4. ■

Now, we note that the practical evaluation of  $B_k$  in  $(15)_k$  can be carried out in the following fashion:

$$\begin{aligned} B_k &= P_k^* B_{k-1} P_k \\ &= (I - 2\omega\omega^*) B_{k-1} (I - 2\omega\omega^*) \\ &= B_{k-1} - 2\omega\omega^* B_{k-1} - 2B_{k-1}\omega\omega^* + 4\omega\omega^* B_{k-1}\omega\omega^*. \end{aligned}$$

Therefore,

$$(16)_k \quad B_k = B_{k-1} - 2\beta^2(\mathbf{v}\mathbf{u}^* + \mathbf{u}\mathbf{v}^*),$$

where

$$\mathbf{u} \equiv \xi - \alpha\mathbf{v},$$

with

$$\xi \equiv B_{k-1}\mathbf{v},$$

$$\alpha \equiv \beta^2\mathbf{v}^*\xi.$$

Observe that in  $(16)_k$ ,  $K$  need not be evaluated, and therefore only one square root operation is required in each stage of Householder's method. Furthermore, given  $\mathbf{v}$ , the evaluation of  $\xi$  requires

$(n - k)(n - k - 1)$  multiplications;

of  $\alpha$  requires

$n - k$  multiplications;

of  $\mathbf{u}\mathbf{v}^*$  requires

$(n - k)^2$  multiplications.

Hence we have shown that the number of multiplications required to produce  $B_k$  is approximately  $2(n - k)^2$ .

**COROLLARY.** *Householder's method reduces a real symmetric matrix to tridiagonal form with the use of  $\frac{2}{3}n^3 + \mathcal{O}(n^2)$  multiplications.*

*Proof.* The result follows from the formula

$$\sum_{k=1}^{n-2} k^2 = \frac{n^3}{3} + \mathcal{O}(n^2). \quad \blacksquare$$

We now remark that both Givens' method and Householder's method can be employed to reduce any real matrix  $A$  to *lower Hessenberg form*. The matrix  $B \equiv (b_{ij})$  is in lower Hessenberg form iff  $b_{i,s} = 0$  for  $i + 2 \leq s \leq n$ .

Finally, we give the treatment of Givens for finding the eigenvalues of the symmetric tridiagonal matrix  $B$ .

Let  $B \equiv (b_{ij})$  be real symmetric and tridiagonal, i.e.,

$$b_{ij} = b_{ji}, \quad 1 \leq i, j \leq n;$$

$$b_{ij} = 0, \quad j \neq i - 1, i, i + 1; \quad 1 \leq i \leq n.$$

Set

$$a_i \equiv b_{ii}, \quad 1 \leq i \leq n;$$

(17)

$$c_i = b_{i,i+1} = b_{i+1,i}, \quad 1 \leq i \leq n - 1.$$

Recall that

$$p_B(\lambda) \equiv \det(\lambda I - B).$$

Givens observed that the principal minor determinants of  $\lambda I - B$  formed a sequence of polynomials having properties similar to those of a Sturm sequence (see Chapter 3, Subsection 4.2). That is, let

$$(18) \quad \begin{aligned} p_0(\lambda) &\equiv 1, \\ p_1(\lambda) &\equiv (\lambda - a_1), \\ p_2(\lambda) &\equiv (\lambda - a_2)(\lambda - a_1) - c_1^2, \\ &\vdots \\ p_i(\lambda) &\equiv (\lambda - a_i)p_{i-1}(\lambda) - (c_{i-1})^2 p_{i-2}(\lambda); \quad i = 3, 4, \dots, n. \end{aligned}$$

In Problem 5 it is shown that  $p_n(\lambda) \equiv p_B(\lambda)$ . If any  $c_i = 0$ , then the determination of the eigenvalues of  $B$  is reduced to the eigenvalue problem for a smaller matrix. Hence we assume  $c_i \neq 0$ ,  $1 \leq i \leq n - 1$ . We have then,

**THEOREM 4 (GIVENS).** *Let the tridiagonal, real symmetric matrix  $B$  be defined by (17), with all  $c_i \neq 0$ . Then the zeros of each  $p_i(\lambda)$ ,  $i = 2, 3, \dots, n$ , are distinct and are separated by the zeros of  $p_{i-1}(\lambda)$ ; and, if  $p_n(\gamma) \neq 0$ , the number of eigenvalues of  $B$  that are greater than  $\gamma$  is equal to the number of sign variations in the sequence  $p_n(\gamma), p_{n-1}(\gamma), \dots, p_1(\gamma), 1$ .*

*Proof.* Since  $c_i \neq 0$ , no two successive polynomials  $p_i(\lambda)$  and  $p_{i-1}(\lambda)$  can have a common zero. Otherwise, from (18),  $p_{i-2}(\lambda), \dots, p_0(\lambda)$  would also have that zero. By mathematical induction, we can now prove the separation property. That is, a simple plot of  $p_2(\lambda)$  shows that the simple zero of  $p_1(\lambda)$  separates the two simple zeros of  $p_2(\lambda)$ . Assume that the  $i - 2$  simple zeros of  $p_{i-2}(\lambda)$  separate the  $i - 1$  simple zeros of  $p_{i-1}(\lambda)$ . Now, from (18), at each zero of  $p_{i-1}(\lambda)$ , the sign of  $p_i(\lambda)$  is opposite to the sign of  $p_{i-2}(\lambda)$ . But, by the induction hypothesis,  $p_{i-2}(\lambda)$  changes sign between each pair of neighboring zeros of  $p_{i-1}$ . Therefore,  $p_i(\lambda)$  also changes sign and hence has a zero *between* each neighboring pair of zeros of  $p_{i-1}(\lambda)$ . Now

$$p_i(+\infty) = +\infty, \quad p_i(-\infty) = (-1)^i \infty, \quad i = 1, 2, \dots$$

Therefore  $p_i(\lambda)$  has a zero to the right of the largest zero of  $p_{i-1}(\lambda)$  and a zero to the left of the smallest zero of  $p_{i-1}(\lambda)$ . On the other hand,  $p_i(\lambda)$  can have no more than  $i$  zeros. Therefore, we have shown that the  $i - 1$  simple zeros of  $p_{i-1}(\lambda)$  separate the  $i$  simple zeros of  $p_i(\lambda)$ . This *separation property* is all that is needed to verify the rest of the theorem's conclusion, by the kind of argument we used for treating Sturm sequences in Chapter 3, Subsection 4.2 (see Problem 6). ■

The evaluation of the characteristic polynomial of  $B$ , once we have calculated  $\{c_i^2\}$ , is then carried out by using (18). Thus a sequence of  $2n - 3$  multiplications is required for each determination of  $p_n(\lambda)$ . If all of the

eigenvalues of  $B$  are in the interval  $[a, b]$ , then we may locate the eigenvalues more precisely by halving the interval and using the theorem to find out how many zeros of  $p_B(\lambda)$  are in each half, i.e., pick  $\gamma = (a + b)/2$ . In this way, after  $t$  halvings, we will know the location of an eigenvalue to within  $\pm (a + b)2^{-(t+1)}$ , at the expense of about  $2tn$  multiplications.

Once an eigenvalue  $\lambda$  of  $B$  is found, a corresponding eigenvector can be determined by using the fact that

$$B - \lambda I$$

has rank  $\leq n - 1$ .

If none of the off-diagonal terms  $c_i$  vanish,<sup>†</sup> then the equations

$$(B - \lambda I)\mathbf{x} = \mathbf{0}$$

may be solved simply by applying Gaussian elimination. That is, with the use of maximal column pivots, we proceed to eliminate  $x_1, x_2, \dots, x_n$ . We neglect the last  $r$  equations of the essentially triangular system that results, if the rank is  $n - r$ . We give arbitrary non-zero values to the variables  $x_{n-r+1}, \dots, x_n$ , that appear in the last  $r$  equations; and solve the first  $n - r$  equations of the triangular system. If the maximal column pivots do not occur in order along the diagonal, we list the pivotal equations in the order that we find them. In this case, the then resulting upper triangular matrix,  $U$ , may not be codiagonal. That is, the only non-trivial elements in row  $i$  may be the elements  $u_{ii}$ ,  $u_{i,i+1}$ , and  $u_{i,i+2}$ .

In Problem 8 we see that  $\max_{i,j} |u_{ij}| \leq 5b$  where  $b = \max_{i,j} |b_{ij}|$ . Hence by the theory of a priori estimates in Subsection 1.2, Chapter 2, though the matrix  $B - \lambda I$  is singular, the first  $n - r$  equations of  $U$ , when computed with finite precision, give an accurate representation of the coefficients of the unknowns  $x_1, x_2, \dots, x_{n-r}$  that would arise by exact elimination. Furthermore, the precise determination of  $r$  is not necessary!

If  $\mathbf{x}$  is an eigenvector of  $B$  corresponding to the eigenvalue  $\lambda$  then

$$\mathbf{y} = P\mathbf{x},$$

is the corresponding eigenvector of  $A$ , since

$$B = P^{-1}AP.$$

For determining all of the eigenvalues and all of the eigenvectors of  $B$ , experience has shown that the  $QR$  method (see end of section) for the symmetric, tridiagonal matrix  $B$  is a more efficient procedure than Givens' method.

<sup>†</sup> If any  $c_i = 0$ , the system splits into two disjoint systems.

We now turn our attention to the case of the general real matrix  $A$ . If we are interested in determining all of the eigenvalues of  $A$ , then a preliminary simplification to a similar Hessenberg form  $B$  is appropriate, since the iterative operations on  $B$  will then require fewer calculations. For example, as remarked after the corollary to Theorem 3, Givens' or Householder's orthogonal similarity transformations might be used to effect the reduction.

Review the discussion, between Theorem 3 and its corollary, of the number of operations involved in using Householder's method. Since the matrix  $B_{k-1}$  is not symmetric when  $A$  is not symmetric, we see that the operational count for the vector  $\xi$  becomes  $(n - k)n + \mathcal{O}(n)$  while the other counts remain the same. Hence, the number of multiplications required to reduce the general real matrix  $A$  to Hessenberg form is  $\frac{5}{3}n^3 + \mathcal{O}(n^2)$  since  $(16)_k$  is not valid.

A convenient and practical technique to evaluate  $p_B(\lambda)$ , when  $B \equiv (b_{ij})$  is in lower Hessenberg form, uses

**LEMMA 3 (HYMAN).** Let  $B$  be in lower Hessenberg form. If  $b_{i,i+1} \neq 0$ ,  $i = 1, 2, \dots, n - 1$ , define the sequence of polynomials  $m_i(\lambda)$

$$(19) \quad \begin{aligned} m_0 &\equiv 1 \\ -b_{i,i+1}m_i &\equiv b_{i1}m_0 + b_{i2}m_1 + \cdots + b_{i,i-1}m_{i-2} + (b_{ii} - \lambda)m_{i-1}, \\ i &= 1, 2, \dots, n - 1. \end{aligned}$$

Then

$$(20) \quad \begin{aligned} g_B(\lambda) &\equiv \det(B - \lambda I) = (-1)^{n+1}b_{12}b_{23}\cdots b_{n-1,n}g(\lambda), \\ g(\lambda) &\equiv b_{n1}m_0 + b_{n2}m_1 + \cdots + b_{n,n-1}m_{n-2} + (b_{nn} - \lambda)m_{n-1}. \end{aligned}$$

*Proof.* Since  $b_{i,i+1} \neq 0$ , we can successively add multiples of the columns of  $B - \lambda I$  to the first column in order to annihilate the first  $n - 1$  elements of the first column. This process defines the polynomials  $\{m_i\}$ ,  $i = 1, 2, \dots, n - 1$ , and does not change the value of the determinant. But the  $(n, 1)$  element is  $g(\lambda)$  and the expansion of the determinant with respect to the elements of the first column results in (20). ■

Clearly, if any  $b_{i,i+1} = 0$ , the  $\det(B - \lambda I)$  can be written as the product of two determinants of submatrices of  $B - \lambda I$ .

In the case  $b_{i,i+1} \neq 0$ , formulas (19) and (20) may be used to calculate  $g(\lambda)$ , with the use of  $n^2/2 + \mathcal{O}(n)$  multiplications. Similarly, by differentiating the formulas (19) with respect to  $\lambda$ , a recursive evaluation of  $g'(\lambda)$  (or higher derivatives) is also simply carried out. Hence we may apply any of the standard iterative methods for finding the roots of the poly-

nomial  $g(\lambda)$  (without evaluating its coefficients). A matter of considerable practical importance for the evaluation of the polynomial  $g(\lambda)$  is the double precision accumulation of the inner products in (19) and (20).

Another family of methods for finding the eigenvalues of the lower Hessenberg matrix  $A$  are called *factorization methods*. First, we note that  $C \equiv A^T$  and  $A$  have the same eigenvalues, but that  $C$  is of upper Hessenberg form. The first factorization method we describe is *Rutishauser's LR method*. That is, the *LR* method consists in constructing (when possible) the factorization of the matrix  $C_1 \equiv C$  in the form

$$(21) \quad C_1 = L_1 R_1,$$

where  $L$  is lower triangular (with unit diagonal elements) and  $R$  is upper triangular. Then Rutishauser considers

$$(22) \quad L_1^{-1} C_1 L_1 = R_1 L_1 \equiv C_2.$$

Next find

$$(23) \quad C_2 = L_2 R_2,$$

again a lower unit and upper triangular factorization.

Now

$$L_2^{-1} C_2 L_2 = L_2^{-1} L_1^{-1} C_1 L_1 L_2 = R_2 L_2 \equiv C_3.$$

In general, a sequence of *similar* matrices  $\{C_k\}$  is constructed and their  $L_k R_k$  factorization via Gaussian elimination is also found, so that

$$(24) \quad \begin{aligned} C_k &= L_k R_k, \\ C_{k+1} &= L_k^{-1} C_k L_k \\ &= L_k^{-1} \cdots L_1^{-1} C_1 L_1 \cdots L_k. \end{aligned}$$

But if we define

$$P_k = L_1 \cdots L_k,$$

$$Q_k = R_k \cdots R_1.$$

then

$$P_k C_{k+1} = C_1 P_k,$$

and therefore

$$(25) \quad \begin{aligned} P_k Q_k &= P_{k-1} C_k Q_{k-1} = C_1 P_{k-1} Q_{k-1} \\ &= C_1^2 P_{k-2} Q_{k-2} \\ &\vdots \\ &= C_1^k. \end{aligned}$$

Hence,  $P_k Q_k$  is the *LR* factorization of  $C_1^k$ . The fact that the matrix  $C_k$  converges to an upper triangular form is shown under special assumptions (which may be weakened considerably).

**THEOREM 5 (RUTISHAUSER).** *Let  $C_1 X = X \Lambda$ , where  $\Lambda \equiv (\lambda_i \delta_{ij})$ . Assume  $|\lambda_1| > |\lambda_2| > \dots > |\lambda_n| > 0$ . Set*

$$Y \equiv X^{-1}.$$

*Assume  $X$  and  $Y$  can be factored in the form*

$$X = L_X R_X, \quad Y = L_Y R_Y,$$

*where  $L_X$  and  $L_Y$  are lower unit triangular and  $R_X$  and  $R_Y$  are upper triangular. Then with  $\{C_k\}$  defined by (21), we have the result:  $C_k$  converges to upper triangular form.*

*Proof (Wilkinson).* Clearly,

$$\begin{aligned} C_1^k &= X \Lambda^k Y = X \Lambda^k L_Y R_Y \\ &= X (\Lambda^k L_Y \Lambda^{-k}) (\Lambda^k R_Y). \end{aligned}$$

But by the strict inequalities satisfied by  $\{\lambda_i\}$ , the lower triangular matrix  $E_k$ , defined by

$$\Lambda^k L_Y \Lambda^{-k} - I \equiv E_k \equiv (e_{ij}^{(k)})$$

satisfies

$$E_k \rightarrow O; \quad e_{ii}^{(k)} = 0, \quad i = 1, 2, \dots, n.$$

Therefore

$$\begin{aligned} C_1^k &= L_X R_X (I + E_k) \Lambda^k R_Y \\ &= L_X (I + R_X E_k R_X^{-1}) R_X \Lambda^k R_Y. \end{aligned}$$

But  $R_X E_k R_X^{-1} \rightarrow O$ . Therefore, the *LR* factors of  $I + R_X E_k R_X^{-1}$  both converge to  $I$ .

Hence, since  $R_X \Lambda^k R_Y$  is upper triangular, we see that the lower triangular factor of  $C_1^k$ , which by (25) is  $P_k$ , converges to  $L_X$ . That is,

$$P_k \rightarrow L_X.$$

Hence  $L_k = P_{k-1}^{-1} P_k$  converges to  $I$ .

But then, since

$$L_k^{-1} C_k = R_k,$$

is upper triangular, it follows that  $C_k$  must converge to upper triangular form. ■

It is easy to verify that the *LR* method preserves the Hessenberg form of the matrices  $\{C_k\}$ . The *LR* method can be made to converge much more rapidly by introducing a shift

$$D_k = C_k - s_k I,$$

and then continuing with the factorization of  $D_k$ .

But we shall not pursue this avenue further. We merely remark that the *LR* method does not always work as well as we have described, because the *LR* factorization, even if possible, may give rise to ever increasing magnitudes of numbers.

Another factorization method which seems not to suffer from the above defect is the *QR* factorization of *Francis* and *Kublanovskaja*. That is, the upper Hessenberg matrix  $C_1$  is written

$$C_1 = Q_1 R_1,$$

whence

$$Q_1^{-1} C_1 Q_1 = R_1 Q_1 \equiv C_2,$$

where  $R_1$  is upper triangular and  $Q_1$  is unitary, i.e.,

$$Q_1^* = Q_1^{-1}.$$

We then factor

$$C_2 = Q_2 R_2.$$

In general, with

$$C_k = Q_k R_k,$$

we set

$$Q_k^{-1} C_k Q_k \equiv C_{k+1} = R_k Q_k.$$

All the matrices  $Q_k$  are unitary and all the matrices  $R_k$  are upper triangular. Again, the Hessenberg form of  $\{C_i\}$  is preserved.

*Francis* and *Kublanovskaja* have given proofs of convergence of  $C_k$  to upper triangular form, in special cases. *Wilkinson* has given a simpler proof using techniques similar to those used in proving Theorem 5.

An important feature of *Francis'* work is that he shows how to use real arithmetic and maintain the real Hessenberg form, even when some eigenvalues are complex and the accelerating shifts indicated above (for the *LR* method) are complex numbers. Since he works with real numbers, when the eigenvalues are distinct but some are complex conjugate, in pairs

$$\lambda_i = \bar{\lambda}_{i+1},$$

then the matrices  $C_k$  converge to a form that is not triangular (i.e., the limiting form has second order matrix blocks on the diagonal).†

The *QR* factorization is unique when  $C_1$  is non-singular. The Gram-Schmidt orthogonalization process is not recommended for carrying out the *QR* factorization. Rather, left-multiplications may be performed upon the matrix  $C_1$ , by unitary matrices of the form  $I - 2\omega\omega^*$ , in order to successively reduce the columns of  $C_1$ . That is,

$$(I - 2\omega_{n-1}\omega_{n-1}^*) \cdots (I - 2\omega_1\omega_1^*) C_1 = R_1.$$

† These real matrices  $C_k$  are real orthogonally similar and therefore can only be expected to converge to the *Murnaghan-Wintner* canonical form rather than the Schur form of Theorem 2.1.

The matrix  $Q_1$  is given by the product of the unitary matrices,

$$Q_1 = (I - 2\omega_1\omega_1^*)(I - 2\omega_2\omega_2^*) \cdots (I - 2\omega_{n-1}\omega_{n-1}^*).$$

When  $C_1$  is in Hessenberg form, these unitary matrices are simple two dimensional rotations.

### PROBLEMS, SECTION 3

1. With  $A_{ij}$ ,  $R$ , and  $B_{ij}$  as defined in Lemma 2, show that  $b_{ij} = b_{ji} = 0$ .
2. Give the details of the proof of Theorem 2.

[Hint: The only elements of  $B_{k-1}$  that are transformed by  $P_k$  are in the rows and columns with indices  $i$  or  $j$ . It is necessary only to examine the components (where  $j > i \geq 2$ )

$$b_{is}^{(k)}, \quad \text{if } s \leq i - 2;$$

and

$$b_{sj}^{(k)}, \quad \text{if } 1 \leq s \leq i - 1.]$$

3. Verify that for any matrix  $P$  of the form (13),  $P^* = P$  and

$$P^*P = PP^* = I.$$

4. Carry out the verification of Theorem 3.

[Hint: Rows  $i = 1, 2, \dots, k - 1$  are unaffected by  $(15)_k$ . Check that row  $k$  is properly reduced.]

5. Verify that  $p_n(\lambda)$  given by (18) satisfies  $p_n(\lambda) \equiv p_B(\lambda)$ .

6. Complete the proof of Theorem 4. That is, use the hypothesis and assume the root separation property to prove the recipe for counting eigenvalues.

7. What recurrence relation is satisfied by the functions  $\{p_i'(\lambda)\}$ , where  $\{p_i(\lambda)\}$  is defined in (18)?

8. Let  $B$  be tridiagonal, of form (17);  $c_i \neq 0$ ,  $i = 1, 2, \dots, n - 1$ ; and  $\lambda$  be an eigenvalue of  $B$ . Use Gaussian elimination with maximal column pivots to solve  $(B - \lambda I)\mathbf{x} = \mathbf{0}$ . Let  $U \equiv (u_{ij})$  be the matrix of the resulting equivalent upper triangular system. If  $\max_{i,j} (|a_i|, |c_j|) = b$ , then  $\max_{i,j} |u_{ij}| \leq 5b$  (Wilkinson).

[Hint: By Gershgorin's theorem  $|\lambda| \leq 3b$ . Use induction and examine the two equations involved in eliminating  $x_i$ .]

9. Let  $A$  be Hermitian of order  $n$  and  $\mu$  be an approximation to the simple eigenvalue  $\lambda$ . Let the eigenvector  $\mathbf{x}$  correspond to  $\lambda$ . For simplicity, suppose  $\max_i |x_i| = x_n = 1$ . The usual way to approximate  $\mathbf{x}$  consists in solving the  $n - 1$  equations

$$(26) \quad (\hat{A} - \mu \hat{I})\hat{\mathbf{x}} = -\hat{\mathbf{c}},$$

where  $\hat{A}$ ,  $\hat{I}$ ,  $\hat{\mathbf{x}}$ ,  $\hat{\mathbf{c}}$  consist in deleting the last row and column of  $A$  and  $I$  and deleting the last component both of  $\mathbf{x}$  and of the last column  $\mathbf{c}$  of  $A$ , respectively. Then define the residual

$$(27) \quad f(\mu) = \hat{\mathbf{c}}^* \hat{\mathbf{x}} + a_{nn} - \mu.$$

(a) From (26), since  $\hat{\mathbf{x}}$  is a function of  $\mu$ , verify by differentiation with respect to  $\mu$  that

$$(\hat{A} - \mu\hat{I}) \frac{d\hat{\mathbf{x}}}{d\mu} - \hat{\mathbf{x}} = \hat{\mathbf{0}}.$$

If  $\mu$  is close enough to  $\lambda$ , and  $\hat{A} - \lambda\hat{I}$  is non-singular, then

$$\frac{d\hat{\mathbf{x}}}{d\mu} = (\hat{A} - \mu\hat{I})^{-1}\hat{\mathbf{x}}.$$

From (27),

$$\frac{df}{d\mu} = \hat{\mathbf{c}}^*(\hat{A} - \mu\hat{I})^{-1}\hat{\mathbf{x}} - 1.$$

(b) Verify, then, that in Newton's method for improving  $\mu$

$$\Delta\mu = -\frac{f(\mu)}{f'(\mu)} = \frac{\mathbf{x}^*\boldsymbol{\eta}}{\mathbf{x}^*\mathbf{x}},$$

where  $\eta_i = 0$  for  $1 \leq i \leq n-1$ ;  $\eta_n = f(\mu)$ .

10. Define the *circulant matrix*  $A \equiv (a_{ij})$ , of order  $n$ , generated from  $(a_1, a_2, \dots, a_n)$  by

$$a_{ij} \equiv \begin{cases} a_{j-i+1}, & i \leq j, \\ a_{n+j-i+1}, & i > j. \end{cases}$$

Show that the eigenvectors  $\{\mathbf{u}_j\}$  and eigenvalues  $\{\lambda_j\}$  are given in terms of the  $n$  roots of unity  $\{\omega_j\}$  [i.e.,  $(\omega_j)^n = 1$ ] by

$$\mathbf{u}_j \equiv (1, \omega_j, \omega_j^2, \dots, \omega_j^{n-1}),$$

$$\lambda_j \equiv a_1 + a_2\omega_j + a_3\omega_j^2 + \dots + a_n\omega_j^{n-1}.$$

11. For the circulant matrix generated by  $a_i = 1$ ,  $1 \leq i \leq 6$ , use Householder's method and Jacobi's method to find the eigenvalues and eigenvectors.

12. Show that Householder's method reduces a real skew-symmetric matrix to a real tridiagonal skew-symmetric matrix. Carry out this procedure for the circulant matrix generated by  $(a_1, a_2, a_3, a_4, a_5, a_6) \equiv (0, 1, 1, 0, -1, -1)$ .

13. If  $C_1$  is of Hessenberg form, then show that each stage of the *LR* transformation requires  $n^2 + \mathcal{O}(n)$  operations, while each stage of the *QR* transformation requires  $4n^2 + \mathcal{O}(n)$  operations.

# 5

## Basic Theory of Polynomial Approximation

### 0. INTRODUCTION

There are numerous reasons for seeking approximations to functions. The type of approximation sought depends upon the application intended as well as the ease or difficulty with which it can be obtained. In any event, the “simplest”† approximating functions would seem to be polynomials, and we devote much of our attention to them in this chapter. Some consideration is also given to approximation by trigonometric functions. We shall study the approximation of continuous (possibly differentiable) functions, in a closed bounded interval.

In general, a polynomial, say  $P_n(x)$  of degree at most  $n$ , may be said to be an approximation to a function,  $f(x)$ , in an interval  $a \leq x \leq b$  if some measure of the deviation of the polynomial from the function in this interval is “small.” This notion of approximation becomes precise only when the measure of deviation and magnitude of smallness have been specified.

To this end, we recapitulate the definition of *norm*, this time for a linear space (not necessarily finite dimensional) whose elements are functions  $\{f(x)\}$  (see Chapter 1, Section 1). The norm, written

$$\text{Norm}(f) \equiv N(f) \equiv \|f\|,$$

is an assignment of a real number to each element of the linear space such that:

† We do not study the theory of approximation by *rational functions* (i.e., quotient of polynomials), even though rational functions are easy to evaluate and, in certain cases, are more efficient than polynomials.

- (i)  $\|f\| \geq 0$ ,
- (ii)  $\|f\| = 0$  iff  $f(x) \equiv 0$ ,
- (iii)  $\|cf\| = |c| \cdot \|f\|$  for any constant  $c$ ,
- (iv)  $\|f + g\| \leq \|f\| + \|g\|$ , the *triangle inequality*.

The notion of linear spaces of functions is of basic importance in the analysis of many approximation procedures. In particular, for present applications we wish to examine not the polynomial approximation of a particular function but, in fact, the properties of such approximations for any function in an appropriate linear space.

A *measure of the deviation* or *error* in the approximation of  $f(x)$  by  $P_n(x)$  will be denoted generically by

$$\|f(x) - P_n(x)\|$$

(sometimes with an appropriate subscript or superscript). This measure of deviation will be required to satisfy the properties (i), (iii), and (iv) of a *norm*, but not necessarily property (ii), because  $\|f\| = 0$  may not imply  $f(x) \equiv 0$ . Such a measure is actually called a *semi-norm*. If we were to introduce here equivalence classes of functions, i.e., identify  $f(x)$  and  $g(x)$  if  $\|f - g\| = 0$ , then the measure  $\|\cdot\|$  becomes a norm in a natural way in the linear space composed of these equivalence classes. For simplicity, we refer to  $\|\cdot\|$  as a norm in this chapter, even though we do not formally introduce the equivalence classes of functions. Once such a norm has been defined three questions are naturally suggested:

- (a) Does a polynomial exist, of a specified maximum degree, which minimizes the error?
- (b) If such a polynomial exists, is it unique?
- (c) If such a polynomial exists, how can it be determined?

With the convention that, unless otherwise specified,  $P_n(x)$  represents a polynomial of degree at most  $n$ , it is clear that

$$(1) \quad \text{g.l.b. } \underset{\{P_n(x)\}}{\|f(x) - P_n(x)\|} \equiv d_n \geq 0,$$

is a monotonic non-increasing function of  $n$ . If there exists a unique polynomial  $P_n(x)$  for which the minimum error is attained, we may then investigate methods for determining  $P_n(x)$  and the magnitude of  $d_n$  (or any other norm of the deviation). In particular, we are most interested in those approximation methods for which  $d_n \rightarrow 0$  as  $n \rightarrow \infty$ .

An example for which questions (a), (b), and (c) are easily answered is furnished by a well-known polynomial approximation: the first  $m + 1$  terms in the Taylor expansion of  $f(x)$  about  $x_0$ . That is, we consider the

linear space composed of functions  $f(x)$  which have derivatives of order  $n$  at  $x = x_0$ . For any given  $f(x)$  in this space we define

$$(2) \quad P_m(x) \equiv f(x_0) + \frac{(x - x_0)}{1!} f^{(1)}(x_0) + \frac{(x - x_0)^2}{2!} f^{(2)}(x_0) \\ + \cdots + \frac{(x - x_0)^m}{m!} f^{(m)}(x_0), \quad m \leq n.$$

This polynomial clearly minimizes the measure of error

$$(3) \quad \|g(x)\|^{(n)} \equiv \sum_{k=0}^n |g^{(k)}(x_0)|,$$

with  $g(x) \equiv f(x) - Q_m(x)$  among all polynomials  $\{Q_m\}$  of degree at most  $m$ , since

$$\|f(x) - P_m(x)\|^{(n)} = \begin{cases} 0, & \text{if } m = n, \\ \sum_{k=m+1}^n |f^{(k)}(x_0)|, & \text{if } m < n. \end{cases}$$

Thus existence and explicit construction of a best approximating polynomial for this somewhat contrived norm† are demonstrated. Uniqueness of  $P_m(x)$  for a given  $m \leq n$  can be proven by assuming that there is some other polynomial of degree at most  $m$ , say  $Q_m(x)$ , which also minimizes the error. By expressing  $Q_m(x)$  as a polynomial in  $(x - x_0)$  we find that the coefficients  $Q_m^{(k)}(x_0)/k!$  must be identical with those of  $P_m(x)$  given in (2), since otherwise,  $\|f - Q_m\|^{(n)} > \|f - P_m\|^{(n)}$ .

Now, if  $f(x)$  has an  $(n+1)$ st derivative in some interval about  $x_0$ , say  $|x - x_0| \leq a$ , then by Taylor's theorem the remainder in the expansion (or we may call it the pointwise error in the approximation) is given by

$$(4) \quad R_n(x) \equiv f(x) - P_n(x) = \frac{(x - x_0)^{n+1}}{(n+1)!} f^{(n+1)}(\xi),$$

where  $|x - x_0| \leq a$  and  $\xi = \xi(x)$  is some point in the open interval  $(x, x_0)$ . For the special function  $f(x) \equiv 1/(1+x)$  in the interval  $-\frac{1}{2} \leq x \leq 2$  and  $x_0 = 0$ , we find  $P_n(x) \equiv 1 - x + \cdots + (-1)^n x^n$  and

$$R_n(x) \equiv \frac{(-1)^{n+1} x^{n+1}}{1+x}.$$

In this case, although  $\|R_n(x)\|^{(n)} = 0$ , we note that for the *maximum norm*

$$\|R_n(x)\|_\infty \equiv \underset{-\frac{1}{2} \leq x \leq 2}{\text{l.u.b.}} |R_n(x)| \geq \frac{2^{n+1}}{3}$$

† Note that we may have  $\|g(x)\|^{(n)} = 0$  but  $g(x) \not\equiv 0$ , e.g.,  $g(x) \equiv (x - x_0)^{n+1}$ . Thus, on the indicated space, (3) is an example of a semi-norm which is not a norm.

and hence some caution must be used. What may be a good approximation when measured in one norm [i.e., in  $\|\cdot\|^{(n)}$  of (3)] may be a very poor approximation in another norm (i.e.,  $\|\cdot\|_\infty$ ). But, in this example, if the interval were  $-\frac{1}{2} \leq x \leq \frac{1}{2}$ , then

$$\|R_n(x)\|_\infty \equiv \underset{-\frac{1}{2} \leq x \leq \frac{1}{2}}{\text{l.u.b.}} |R_n(x)| \leq 2^{-n}$$

and we find that the Taylor series,  $P_n(x)$ , converges uniformly with respect to  $x$  for this function. In fact, the series converges uniformly in any closed interval contained in the open interval  $(-1, 1)$ . The latter property of Taylor series is typical and is more fully treated in the study of analytic functions of a complex variable.

The questions (a), (b), and (c) can be answered in many other specific cases and we do this for several different norms in this chapter. However, question (a), of existence, can be given an affirmative answer quite generally. We do this in Theorem 1. The answer to question (b), on uniqueness, is a qualified yes given in Theorem 2. (For all the specific approximation problems treated in this chapter the answer is yes.) A general answer to question (c) is not known but we show how to construct the minimizing polynomial for several norms.

For the general results to be presented we assume that the polynomials and the functions,  $f(x)$ , to be approximated are in a linear space  $C[a, b]$  of functions defined on the closed bounded interval,  $[a, b]$ . Then we have

**THEOREM 1.** *Let the measure of deviation  $\|\cdot\|$  be defined in  $C[a, b]$ , and let there exist positive numbers  $m_n$  and  $M_n$  which satisfy*

$$(5) \quad 0 < m_n \leq \left| \sum_{j=0}^n b_j x^j \right| \leq M_n, \quad n = 0, 1, \dots,$$

for all  $\{b_j\}$  such that

$$\sum_{j=0}^n |b_j|^2 = 1.$$

Then for any integer  $n$  and  $f(x)$  in  $C[a, b]$  there exists a polynomial of degree at most  $n$  for which

$$\|f(x) - P_n(x)\|$$

attains its minimum over all such polynomials.

*Proof.* Write the general  $n$ th degree polynomial as

$$P_n(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

and consider the function of the  $n + 1$  coefficients

$$\phi(a_0, a_1, \dots, a_n) \equiv \|f(x) - P_n(x)\|.$$

By the properties (iii) and (iv) of norms and the hypothesis (5), we obtain the continuity of  $\phi$ , as follows:

$$\begin{aligned}\phi(a_0 + \epsilon_0, a_1 + \epsilon_1, \dots, a_n + \epsilon_n) &= \left\| f(x) - P_n(x) - \sum_{j=0}^n \epsilon_j x^j \right\|, \\ &\leq \phi(a_0, a_1, \dots, a_n) + \left\| \sum_{j=0}^n \epsilon_j x^j \right\|, \\ &\leq \phi(a_0, a_1, \dots, a_n) + \sum_{j=0}^n |\epsilon_j| \cdot \|x^j\|, \\ &\leq \phi(a_0, a_1, \dots, a_n) + M_n \sum_{j=0}^n |\epsilon_j|.\end{aligned}$$

Similarly, we find

$$\begin{aligned}\phi(a_0, a_1, \dots, a_n) &= \|f(x) - P_n(x) - \sum_{j=0}^n \epsilon_j x^j + \sum_{j=0}^n \epsilon_j x^j\| \\ &\leq \phi(a_0 + \epsilon_0, a_1 + \epsilon_1, \dots, a_n + \epsilon_n) + M_n \sum_{j=0}^n |\epsilon_j|.\end{aligned}$$

Hence for any  $\{a_j\}$  and  $\{\epsilon_j\}$

$$(6) \quad \begin{aligned}|\phi(a_0 + \epsilon_0, a_1 + \epsilon_1, \dots, a_n + \epsilon_n) - \phi(a_0, a_1, \dots, a_n)| \\ \leq M_n \sum_{j=0}^n |\epsilon_j|.\end{aligned}$$

This demonstrates that  $\phi(a_0, a_1, \dots, a_n)$  is a continuous function of the coefficients  $(a_0, a_1, \dots, a_n)$ . (Compare Lemma 1.1 of Chapter 1.)

Since  $\phi(a_0, a_1, \dots, a_n) \geq 0$ , the “minimum deviation” in (1) can be characterized as:

$$(7) \quad \underset{(a_0, a_1, \dots, a_n)}{\text{g.l.b.}} \phi(a_0, a_1, \dots, a_n) \equiv d_n \geq 0.$$

Thus, the existence problem is reduced to showing that there is a set of coefficients, say  $(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_n)$ , such that

$$\phi(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_n) = d_n.$$

However, since  $\phi$  is continuous, the result will follow from a theorem of Weierstrass if we can show that  $d_n$  is the g.l.b. of  $\phi$  in an appropriate closed bounded domain of the coefficients. That is, we will show that for some  $R > 0$ ,

$$(8) \quad d_n = \underset{\sum_{j=0}^n |a_j|^2 \leq R^2}{\text{g.l.b.}} \{\phi(a_0, a_1, \dots, a_n)\}.$$

Then since the continuous function  $\phi(a_0, a_1, \dots, a_n)$  attains its minimum  $d_n$  on the closed bounded set  $\sum_{j=0}^n |a_j|^2 \leq R^2$ , the theorem follows.

To verify (8) we observe, using (iii) and (iv), that

$$\|f - g\| \leq \|f\| + \|g\|,$$

and setting  $f \equiv u + g$  we get

$$(9) \quad \|u + g\| \geq \|u\| - \|g\|.$$

Therefore, for any constant  $\mu > 0$ , by (9) and (iii),

$$\begin{aligned} \|f - P_n(x)\| &\geq \|P_n(x)\| - \|f\|, \\ &\geq \mu \left| \frac{1}{\mu} P_n(x) \right| - \|f\|. \end{aligned}$$

Let us pick  $\mu$  such that  $\sum_{j=0}^n |a_j/\mu|^2 = 1$ . Then by (5) in the hypothesis of the theorem

$$\left| \frac{1}{\mu} P_n(x) \right| = \left| \sum_{j=0}^n \frac{a_j}{\mu} x^j \right| \geq m_n,$$

and the previous inequality implies

$$\|f - P_n(x)\| \geq \mu m_n - \|f\|.$$

So, if  $\mu$  satisfies

$$\mu \geq \frac{\|f\| + d_n + 1}{m_n},$$

then

$$\|f - P_n(x)\| \geq d_n + 1.$$

Thus, we conclude that (8) is valid with the choice  $R = (\|f\| + d_n + 1)/m_n$  and the proof is ended. ■

Observe that the function  $f(x)$  need not be continuous. Furthermore, note that this theorem gives no estimate of the magnitude of  $d_n$ . The semi-norm (3) satisfies condition (5) for  $x_0 = 0$ , with  $m_n = (n+1)^{-\frac{1}{2}}$ ,  $M_n = \left( \sum_{j=0}^n (j!)^2 \right)^{\frac{1}{2}}$ ; see Problem 2.

For the uniqueness result, we require the measure of deviation to be *strict*. By definition a norm  $\|\cdot\|$  is strict if

$$\|f + g\| = \|f\| + \|g\|$$

implies there exist constants  $\alpha, \beta$  such that  $|\alpha| + |\beta| \neq 0$  and

$$\alpha f(x) + \beta g(x) \equiv 0.$$

Now we state

**THEOREM 2.** *To the hypothesis of Theorem 1 add the requirement that  $|\cdot|$  is strict. Then the minimizing polynomial, say  $P_n(x)$ , is unique.*

*Proof.* Assume, for a given  $f(x)$ , that  $P_n(x)$  and  $Q_n(x)$  both minimize (1). Then from (7), (iii), and (iv), we find

$$d_n \leq \|f - \frac{1}{2}(P_n + Q_n)\| \leq \frac{1}{2}\|f - P_n\| + \frac{1}{2}\|f - Q_n\| = d_n.$$

Hence, equality holds throughout, and since  $|\cdot|$  is strict there exist non-trivial constants  $\alpha$  and  $\beta$  such that

$$\frac{\alpha}{2}[f(x) - P_n(x)] + \frac{\beta}{2}[f(x) - Q_n(x)] \equiv 0.$$

Now if  $\alpha = -\beta \neq 0$  then  $P_n(x) \equiv Q_n(x)$ . Otherwise,  $\alpha \neq -\beta$  and then  $f(x)$  must be a polynomial of degree at most  $n$ . In this case,  $d_n = 0$ , and using (5) it follows that  $P_n(x) \equiv f(x) \equiv Q_n(x)$ . ■

Again, observe that the function  $f(x)$  need not be continuous, as remarked after the proof of Theorem 1. This theorem is valid for a semi-norm which satisfies the appropriate additional conditions [i.e., (5) and strictness]. Of course, the minimizing polynomial may be unique even though the norm is not strict. Such an instance is furnished by the semi-norm (3) which is not strict. Further examples follow.

## PROBLEMS, SECTION 0

1. Show that if  $|\cdot|$  is a norm defined for polynomials and satisfies (i)–(iv), then

$$\text{g.l.b.} \left\| \sum_{j=0}^n b_j x^j \right\| \equiv m_n > 0.$$

[Hint: Prove that

$$\psi(a_0, a_1, \dots, a_n) \equiv \left\| \sum_{j=0}^n a_j x^j \right\|$$

is a continuous function of  $(a_0, a_1, \dots, a_n)$ .]

2. Verify that for the semi-norm

$$\|g(x)\|^{(n)} \equiv \sum_{k=0}^n |g^{(k)}(0)|,$$

(5) is satisfied with

$$m_n = (n+1)^{-\frac{1}{2}}, \quad M_n = \left[ \sum_{j=0}^n (j!)^2 \right]^{\frac{1}{2}}.$$

[Hint: Note that some  $b_j$  satisfies  $|b_j| \geq (n+1)^{-\frac{1}{2}}$ . On the other hand, apply Schwarz' inequality to estimate  $\sum_{j=0}^n (j!)|b_j|$ .]

## 1. WEIERSTRASS' APPROXIMATION THEOREM AND BERNSTEIN POLYNOMIALS

We are justified in seeking a close polynomial approximation to a continuous function throughout a finite interval because of the fundamental

WEIERSTRASS' APPROXIMATION THEOREM. *Let  $f(x)$  be any function continuous in the (closed) interval  $[a, b]$ . Then for any  $\epsilon > 0$  there exists an integer  $n = n(\epsilon)$  and a polynomial  $P_n(x)$  of degree at most  $n$  such that*

$$|f(x) - P_n(x)| < \epsilon,$$

for all  $x$  in  $[a, b]$ .

This theorem guarantees that arbitrarily close polynomial approximations are possible throughout a closed bounded interval provided only that the function being approximated is continuous. The statement of the theorem is one of existence and gives no hint about constructing such approximations. However, a simple and elegant constructive proof of this result is due to Bernstein and we shall present it in Theorem 1.

First, a basic notion in analysis must be recalled and some preliminary identities will be introduced. If  $f(x)$  is a continuous function in a closed interval, say  $[0, 1]$ ,† then the *modulus of continuity of  $f(x)$  in  $[0, 1]$  is defined as*

$$(1) \quad \omega(f; \delta) \equiv \underset{\substack{\{x, x' \text{ in } [0, 1] \\ |x - x'| \leq \delta\}}}{\text{l.u.b.}} |f(x) - f(x')|.$$

Since  $f(x)$  is continuous in a closed interval, and hence uniformly continuous, it follows that

$$\lim_{\delta \rightarrow 0} \omega(f; \delta) = 0.$$

If, in addition,  $f(x)$  satisfies a *Lipschitz condition* in  $[0, 1]$ , i.e., if

$$|f(x) - f(x')| \leq \lambda|x - x'|,$$

for  $x, x'$  in  $[0, 1]$  and some constant  $\lambda$ , then from (1):

$$\omega(f; \delta) \leq \lambda\delta.$$

The concept of a modulus of continuity is generally useful in analysis and its use will recur in our study.

† We need only consider this case since:

An arbitrary finite interval  $a \leq y \leq b$  is mapped 1-1 onto the unit interval  $0 \leq x \leq 1$  by the continuous change of variable:  $x = (a - y)/(a - b)$  or  $y = (b - a)x + a$ . Hence, if  $g(y)$  is continuous in  $[a, b]$ ,  $f(x) \equiv g((b - a)x + a)$  is continuous in  $[0, 1]$ . Now, if  $P_n(x)$  approximates  $f(x)$  in  $[0, 1]$  to within  $\epsilon$ , then  $Q_n(y) \equiv P_n((a - y)/(a - b))$  is a polynomial of degree at most  $n$  that is within  $\epsilon$  of  $g(y)$  in  $[a, b]$ .

The identities required all follow from the well-known binomial expansion

$$(2a) \quad (a + b)^n = \sum_{j=0}^n \binom{n}{j} a^j b^{n-j}.$$

Upon forming  $\partial(a + b)^n / \partial a$  and  $\partial^2(a + b)^n / \partial a^2$  we obtain the further identities

$$(2b) \quad a(a + b)^{n-1} = \sum_{j=0}^n \frac{j}{n} \binom{n}{j} a^j b^{n-j},$$

$$(2c) \quad \left(1 - \frac{1}{n}\right) a^2 (a + b)^{n-2} = \sum_{j=0}^n \left(\frac{j^2}{n^2} - \frac{j}{n^2}\right) \binom{n}{j} a^j b^{n-j}.$$

Now set  $a = x$ , and  $b = 1 - x$ ; define the  $n$ th degree polynomials

$$(3) \quad \beta_{n,j}(x) \equiv \binom{n}{j} x^j (1-x)^{n-j}, \quad j = 0, 1, \dots, n;$$

and the identities (2) become

$$(4a) \quad \sum_{j=0}^n \beta_{n,j}(x) = 1,$$

$$(4b) \quad \sum_{j=0}^n \frac{j}{n} \beta_{n,j}(x) = x,$$

$$(4c) \quad \sum_{j=0}^n \frac{j^2}{n^2} \beta_{n,j}(x) = \left(1 - \frac{1}{n}\right) x^2 + \frac{1}{n} x.$$

It should be noted that for  $x$  in  $[0, 1]$ ,  $\beta_{n,j}(x) \geq 0$ .

Let the unit interval  $[0, 1]$  be subdivided into  $n$  equal subintervals with the endpoints

$$(5) \quad x_j = \frac{j}{n}, \quad j = 0, 1, \dots, n.$$

We finally introduce the *Bernstein polynomial of degree  $n$  for the function  $f(x)$  on  $[0, 1]$*  by the definition:

$$(6) \quad B_n(f; x) \equiv \sum_{j=0}^n f(x_j) \beta_{n,j}(x).$$

This is, from (3), clearly a polynomial of degree at most  $n$  and has coefficients depending upon the values of  $f(x)$  at  $n + 1$  equally spaced points in  $[0, 1]$ .

**THEOREM 1.** *Let  $f(x)$  be any continuous function defined on  $[0, 1]$ . Then for all  $x$  in  $[0, 1]$ , and any positive integer  $n$ ,*

$$(7) \quad |f(x) - B_n(f; x)| \leq \frac{9}{4}\omega(f; n^{-\frac{1}{2}}).$$

*Proof.* From (4a) and (6) we may write

$$f(x) - B_n(f; x) = \sum_{j=0}^n [f(x) - f(x_j)]\beta_{n,j}(x) \equiv S_1(x) + S_2(x),$$

where we define, for any  $\delta > 0$ ,

$$S_1(x) \equiv \sum_{j: |x-x_j| \leq \delta} \dots, \quad S_2(x) \equiv \sum_{j: |x-x_j| > \delta} \dots.$$

Thus by the definition of  $\omega(f; \delta)$  and (4a),

$$\begin{aligned} |S_1(x)| &\leq \omega(f; \delta) \sum_{j: |x-x_j| \leq \delta} \beta_{n,j}(x) \\ &\leq \omega(f; \delta) \sum_{j=0}^n \beta_{n,j}(x) = \omega(f; \delta). \end{aligned}$$

For the remaining sum, since  $|x - x_j| > \delta$ , we note

$$\begin{aligned} f(x) - f(x_j) &= [f(x) - f(\xi_1)] + [f(\xi_1) - f(\xi_2)] + \dots \\ &\quad + [f(\xi_{p-1}) - f(\xi_p)] + [f(\xi_p) - f(x_j)], \end{aligned}$$

where  $p \equiv [|x - x_j|/\delta]$ ,† and  $\xi_1, \xi_2, \dots, \xi_p$  are  $p$  points inserted uniformly between  $(x, x_j)$  where each of the  $p + 1$  successive intervals is of length  $|x - x_j|/(p + 1) < \delta$ . Hence,

$$|f(x) - f(x_j)| \leq (p + 1)\omega(f; \delta) \leq \left(1 + \frac{|x - x_j|}{\delta}\right)\omega(f; \delta).$$

Therefore,

$$\begin{aligned} |S_2(x)| &\leq \omega(f; \delta) \left[ \sum_{j: |x-x_j| > \delta} \beta_{n,j}(x) + \frac{1}{\delta} \sum_{j: |x-x_j| > \delta} |x - x_j| \beta_{n,j}(x) \right] \\ &\leq \omega(f; \delta) \left[ 1 + \frac{1}{\delta^2} \sum_{j: |x-x_j| > \delta} (x - x_j)^2 \beta_{n,j}(x) \right] \\ &\leq \omega(f; \delta) \left[ 1 + \frac{1}{\delta^2} \sum_{j=0}^n (x - x_j)^2 \beta_{n,j}(x) \right]. \end{aligned}$$

From (5) and (4)

$$\sum_{j=0}^n (x - x_j)^2 \beta_{n,j}(x) = \frac{x(1-x)}{n} \leq \frac{1}{4n}.$$

†  $p \equiv [x]$  for  $x > 0$ , means  $p$  is the largest integer satisfying  $p \leq x$ .

Therefore,

$$|S_2(x)| \leq \omega(f, \delta) \left(1 + \frac{1}{4n\delta^2}\right)$$

and finally,

$$\begin{aligned} |f(x) - B_n(f; x)| &\leq |S_1(x)| + |S_2(x)| \\ &\leq \omega(f, \delta) \left(2 + \frac{1}{4n\delta^2}\right). \end{aligned}$$

The error estimate (7) is obtained, if we choose  $\delta = n^{-\frac{1}{2}}$ . ■

Weierstrass' theorem follows by picking  $n$  so large that  $\omega(f; n^{-\frac{1}{2}}) < 4\epsilon/9$ .

If  $f(x)$  satisfies a Lipschitz condition, we find easily:

**COROLLARY.** *Let  $f(x)$  satisfy a Lipschitz condition*

$$|f(x) - f(y)| \leq \lambda|x - y|,$$

for all  $x, y$  in  $[0, 1]$ . Then for all  $x$  in  $[0, 1]$ ,

$$(8) \quad |f(x) - B_n(f; x)| \leq \frac{9}{4}\lambda n^{-\frac{1}{2}}. \quad \blacksquare$$

It can be shown that the approximation given by  $B_n(f; x)$  may be better than is implied in this result (see Problem 1). However, in general, even if  $f(x)$  has  $p$  derivatives, the convergence is, at best, of order  $1/n$ . In fact, it can be shown that

$$\lim_{n \rightarrow \infty} n[B_n(f; x) - f(x)] = \frac{1}{2}f''(x)x(1-x), \quad \text{if } p \geq 2.$$

As such convergence is quite slow compared to that of many other polynomial approximation methods (see Theorem 9 of Section 3), the Bernstein polynomials are seldom used in practice. It should be emphasized, however, that they converge (uniformly) for any continuous function when many of the other polynomial approximations do not.

The Weierstrass approximation theorem is valid for functions of several variables which are continuous on appropriate sets. In fact, the Bernstein polynomials can again be employed to yield a constructive proof in various cases (see Problem 2).

## PROBLEMS, SECTION 1

1. If  $f(x)$  satisfies a Lipschitz condition with constant  $\lambda$  in  $[0, 1]$ , show  $E_n \equiv |f(x) - B_n(f; x)| \leq (\lambda/2)n^{-\frac{1}{2}}$ .

[Hint: Use Schwarz' inequality to get

$$\begin{aligned} E_n &= \left| \sum [f(x) - f(x_i)](\beta_{n,i}(x))^{\frac{1}{2}}(\beta_{n,i}(x))^{\frac{1}{2}} \right| \\ &\leq \left\{ \sum [f(x) - f(x_i)]^2 \beta_{n,i}(x) \right\}^{\frac{1}{2}}. \end{aligned}$$

Now estimate

$$\begin{aligned} \sum [f(x) - f(x_j)]^2 \beta_{n,j}(x) &\leq \lambda^2 \sum (x - x_j)^2 \beta_{n,j}(x) \\ &\leq \frac{\lambda^2}{4n}. \end{aligned}$$

**2.** Let  $f(x, y)$  be continuous on the closed unit square:  $0 \leq x \leq 1$ ,  $0 \leq y \leq 1$ . Then prove Weierstrass' theorem for this case by employing the polynomials:

$$B_{m,n}(f; x, y) \equiv \sum_{j=0}^m \sum_{k=0}^n f\left(\frac{j}{m}, \frac{k}{n}\right) \beta_{m,j}(x) \beta_{n,k}(y).$$

Show how to extend this theorem to functions of more variables continuous in arbitrary "cubes" with faces parallel to the coordinate planes.

[Hint: Let  $R(x, y) \equiv f(x, y) - B_{m,n}(f; x, y)$ . Show that

$$\begin{aligned} R(x, y) &= \sum_{j,k} \left[ f(x, y) - f\left(\frac{j}{m}, \frac{k}{n}\right) \right] \beta_{m,j}(x) \beta_{n,k}(y), \\ |R| &\leq \left| \sum_{j,k} \sum_{\substack{|x-x_j| \leq \delta \\ |y-y_k| \leq \delta}} \right| + \left| \sum_{j} \sum_{|x-x_j| > \delta} \right| + \left| \sum_{k} \sum_{|y-y_k| > \delta} \right| \end{aligned}$$

and use the reasoning of the one variable case.

To prove the theorem when  $f(x, y)$  is continuous in a square,  $0 \leq x - a \leq c$ ,  $0 \leq y - b \leq c$ : define  $g(u, v) \equiv f(cu + a, cv + b)$  for  $0 \leq u, v \leq 1$ . Then construct  $B_{m,n}(g; u, v)$ . Finally, set

$$P_{m,n}(x, y) \equiv B_{m,n}\left(g; \frac{x-a}{c}, \frac{y-b}{c}\right)$$

and observe that  $|f(x, y) - P_{m,n}(x, y)|$  can be made small.]

**3.\*** If  $f(x)$  has a continuous first derivative in  $[0, 1]$ , show that the first derivatives of the Bernstein polynomials which approximate  $f(x)$  converge to  $f'(x)$  uniformly on  $[0, 1]$ .

[Hint: Verify that

$$\begin{aligned} \beta'_{n,j} &= n(\beta_{n-1,j-1} - \beta_{n-1,j}) \quad \text{for } j = 1, \dots, n-1, \\ \beta'_{n,n} &= n\beta_{n-1,n-1}, \quad \beta'_{n,0} = -n\beta_{n-1,0}. \end{aligned}$$

Then regroup the sum in terms of  $\beta_{n-1,k}$  for  $k = 0, 1, \dots, n-1$ .]

## 2. THE INTERPOLATION POLYNOMIALS

An approximation polynomial which is equal to the function it approximates at a number of *specified* points is called an *interpolation polynomial*. Given the  $n+1$  *distinct* points  $x_i$ ,  $i = 0, 1, \dots, n$  and corresponding

function values  $f(x_i)$ , the interpolation polynomial of degree at most  $n$  minimizes the norm:<sup>†</sup>

$$(1) \quad \|f(x) - P_n(x)\|_{\mathcal{S}} \equiv \sum_{i=0}^n |f(x_i) - P_n(x_i)|.$$

We shall show that such a polynomial exists (by explicit construction) and is unique; in fact, that the minimum value of the norm in (1) is 0.

The least possible value of  $\|f - P_n\|_{\mathcal{S}}$  is, of course, zero. Thus, we seek a polynomial

$$(2) \quad Q_n(x) = \sum_{k=0}^n a_k x^k,$$

for which  $Q_n(x_i) = f(x_i)$ . By considering the coefficients  $a_k$  in (2) as unknowns, we have the system of  $n + 1$  linear equations

$$(3) \quad Q_n(x_i) = a_0 + a_1 x_i + \cdots + a_n x_i^n = f(x_i), \quad i = 0, 1, \dots, n.$$

This system has a unique solution if the coefficient matrix is non-singular. The determinant of this matrix is called a *Vandermonde* determinant and can be easily evaluated (see Problem 1) to yield

$$(4) \quad \begin{vmatrix} 1 & x_0 & \cdots & x_0^n \\ 1 & x_1 & \cdots & x_1^n \\ \vdots & \vdots & & \vdots \\ 1 & x_n & \cdots & x_n^n \end{vmatrix} = \prod_{i>j} (x_i - x_j) \equiv \prod_{j=0}^{n-1} \left[ \prod_{i=j+1}^n (x_i - x_j) \right].$$

Since the  $\{x_i\}$  are distinct points, the determinant does not vanish and (3) may be uniquely solved for the  $a_i$  to determine the interpolation polynomial. Another proof of uniqueness is given in Lemma 1.

Rather than solve the system (3), we may use an alternative procedure to obtain the interpolation polynomial directly. Set

$$(5a) \quad P_n(x) = \sum_{j=0}^n f(x_j) \phi_{n,j}(x),$$

where the  $n + 1$  functions  $\phi_{n,j}(x)$  are  $n$ th degree polynomials. We note that  $P_n(x_i) = f(x_i)$  if the polynomials  $\phi_{n,j}(x)$  satisfy:

$$\phi_{n,j}(x_i) = \delta_{ij}, \quad i, j = 0, 1, \dots, n.$$

<sup>†</sup> This is only a semi-norm. Of course, Theorem 0.1 shows that there is a  $P_n(x)$  which minimizes the norm in (1), but we prove more here, namely that  $d_n = 0$ . We also prove uniqueness which is not covered by Theorem 0.2, since (1) is not strict.

Such polynomials are easily constructed, since the  $\{x_i\}$  are distinct, i.e.,

$$(6a) \quad \phi_{n,j}(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_{j-1})(x - x_{j+1}) \cdots (x - x_n)}{(x_j - x_0)(x_j - x_1) \cdots (x_j - x_{j-1})(x_j - x_{j+1}) \cdots (x_j - x_n)},$$

$$j = 0, 1, \dots, n.$$

By introducing  $\omega_n(x) \equiv (x - x_0)(x - x_1) \cdots (x - x_n)$  we find that (6a) can be written in the brief form [where  $\omega_n'(x_j) = (d\omega_n(x)/dx)_{x=x_j}$ ]:

$$(6b) \quad \phi_{n,j}(x) = \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)}.$$

The interpolation polynomial, especially when in the form (5), is called the *Lagrange interpolation polynomial* and the polynomials (6) are called the *Lagrange interpolation coefficients*. We can use the product notation for  $\phi$  which yields

$$(5b) \quad P_n(x) = \sum_{j=0}^n f(x_j) \prod_{\substack{k=0 \\ k \neq j}}^n \frac{x - x_k}{x_j - x_k}.$$

That the Lagrange interpolation polynomial is identical to the polynomial defined by (2) and (3) is a consequence of the following

**LEMMA 1.** *Let  $P_n(x)$  and  $Q_n(x)$  be any two polynomials, of degree at most  $n$ , for which*

$$P_n(x_i) = Q_n(x_i), \quad i = 0, 1, 2, \dots, n,$$

*where the  $n + 1$  points  $\{x_i\}$  are distinct. Then  $P_n(x) \equiv Q_n(x)$ .*

*Proof.* Define the polynomial

$$D_n(x) \equiv P_n(x) - Q_n(x),$$

which is of degree at most  $n$ . This polynomial has at least  $n + 1$  distinct roots:

$$D_n(x_i) = 0, \quad i = 0, 1, \dots, n.$$

However, the only polynomial of degree at most  $n$  with more than  $n$  roots is the identically vanishing “polynomial”  $D_n(x) \equiv 0$ . ■

In summary, there is only one polynomial of degree at most  $n$  for which (1) vanishes and it is given by (5) and (6). Of course, there are many other ways of representing this polynomial; since such considerations are of great practical interest they form a large part of the next chapter.

## 2.1. The Pointwise Error in Interpolation Polynomials

The *pointwise error* between a function,  $f(x)$ , and some polynomial approximation to it,  $P_n(x)$ , is defined as

$$(7) \quad R_n(x) \equiv f(x) - P_n(x).$$

It is, of course, quite useful to have an explicit expression for this error and, if possible, simple bounds on it. Such information may yield, as in the example of Section 0, an estimate of the rapidity of convergence of  $P_n(x)$  to  $f(x)$  as  $n \rightarrow \infty$  (or of divergence). Further, it facilitates a comparison of the different types of polynomial approximation.

For interpolation polynomials, a useful representation of  $R_n(x)$  is readily obtained. This result may be stated as

**THEOREM 1.** *Let  $f(x)$  have an  $(n + 1)$ st derivative,  $f^{(n+1)}(x)$ , in an interval  $[a, b]$ . Let  $P_n(x)$  be the interpolation polynomial for  $f(x)$  with respect to  $n + 1$  distinct points  $x_i$ ,  $i = 0, 1, \dots, n$  in the interval  $[a, b]$  (i.e.,  $P_n(x_i) = f(x_i)$  and  $x_i \in [a, b]$ ). Then for each  $x \in [a, b]$  there exists a point  $\xi = \xi(x)$  in the open interval:*

$$(8) \quad \min(x_0, x_1, \dots, x_n, x) < \xi < \max(x_0, x_1, \dots, x_n, x),$$

such that

$$(9) \quad f(x) - P_n(x) \equiv R_n(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_n)}{(n + 1)!} f^{(n+1)}(\xi) \\ \equiv \frac{\omega_n(x)}{(n + 1)!} f^{(n+1)}(\xi).$$

*Proof.* Since

$$R_n(x_0) = R_n(x_1) = \cdots = R_n(x_n) = 0,$$

we define  $S_n(x)$ , for any  $x \neq x_i$ , by setting

$$(10) \quad R_n(x) \equiv (x - x_0)(x - x_1) \cdots (x - x_n) S_n(x) = \omega_n(x) S_n(x).$$

Considering  $x$  to be fixed as above we also define a function  $F(z)$  by

$$F(z) \equiv f(z) - P_n(z) - \omega_n(z) S_n(x).$$

Clearly, this function and its derivatives with respect to  $z$  are defined and continuous wherever  $f(z)$  and its derivatives are defined and continuous; thus,  $F^{(n+1)}(z)$  is defined in  $[a, b]$ . (See Problem 5 for a mild generalization.)

We see that  $F(z)$  vanishes at  $n + 2$  distinct points in  $[a, b]$ , namely

$$F(x_0) = F(x_1) = \cdots = F(x_n) = F(x) = 0.$$

Thus, there are  $n + 1$  adjacent intervals in  $[a, b]$  at whose endpoints  $F(z)$  vanishes. Rolle's theorem is now applicable, since  $F'(z)$  is defined in  $[a, b]$ . Therefore, in the interior of each of these intervals, there is at least one point at which  $F'(z)$  vanishes. Thus, there are at least  $n + 1$  distinct points in the interval (8) at which  $F'(z) = 0$ . They form at least  $n$  intervals

such that in the interior of each, by another application of Rolle's theorem, the derivative of  $F'(z)$  vanishes. That is,  $F''(z) = 0$  for at least  $n$  distinct points in (8). By continuing this process we find that there is some point, say  $\xi$ , in (8) at which the  $(n + 1)$ st derivative of  $F(z)$  vanishes.

However, since  $P_n(z)$  is an  $n$ th degree polynomial,

$$\frac{d^{n+1}P_n(z)}{dz^{n+1}} = 0,$$

and a simple calculation yields

$$\frac{d^{n+1}}{dz^{n+1}} [\omega_n(z)S_n(x)] = (n + 1)!S_n(x).$$

Thus,

$$F^{(n+1)}(z) = f^{(n+1)}(z) - (n + 1)!S_n(x),$$

and since  $F^{(n+1)}(\xi) = 0$  we obtain

$$(11) \quad S_n(x) = \frac{1}{(n + 1)!} f^{(n+1)}(\xi), \quad \text{for } x \neq x_0, x_1, \dots, x_n.$$

With this in (10) the theorem follows. It should be pointed out that, although  $S_n(x)$  is not defined for  $x = x_i$ , the final result (9) is valid for all  $x$  in  $[a, b]$  [in fact, since  $R_n(x_i) = 0$ ,  $\xi$  for these values of  $x$  may be picked arbitrarily]. ■

If the maximum and minimum of  $f^{(n+1)}(x)$  in  $[a, b]$  can be determined, (9) will yield bounds on the error. It should be noted that the error (9) for interpolation polynomials is similar to the remainder in Taylor's expansion (0.4). In fact, we might naively assume that if  $|x - x_i| < |x - x_0|$  for  $i = 1, 2, \dots, n$  then the interpolation polynomial error is smaller than the error in Taylor's expansion about the point  $x_0$ . This assumption is not always justified since the terms  $f^{(n+1)}(\xi)$  in (0.4) and (9) are not evaluated at the same point  $\xi$  for a given  $x$ .

Does the sequence of interpolation polynomials  $\{P_n(x)\}$  converge to  $f(x)$  in  $[a, b]$  if  $\{(x_0^{(n)}, \dots, x_n^{(n)})\}$  covers  $[a, b]$ ? This is a question that is not completely answered. In the case of uniform spacing [i.e.,  $x_0^{(n)} = a$ ,  $x_j^{(n)} = x_0 + jh_n$ ,  $h_n = (b - a)/n$ ], we illustrate the fact that divergence is to be expected by studying Runge's example,  $f(x) = 1/(1 + x^2)$  over  $[-5, 5]$  (see Chapter 6, Subsection 3.4). On the other hand, in Corollary 2, Theorem 2 of Section 5, we exhibit a sequence of non-uniformly spaced points,  $\{(x_0^{(n)}, \dots, x_n^{(n)})\}$ , for which uniform convergence of  $\{P_n(x)\}$  to  $f(x)$  may be established for any function  $f(x)$  with continuous second derivatives.†

† Amazingly, for any sequence  $\{(x_0^{(n)}, x_1^{(n)}, \dots, x_n^{(n)})\}$ , there exists a *continuous* function  $f(x)$  for which  $|P_n(x) - f(x)| \rightarrow 0$ !

From the remainder theorem we can deduce some interesting and useful *identities* satisfied by the Lagrange interpolation coefficients. Since  $d^{n+1}x^m/dx^{n+1} = 0$  for  $m = 0, 1, \dots, n$ , the interpolation polynomials of degree  $n$  represent exactly all polynomials of degree at most  $n$ . Thus, with  $f(x) \equiv x^m$  in (5), equation (9) yields

$$(12) \quad \sum_{j=0}^n x_j^m \phi_{n,j}(x) = x^m, \quad m = 0, 1, \dots, n.$$

The case  $m = 0$  is particularly useful. [Compare (12) with equation (1.4) for the Bernstein polynomials.]

## 2.2. Hermite or Osculating Interpolation

The *osculating polynomial*, a generalization of the interpolation polynomial, is obtained by requiring agreement at the distinct points of interpolation,  $x_j$ , with the first  $r_j - 1$  derivatives of  $f(x)$ . (This polynomial arises also as the limit of the interpolation polynomials when  $r_j$  points of ordinary interpolation approach each other at the point  $x_j$ .) This procedure contains, as special cases, Taylor's expansion and ordinary interpolation. The number of combinations is boundless, but, in fact, a representation of the osculating polynomial can be found together with an expression for the pointwise error (see Chapter 6, Section 1, Problem 10). We shall consider in detail the case in which the function and its first derivative are to be assigned at each point of interpolation. This special procedure is usually called *Hermite or osculatory interpolation*.

The problem is to find a polynomial of least degree, say  $H_{2n+1}(x)$ , such that:

$$(13) \quad \begin{aligned} f(x_j) &= H_{2n+1}(x_j), \\ &\quad j = 0, 1, \dots, n. \\ f'(x_j) &= H'_{2n+1}(x_j), \end{aligned}$$

By counting the data (i.e.,  $2n + 2$  conditions), we find that a polynomial of degree  $2n + 1$  has the required number of undetermined coefficients. Thus, in analogy with the Lagrange interpolation formula (5), we seek a representation in the form

$$(14) \quad H_{2n+1}(x) = \sum_{j=0}^n f(x_j) \psi_{n,j}(x) + \sum_{j=0}^n f'(x_j) \Psi_{n,j}(x).$$

Here the polynomials  $\psi_{n,j}(x)$  and  $\Psi_{n,j}(x)$  are required to be of degree at most  $2n + 1$  and to satisfy

$$(15) \quad \begin{aligned} \psi_{n,j}(x_i) &= \delta_{ij}, & \Psi_{n,j}(x_i) &= 0, \\ &\quad i, j = 0, 1, \dots, n. \\ \psi'_{n,j}(x_i) &= 0, & \Psi'_{n,j}(x_i) &= \delta_{ij}, \end{aligned}$$

Such polynomials are given in terms of the Lagrange interpolation coefficients,  $\phi_{n,j}(x)$ , as:

$$(16) \quad \begin{aligned} \psi_{n,j}(x) &\equiv [1 - 2\phi'_{n,j}(x_j)(x - x_j)]\phi_{n,j}^2(x), \\ \Psi_{n,j}(x) &\equiv (x - x_j)\phi_{n,j}^2(x). \end{aligned}$$

The error in using (14) to approximate  $f(x)$  is

$$(17) \quad f(x) - H_{2n+1}(x) = \frac{\omega_n^2(x)}{(2n+2)!} f^{(2n+2)}(\xi),$$

provided  $f(x)$  has a continuous derivative of order  $2n + 2$ . The point  $\xi = \xi(x)$  is again in the interval determined by the points  $x, x_0, \dots, x_n$ . The proof of formula (17) is left to Problem 3.

From equation (17) we easily deduce, in analogy with (12), the *identities*:

$$(18) \quad \sum_{j=0}^n x_j^m \psi_{n,j}(x) + m \sum_{j=0}^n x_j^{m-1} \Psi_{n,j}(x) = x^m,$$

$$m = 0, 1, \dots, 2n + 1.$$

## PROBLEMS, SECTION 2

1. Evaluate the Vandermonde determinant to verify (4).

[Hint: Let each  $x_j$ ,  $j = 0, 1, \dots, n$ , in order, be considered variable and determine all the roots of the resulting polynomial. The remaining scalar factor is obtained by evaluating the coefficient of any specific term, say  $(1 \cdot x_1 \cdot x_2^2 \cdot \dots \cdot x_n^n)$ . An alternative proof could be given by using mathematical induction and expanding the determinant with respect to the elements of the last column.]

2. Prove that the system (3) is non-singular by assuming that the homogeneous system has a non-trivial solution and using Lemma 1 to obtain a contradiction.

3. Derive the error formula, equation (17), for Hermite interpolation if  $f(x)$  is sufficiently differentiable.

[Hint: Proceed exactly as in the derivation of the interpolation error and define:  $F(z) \equiv f(z) - H_{2n+1}(z) - \omega_n^2(z)S_n(x)$ . After the first application of Rolle's theorem, however,  $F'(z)$  will have  $2n + 2$  distinct zeros.]

4. Formulate the definition of the Hermite interpolation polynomial as a minimizing polynomial for the appropriate semi-norm. Does this semi-norm satisfy hypothesis (0.5) of Theorem 0.1? Is it strict?

5. Show that the conclusion of Theorem 1 follows under the weaker assumption:  $f(x)$  is continuous in the closed interval  $[a, b]$ , but has the requisite derivatives only in the open interval  $(a, b)$ .

### 3. LEAST SQUARES APPROXIMATION

A property which is frequently used to determine an approximating polynomial,

$$(1) \quad Q_n(x) = a_0 + a_1x + \cdots + a_nx^n,$$

is that the  $L_2$  norm, or mean square error,

$$(2) \quad \|f(x) - Q_n(x)\|_2 \equiv \left\{ \int_a^b [f(x) - Q_n(x)]^2 dx \right\}^{1/2}$$

be a minimum. For the general polynomial (1) and any appropriate† function  $f(x)$ , we define the function of  $n + 1$  variables,

$$(3) \quad \phi(a_0, a_1, \dots, a_n) \equiv \int_a^b [f(x) - Q_n(x)]^2 dx.$$

The least squares polynomial approximation to  $f(x)$  of degree at most  $n$  is then determined by finding a point  $(\hat{a}_0, \hat{a}_1, \dots, \hat{a}_n)$  in the  $n + 1$  dimensional space for which  $\phi$  is a minimum.

**THEOREM 1.** *For each appropriate function  $f(x)$ , there is a unique least squares polynomial approximation of degree at most  $n$  which minimizes (2).*

*Proof.* The hypotheses of Theorems 0.1 and 0.2 are satisfied by  $\|\cdot\|_2$  (see Problem 10), whence existence and uniqueness are established. ■

We now give an analytical description of a method for calculating the coefficients  $\hat{\mathbf{a}} \equiv (\hat{a}_0, \hat{a}_1, \dots, \hat{a}_n)$  of the polynomial that minimizes  $\phi(\mathbf{a})$  in (3).

Since

$$(4) \quad \begin{aligned} \phi(a_0, a_1, \dots, a_n) &= \int_a^b f^2(x) dx - 2 \sum_{i=0}^n a_i \int_a^b x^i f(x) dx \\ &\quad + \sum_{i=0}^n \sum_{j=0}^n a_i a_j \int_a^b x^{i+j} dx, \end{aligned}$$

$\phi$  is a quadratic function in the variables  $a_i$ . Now, at the minimum of  $\phi$  the coefficients  $\hat{\mathbf{a}}$  must satisfy

$$\frac{\partial \phi(a_0, a_1, \dots, a_n)}{\partial a_k} \Bigg|_{\mathbf{a}=\hat{\mathbf{a}}} = 0, \quad k = 0, 1, \dots, n.$$

† The given function,  $f(x)$ , for this purpose need only be restricted so that it and its square are integrable over  $[a, b]$ . The *complete* linear space for which (2) is a norm consists of all such functions, if we identify two functions which differ only on a set of measure zero in  $[a, b]$ , and use the Lebesgue integral. But we do not pursue this avenue of generalization and shall, unless otherwise noted, consider only functions that are continuous, except at a finite number of points, where certain conditions will be specified.

From (4) this necessary condition becomes

$$\begin{aligned}
 0 &= \frac{\partial \phi}{\partial a_k} \Big|_{\hat{a}} = 0 - 2 \int_a^b x^k f(x) dx + \sum_{i=0}^n \hat{a}_i \int_a^b x^{i+k} dx \\
 (5a) \quad &\quad \quad \quad + \sum_{j=0}^n \hat{a}_j \int_a^b x^{k+j} dx \\
 &= 2 \left[ \sum_{i=0}^n \hat{a}_i \int_a^b x^{i+k} dx - \int_a^b x^k f(x) dx \right], \\
 &\quad k = 0, 1, \dots, n.
 \end{aligned}$$

A system of  $n+1$  linear equations for the determination of the  $\{\hat{a}_i\}$  is defined in (5a); it is frequently called the *normal system*.

We write the normal system (5a) in the form:

$$(5b) \quad \sum_{j=0}^n h_{ij} \hat{a}_j = c_i, \quad i = 0, 1, \dots, n;$$

where the coefficient matrix and right-hand side are given by

$$(5c) \quad H_{n+1}(a, b) \equiv (h_{ij}), \quad h_{ij} \equiv \int_a^b x^{i+j} dx; \quad c_i \equiv \int_a^b x^i f(x) dx.$$

Now we have

**THEOREM 2.** *The coefficient matrix  $H_{n+1}(a, b)$  is non-singular.*

*Proof.* For a given arbitrary vector  $(c_0, c_1, \dots, c_n)$  it is possible to find a polynomial  $f(x)$ , such that

$$\int_a^b x^k f(x) dx = c_k, \quad k = 0, 1, \dots, n.$$

In fact, the polynomial can be of degree at most  $n$  and we leave this construction to Problem 11. If

$$f(x) \equiv \sum_{i=0}^n a_i x^i,$$

then (5a) has the solution  $\hat{a}_i = a_i$ ,  $i = 0, 1, \dots, n$ . Therefore, the system (5b) has at least one solution for any right-hand side and this implies that the system is non-singular. ■

In the special case  $[a, b] \equiv [0, 1]$ , we get from (5c):

$$(6) \quad H_{n+1}(0, 1) \equiv \begin{Bmatrix} 1 & 1 & \cdots & \frac{1}{n+1} \\ \frac{1}{2} & \frac{1}{3} & \cdots & \frac{1}{n+2} \\ \vdots & \vdots & & \vdots \\ \frac{1}{n+1} & \frac{1}{n+2} & \cdots & \frac{1}{2n+1} \end{Bmatrix} \equiv (h_{ij}),$$

where  $h_{ij} = 1/(i + j - 1)$  and  $i, j = 1, 2, \dots, n + 1$ , a so-called *Hilbert segment* matrix. It is difficult to solve numerically a system of equations with the matrix  $H_n(0, 1)$ . (E.g., solve  $H_4 \mathbf{x} = (0, 0, 0, 1)^T$ : (a) exactly; (b) using four-decimal-place arithmetic.) However, it is possible to find the explicit inverse of  $H_n(0, 1)$  with the aid of some properties of the Lagrange interpolation coefficients (see Problems 1, 2). As a consequence of the non-singularity of  $H_n(a, b)$ , and the fact that (5) is a necessary condition for a minimum of  $\phi(a_0, a_1, \dots, a_n)$ , we have another proof that the least squares polynomial approximation of degree  $n$  is unique.

We may, by the linear change of variable  $x = a + (b - a)y$ , transform the problem of fitting  $f(x)$  by  $Q_n(x)$  in  $[a, b]$ , to that of fitting  $f(a + (b - a)y) \equiv g(y)$  by  $P_n(y)$  in  $[0, 1]$ . Afterwards, by setting

$$Q_n(x) \equiv P_n\left(\frac{x - a}{b - a}\right),$$

we have the least squares polynomial that minimizes the norm in (2).

The least squares polynomial can be determined in a way which avoids the difficulties inherent in directly solving the system (5). (This alternative procedure is but a special case of the general theory of approximation by *orthogonal functions*.)

Consider a set of  $n + 1$  polynomials  $\{P_k(x)\}$ ,  $k = 0, 1, \dots, n$ , where  $P_k(x)$  is of degree†  $k$  in  $x$ . Then without loss of generality, we may let  $Q_n(x)$  be a linear combination of these polynomials, say

$$(7) \quad Q_n(x) = \sum_{j=0}^n c_j P_j(x).$$

Now the mean square error (2) defines a function

$$(8) \quad J(c_0, c_1, \dots, c_n) = \int_a^b [f(x) - Q_n(x)]^2 dx,$$

of the  $n + 1$  variables  $\{c_k\}$ . As before, this function is quadratic in the  $c_k$  and at a minimum we must have

$$0 = \frac{\partial J}{\partial c_k} = 0 - 2 \int_a^b P_k(x)f(x) dx + 2 \sum_{j=0}^n c_j \int_a^b P_j(x)P_k(x) dx;$$

or the normal system

$$(9) \quad \sum_{j=0}^n c_j \int_a^b P_j(x)P_k(x) dx = \int_a^b P_k(x)f(x) dx, \quad k = 0, 1, \dots, n.$$

† Here we require that  $P_k(x)$  have *exactly* degree  $k$ , in order that the set  $\{P_k(x)\}$  for  $k = 0, 1, 2, \dots$ , be linearly independent.

There seems to be no apparent gain in replacing the system (5) by (9). However, the main point of the expansion in polynomials, rather than in powers of  $x$ , is to choose appropriate  $P_k(x)$  so that the system (9) is easily solved. In fact, the simplest choice would be one which makes the coefficient matrix diagonal, or, better still, the identity matrix. This requires

$$(10) \quad \int_a^b P_j(x)P_k(x) dx = \delta_{jk}.$$

In the next subsection, we define such a sequence  $\{P_k(x)\}$ . A set of polynomials (or any functions) which satisfy (10) are called *orthonormal over*  $[a, b]$ . The coefficients  $c_j$  are now simply given by

$$(11) \quad c_j = \int_a^b P_j(x)f(x) dx, \quad j = 0, 1, \dots, n.$$

An additional advantage of the expansion in orthonormal polynomials is that the accuracy of the approximation (7) can be improved by adding an additional term,  $c_{n+1}P_{n+1}(x)$ , without having to recompute the previously determined coefficients,  $c_0, c_1, \dots, c_n$ . [It is also clear that (7), with the coefficients (11), represents an approximation of least mean square error for any set of *orthonormal functions*, not necessarily polynomials, which satisfy (10).] For the approximation determined by (7) and (11), it easily follows from (10) in (8) that

$$(12a) \quad J(c_0, c_1, \dots, c_n) = \int_a^b f^2(x) dx - \sum_{j=0}^n c_j^2 \geq 0.$$

If we let  $n \rightarrow \infty$  it follows from (12a) that  $\sum_{j=0}^{\infty} c_j^2$  converges. Hence, we deduce that  $\lim_{j \rightarrow \infty} c_j = 0$ . This is a conclusion about the integrals of form (11) for general orthonormal functions,  $P_j(x)$ . The inequality (12a) is known as *Bessel's inequality*.

*Convergence in the mean* of the least squares polynomial approximation to a continuous function is easily demonstrated. Specifically we have

**THEOREM 3.** *Let  $f(x)$  be continuous on  $[a, b]$  and  $Q_n(x)$ ,  $n = 0, 1, 2, \dots$ , be the least squares polynomial approximations to  $f(x)$  on  $[a, b]$  determined by (7) and (11). Then*

$$\lim_{n \rightarrow \infty} J_n \equiv \lim_{n \rightarrow \infty} \int_a^b [f(x) - Q_n(x)]^2 dx = 0,$$

and we have *Parseval's equality*

$$(12b) \quad \int_a^b f^2(x) dx = \sum_{j=0}^{\infty} c_j^2.$$

*Proof.* For a proof by contradiction assume that  $\lim_{n \rightarrow \infty} J_n = \delta > 0$ . Then we pick  $\epsilon > 0$  such that  $\epsilon^2 = \delta/[2(b-a)]$  and by the Weierstrass theorem there is some polynomial  $P_m(x)$  such that  $|f(x) - P_m(x)| \leq \epsilon$  in  $a \leq x \leq b$ . For this polynomial,

$$\int_a^b [f(x) - P_m(x)]^2 dx \leq \epsilon^2(b-a) = \frac{\delta}{2}.$$

However, by (12a)  $J_n$  is a non-increasing function of  $n$ ; hence, the least squares approximation of degree  $m$ , say  $Q_m(x)$ , satisfies

$$\delta/2 \geq \int_a^b [f(x) - Q_m(x)]^2 dx \geq \delta.$$

This is a contradiction unless  $\delta = 0$ . Of course, this mean convergence implies (12b), the Parseval equality. ■

Unfortunately, these simple results yield no information about the pointwise approximation of  $f(x)$  by the least squares approximation  $Q_n(x)$ . In order to estimate  $R_n(x) \equiv f(x) - \sum_{j=0}^n c_j P_j(x)$ , with  $c_j$  defined by (11) and  $\{P_j(x)\}$  orthonormal, we write

$$\begin{aligned} R_n(x) &= f(x) - \sum_{j=0}^n P_j(x) \int_a^b P_j(\xi) f(\xi) d\xi \\ &= f(x) - \int_a^b G_n(x, \xi) f(\xi) d\xi, \end{aligned}$$

where

$$(13a) \quad G_n(x, \xi) \equiv \sum_{j=0}^n P_j(x) P_j(\xi).$$

From the orthogonality property, we observe that

$$\int_a^b G_n(x, \xi) d\xi = 1.$$

Therefore, we may rewrite  $R_n(x)$  as

$$(13b) \quad R_n(x) = \int_a^b G_n(x, \xi) [f(x) - f(\xi)] d\xi.$$

Now, the rate at which  $R_n(x) \rightarrow 0$ , as  $n \rightarrow \infty$ , depends on the nature of the *kernel*,  $G_n(x, \xi)$ , and on the function  $f(x)$ .

A direct verification of convergence is possible if the sequence  $Q_n(x) \equiv \sum_{j=0}^n c_j P_j(x)$  converges in the mean to  $f(x)$  and simultaneously converges uniformly in  $[a, b]$ . That is, if we define

$$g(x) \equiv \lim_{n \rightarrow \infty} Q_n(x),$$

the function  $g(x)$  will be continuous in  $[a, b]$ , since it is the uniform limit of a sequence of continuous functions. On the other hand, because of the uniformity of convergence, we may pass to the limit under the integral sign, in the statement of mean convergence, to find

$$\int_a^b [f(x) - g(x)]^2 dx = 0.$$

Therefore,  $f(x) = g(x)$ , since they are both continuous.

Now, it is possible to show that the sequence  $Q_n(x)$  converges uniformly if  $f(x)$  has two continuous derivatives† in  $[a, b]$ . We carry out the details for the interval  $[a, b] \equiv [-1, 1]$ , in Subsection 3.4. On the other hand, if the function  $f(x)$  is merely continuous, the sequence  $Q_n(x)$  need not converge.

### 3.1. Construction of Orthonormal Functions

The method by which a set of orthonormal polynomials  $\{P_k(x)\}$  can be determined is a special case of a general procedure in which an orthonormal set of functions is constructed from an arbitrary linearly independent set.†† This process is known as the *Gram-Schmidt orthonormalization method* and is described as follows.

We begin by defining the *inner product* of any pair of real valued functions  $f(x)$ ,  $g(x)$  by

$$(14) \quad (f, g) = (g, f) = \int_a^b f(x)g(x) dx.$$

Now, let  $\{g_i(x)\}$ ,  $i = 0, 1, \dots, n$ , be  $n + 1$  linearly independent and square integrable functions over  $[a, b]$ . Consider the functions

$$(15) \quad \begin{aligned} f_0(x) &= d_0[g_0(x)], \\ f_1(x) &= d_1[g_1(x) - c_{01}f_0(x)], \\ &\vdots \\ f_n(x) &= d_n[g_n(x) - c_{0n}f_0(x) - \cdots - c_{n-1,n}f_{n-1}(x)]. \end{aligned}$$

† This requirement may be weakened.

†† In analogy with the definition of linear independence for vectors, the set  $\{g_i(x)\}$ ,  $i = 0, 1, \dots, n$ , of functions are linearly independent in some interval  $[a, b]$  if and only if the only linear combination  $\sum_{i=0}^n a_i g_i(x)$  that vanishes identically in  $[a, b]$  has  $a_i = 0$ ,  $i = 0, 1, \dots, n$ .

We seek coefficients  $d_k, c_{jk}$  such that the set  $\{f_i(x)\}$  is orthonormal over  $[a, b]$ ; i.e., using the inner product notation:

$$(f_i, f_j) = \delta_{ij}, \quad i, j = 0, 1, \dots, n.$$

To normalize  $f_0(x)$  we need only require

$$(f_0, f_0) = d_0^2(g_0, g_0) = 1.$$

Since  $g_0(x) \not\equiv 0$ , or the set  $\{g_i(x)\}$  could not be linearly independent, we may define

$$(16)_0 \quad d_0 = \frac{1}{\sqrt{(g_0, g_0)}}.$$

In order that

$$\begin{aligned} 0 &= (f_0, f_1) = d_1(f_0, g_1 - c_{01}f_0) \\ &= d_1[(f_0, g_1) - c_{01}], \end{aligned}$$

we require

$$(16)_{01} \quad c_{01} = (f_0, g_1).$$

To normalize  $f_1$  we set

$$(f_1, f_1) = d_1^2(g_1 - c_{01}f_0, g_1 - c_{01}f_0) = 1.$$

The inner product on the right cannot vanish, by the assumed linear independence of the  $\{g_i(x)\}$ , and thus  $d_1$  is determined to within its sign. As in  $(16)_0$  we adopt the convention of using the positive square root.

In general, then, if  $(f_i, f_j) = \delta_{ij}$  for all  $i, j = 0, 1, \dots, k - 1$ , we find  $(f_j, f_k) = 0$  for all  $j = 0, 1, \dots, k - 1$ , when we define  $f_k$  as in (15) with

$$(16)_{jk} \quad c_{jk} = (f_j, g_k), \quad j \leq k - 1.$$

The normalization constant,  $d_k$ , is easily obtained as before by setting  $(f_k, f_k) = 1$ .

To apply the Gram-Schmidt procedure to the problem of obtaining orthonormal polynomials  $\{P_j(x)\}$  over an interval  $[a, b]$ , we observe that the  $n + 1$  polynomials

$$g_j(x) \equiv x^j, \quad j = 0, 1, \dots, n,$$

are linearly independent over any interval. The proof of their independence follows from the fundamental theorem of algebra used in the proof of Lemma 2.1. As in (15) we form

$$P_0(x) = d_0(1),$$

$$P_1(x) = d_1[x - c_{01}P_0(x)],$$

$$\vdots$$

$$P_n(x) = d_n[x^n - c_{0n}P_0(x) - c_{1n}P_1(x) - \cdots - c_{n-1,n}P_{n-1}(x)].$$

Then

$$1 = \int_a^b P_0^2(x) dx = d_0^2 \int_a^b dx = d_0^2(b - a),$$

or

$$(17)_0 \quad d_0 = \frac{1}{\sqrt{b - a}}.$$

By either repeating the previous derivation or simply applying the formulae  $(16)_{jk}$  we have

$$(17)_{01} \quad c_{01} = \int_a^b d_0 x dx = \frac{1}{\sqrt{b - a}} \frac{b^2 - a^2}{2} = \sqrt{b - a} \left( \frac{b + a}{2} \right).$$

Normalizing  $P_1(x) = d_1[x - c_{01}d_0]$  yields

$$\begin{aligned} 1 &= \int_a^b P_1^2(x) dx = d_1^2 \int_a^b \left[ x^2 - (b + a)x + \frac{(b + a)^2}{4} \right] dx \\ &= d_1^2 \left[ \frac{b^3 - a^3}{3} - (b + a) \frac{b^2 - a^2}{2} + \frac{(b + a)^2}{4} (b - a) \right] \\ &= \frac{d_1^2}{12} (b - a)^3 \end{aligned}$$

or explicitly

$$d_1 = 2\sqrt{3} (b - a)^{-\frac{3}{2}}.$$

The first two polynomials are thus

$$P_0(x) = \frac{1}{\sqrt{b - a}},$$

$$P_1(x) = 2\sqrt{3} (b - a)^{-\frac{3}{2}} \left( x - \frac{b + a}{2} \right),$$

and any number of them can be obtained by continuing this procedure. Let us denote by  $P_n(x; a, b)$ , the sequence of orthonormal polynomials over  $[a, b]$ . Then it is easily verified that

$$P_n(x; a, b) \equiv \sqrt{\frac{2}{b - a}} P_n \left( \frac{2x - (a + b)}{b - a}; -1, 1 \right)$$

is their representation in terms of the polynomials orthonormal over  $[-1, 1]$ . In Problem 6, we verify that

$$P_n(x; -1, 1) = (n + \frac{1}{2})^{\frac{n}{2}} \frac{1}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n.$$

The polynomials

$$Q_n(x) \equiv (n + \frac{1}{2})^{-\frac{1}{2}} P_n(x; -1, 1)$$

are called the *Legendre polynomials*.

### 3.2. Weighted Least Squares Approximation

The mean square measure of approximation defined in (2) gives equal “weight” at each point in  $[a, b]$  to the deviation of the approximating polynomial from the function,  $f(x)$ . For some purposes, it may be required that the approximation be better over some parts of the interval  $[a, b]$  than it is over other parts. This suggests a natural generalization of (2) which is:

$$(18) \quad \|f(x) - Q_n(x)\|_{2,w} \equiv \left\{ \int_a^b [f(x) - Q_n(x)]^2 w(x) dx \right\}^{\frac{1}{2}},$$

where  $w(x) \geq 0$  in  $[a, b]$  and

$$(19) \quad \int_a^b w(x) dx > 0.$$

The non-negative function  $w(x)$  is called the *weight function*; clearly, if  $w(x) \equiv 1$ , the usual least squares approximation results. For convenience, we require that  $w(x)$  be continuous in  $(a, b)$  and have at most isolated zeros in this interval. By choosing an appropriate function  $w(x)$  and finding the corresponding  $Q_n(x)$  which minimizes (18), we may obtain an approximation with good relative accuracy in a specified region of  $[a, b]$ . An extreme example of this is illustrated by the fact that the interpolation problem can be formulated as a special limiting case of (18) (see Problem 4).

If  $Q_n(x)$  is assumed of the form (1) then, as before, we find a system of equations for the determination of the coefficients  $\{a_i\}$ :

$$\sum_{i=0}^n \left[ \int_a^b x^{i+k} w(x) dx \right] a_i = \int_a^b x^k f(x) w(x) dx, \quad k = 0, 1, \dots, n.$$

Again, the necessity for solving this system can be eliminated by introducing a set of polynomials having appropriate properties. Specifically we call a set of functions  $\{P_j(x)\}$  *orthonormal over  $[a, b]$  with respect to the weight  $w(x)$*  if

$$(20) \quad \int_a^b P_j(x) P_k(x) w(x) dx = \delta_{jk}.$$

Then to minimize (18) with a polynomial of the form (7), we find that

$$(21) \quad c_j = \int_a^b P_j(x) f(x) w(x) dx.$$

The construction of orthonormal functions with respect to a weight  $w(x)$  can be accomplished by the procedure of the previous subsection. That is, we introduce, in place of (14), a new definition of inner product

$$(22) \quad (f, g) = (g, f) = \int_a^b f(x)g(x)w(x) dx.$$

The generalizations of Bessel's inequality (12a), the mean convergence proof for polynomial approximations of continuous functions and Parseval's relation (12b) are valid in the present case with essentially no change in argument. An important special example  $w(x) \equiv (1 - x^2)^{-\frac{1}{2}}$ ,  $[a, b] \equiv [-1, 1]$ , gives rise to the Chebyshev polynomials (see Problem 9). The pointwise convergence of weighted mean square approximations is briefly considered in Subsection 3.4. A proof of convergence for the Chebyshev expansion of sufficiently smooth functions is given there.

### 3.3. Some Properties of Orthogonal Polynomials

Let the polynomials  $P_n(x)$ ,  $n = 0, 1, 2, \dots$ , be orthogonal over  $a \leq x \leq b$  with respect to the non-negative weight function  $w(x)$ . Then we have

**THEOREM 4.** *The roots  $x_j$ ,  $j = 1, 2, \dots, n$  of  $P_n(x) = 0$ ,  $n = 1, 2, \dots$ , are all real and simple and lie in the open interval  $a < x_j < b$ .*

*Proof.* Let those roots of  $P_n(x) = 0$  in  $(a, b)$  be  $x_1, x_2, \dots, x_r$ , where any multiple root is repeated the appropriate number of times. Then the polynomial

$$Q_r(x) = (x - x_1)(x - x_2) \cdots (x - x_r)$$

has sign changes wherever  $P_n(x)$  does in  $(a, b)$  and it is of degree  $r \leq n$ . Thus,  $P_n(x)Q_r(x)$  is of one sign in  $(a, b)$  and so

$$\int_a^b P_n(x)Q_r(x)w(x) dx \neq 0.$$

This can only be true if  $r = n$ , since  $P_n(x)$  is orthogonal to all polynomials of lower degree. Now assume some root, say  $x_1$ , is multiple. Then we can write

$$P_n(x) = (x - x_1)^2 p_{n-2}(x),$$

where  $p_{n-2}(x)$  is of degree  $n - 2$ . But

$$P_n(x)p_{n-2}(x) = \left( \frac{P_n(x)}{x - x_1} \right)^2 \geq 0$$

and hence

$$\int_a^b P_n(x)p_{n-2}(x)w(x) dx > 0.$$

But this is a contradiction since  $P_n(x)$  is orthogonal to any lower order polynomial. Hence multiple roots cannot occur. ■

The orthonormal polynomials satisfy a simple recursion formula which is stated in

**THEOREM 5.** *Any three consecutive orthonormal polynomials are related by*

$$(23) \quad P_{n+1}(x) = (A_n x + B_n)P_n(x) - C_n P_{n-1}(x).$$

If  $a_k$  and  $b_k$  represent the coefficients of the terms of degree  $k$  and  $k - 1$  in  $P_k(x)$  then

$$(24) \quad A_n = \frac{a_{n+1}}{a_n}, \quad B_n = \frac{a_{n+1}}{a_n} \left( \frac{b_{n+1}}{a_{n+1}} - \frac{b_n}{a_n} \right), \quad C_n = \frac{a_{n+1}a_{n-1}}{a_n^2}.$$

*Proof.* With  $A_n$  given by (24) it follows that

$$P_{n+1}(x) - A_n x P_n(x) \equiv Q_n(x)$$

is a polynomial of degree at most  $n$ . Hence,  $Q_n(x)$  can be expanded as

$$Q_n(x) = \alpha_n P_n(x) + \cdots + \alpha_0 P_0(x).$$

By the orthogonality, however, we find that

$$\begin{aligned} \alpha_k &= \int_a^b Q_n(x) P_k(x) w(x) dx \\ &= \int_a^b P_{n+1}(x) P_k(x) w(x) dx - A_n \int_a^b P_n(x) P_k(x) x w(x) dx \\ &= 0, \quad \text{for } k = 0, 1, \dots, n-2. \end{aligned}$$

Thus, the form in (23) follows upon setting  $\alpha_n = B_n$  and  $\alpha_{n-1} = -C_n$ .

Now we may write

$$x P_{n-1}(x) \equiv \frac{a_{n-1}}{a_n} P_n(x) + q_{n-1}(x),$$

where  $q_{n-1}(x)$  is of degree at most  $n-1$ . Then it follows that

$$\begin{aligned} C_n &= A_n \int_a^b P_n(x) P_{n-1}(x) x w(x) dx, \\ &= A_n \frac{a_{n-1}}{a_n} \int_a^b P_n^2(x) w(x) dx + A_n \int_a^b P_n(x) q_{n-1}(x) w(x) dx \\ &= A_n \frac{a_{n-1}}{a_n}. \end{aligned}$$

The coefficient  $B_n$  is easily obtained by equating coefficients of the terms of degree  $n$  in (23) and the proof is completed. We observe that (23) and (24) are valid for  $n = 0$  if we define  $a_{-1} \equiv P_{-1}(x) \equiv 0$ . ■

The result in Theorem 5 can be used to derive what is known as the *Christoffel-Darboux relation*. We state this as

**THEOREM 6.** *The orthonormal polynomials satisfy*

$$(25) \quad \frac{a_n}{a_{n+1}} [P_{n+1}(x)P_n(\xi) - P_{n+1}(\xi)P_n(x)] = (x - \xi) \sum_{j=0}^n P_j(x)P_j(\xi).$$

*Proof.* Multiply the recursion formula (23) by  $P_n(\xi)$  to get:

$$P_n(\xi)P_{n+1}(x) = (A_n x + B_n)P_n(\xi)P_n(x) - C_n P_n(\xi)P_{n-1}(x).$$

Since this is an identity, it holds if we interchange the arguments  $x$  and  $\xi$ . Subtracting this interchanged form from the original form and multiplying by  $A_n^{-1}$  yields, with the aid of (24),

$$\begin{aligned} (x - \xi)P_n(x)P_n(\xi) &= A_n^{-1}[P_{n+1}(x)P_n(\xi) - P_{n+1}(\xi)P_n(x)] \\ &\quad - A_{n-1}^{-1}[P_n(x)P_{n-1}(\xi) - P_n(\xi)P_{n-1}(x)]. \end{aligned}$$

We now sum these identities over  $0, 1, \dots, n$  and the theorem follows (for  $n = 0$ , we use the convention  $a_{-1} = 0$ ) since  $A_n^{-1} = a_n/a_{n+1}$ . ■

Theorem 6 gives a convenient representation of the kernel  $G_n(x, \xi)$  defined in (13a).

### 3.4. Pointwise Convergence of Least Squares Approximations

We first consider the ordinary least squares approximation over  $[-1, 1]$  in which case the orthonormal polynomials [essentially the Legendre polynomials, see (29)] can be defined as

$$(26) \quad P_n(x; -1, 1) \equiv P_n(x) \equiv \frac{\sqrt{n + \frac{1}{2}}}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n; \quad n = 0, 1, 2, \dots$$

The derivation of this representation is contained in Problem 6. Given  $f(x)$ , we find the least squares polynomial approximation of degree at most  $n$  to be as in (7) and (11)

$$(27a) \quad Q_n(x) \equiv \sum_{j=0}^n c_j P_j(x);$$

$$(27b) \quad c_j \equiv \int_{-1}^1 f(x)P_j(x) dx.$$

If  $f(x)$  is continuous on  $[-1, 1]$ , it follows from Theorem 3 that

$$(28a) \quad \lim_{n \rightarrow \infty} \int_{-1}^1 [f(x) - Q_n(x)]^2 dx = 0,$$

$$(28b) \quad \lim_{n \rightarrow \infty} c_n = 0.$$

If  $f(x)$  satisfies additional smoothness conditions, we can deduce uniform convergence of  $Q_n(x)$  to  $f(x)$ . In fact, we have

**THEOREM 7.** *Let  $Q_n(x)$  be defined by (27) and let  $f(x)$  have a continuous second derivative on  $[-1, 1]$ . Then for all  $x \in [-1, 1]$  and any  $\epsilon > 0$*

$$|f(x) - Q_n(x)| \leq \frac{\epsilon}{\sqrt{n}}$$

*provided  $n$  is sufficiently large.*

*Proof.* We introduce the Legendre polynomials

$$(29) \quad p_n(x) = (n + \tfrac{1}{2})^{-\frac{1}{2}} P_n(x) = \frac{1}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n; \\ n = 0, 1, 2, \dots$$

some of whose properties are described in Problems 6–8. If we set  $u = x^2 - 1$  in (29), it easily follows that

$$\begin{aligned} 2^{n+1}(n+1)! p'_{n+1}(x) &= \frac{d^{n+2}}{dx^{n+2}} u^{n+1} \\ &= \frac{d^n}{dx^n} [2(n+1)u^{n-1}(u+2nx^2)], \\ &= 2(n+1) \frac{d^n}{dx^n} [(2n+1)u^n + 2nu^{n-1}], \\ &= 2^{n+1}(n+1)! [(2n+1)p_n(x) + p'_{n-1}(x)]. \end{aligned}$$

Thus we have deduced the relation

$$(30) \quad p'_{n+1}(x) = (2n+1)p_n(x) + p'_{n-1}(x), \quad n = 1, 2, \dots;$$

which by (29) can be rewritten for the  $P_n(x)$  as:

$$(31) \quad (n + \tfrac{3}{2})^{-\frac{1}{2}} P'_{n+1}(x) = (n - \tfrac{1}{2})^{-\frac{1}{2}} P'_{n-1}(x) \\ = (2n+1)(n+\tfrac{1}{2})^{-\frac{1}{2}} P_n(x); \quad n = 1, 2, \dots$$

Now we introduce the notation

$$(32) \quad c_k' \equiv \int_{-1}^1 f'(x) P_k(x) dx, \quad c_k'' \equiv \int_{-1}^1 f''(x) P_k(x) dx; \\ k = 0, 1, \dots,$$

and use integration by parts to deduce

$$(33) \quad c'_{k \pm 1} = \int_{-1}^1 f'(x) P_{k \pm 1}(x) dx,$$

$$= [f(x)P_{k \pm 1}(x)] \Big|_{-1}^1 - \int_{-1}^1 f(x)P'_{k \pm 1}(x) dx.$$

Let us assume for the present that  $f(\pm 1) = f'(\pm 1) = 0$ . Then (33) simplifies and with (31) and (27b) yields

$$\begin{aligned} & - (k + \frac{3}{2})^{-\frac{1}{2}} c'_{k+1} - (k - \frac{1}{2})^{-\frac{1}{2}} c'_{k-1} \\ & = (2k + 1)(k + \frac{1}{2})^{-\frac{1}{2}} \int_{-1}^1 f(x)P_k(x) dx, \\ & = (2k + 1)(k + \frac{1}{2})^{-\frac{1}{2}} c_k. \end{aligned}$$

From this, it follows that

$$(34) \quad \begin{aligned} c_k & = - A_k \frac{c'_{k+1}}{k} - B_k \frac{c'_{k-1}}{k}, \\ A_k & \equiv \left( \frac{2k+1}{2k+3} \right)^{\frac{1}{2}} \frac{k}{2k+1}, \quad B_k \equiv \left( \frac{2k+1}{2k-1} \right)^{\frac{1}{2}} \frac{k}{2k+1}. \end{aligned}$$

But since  $f'(x)$  is continuous we may use (28) for  $f'(x)$  and  $c_n'$  to conclude from (34) that, since  $c_k' \rightarrow 0$ ,  $A_k \rightarrow \frac{1}{2}$  and  $B_k \rightarrow \frac{1}{2}$ ,

$$(35) \quad \lim_{k \rightarrow \infty} k c_k = 0.$$

This argument can be repeated with the function  $w(x) \equiv f'(x)$ . By the hypothesis, it follows that  $w'(x) = f''(x)$  is continuous and in place of (35) we get (having assumed that  $w(\pm 1) = f'(\pm 1) = 0$ )

$$\lim_{k \rightarrow \infty} k c_k' = 0.$$

However, by using this result in (34), we find

$$(36) \quad \lim_{k \rightarrow \infty} k^2 c_k = 0.$$

From the property

$$|p_n(x)| \leq 1$$

exhibited in Problem 8 we deduce from (29) that for  $x \in [-1, 1]$

$$(37) \quad |P_k(x)| \leq \sqrt{k + \frac{1}{2}}.$$

By (36) we can pick any  $\epsilon > 0$  and find  $n$  sufficiently large so that  $k^2 |c_k| < \epsilon$

for all  $k \geq n$ . Then from (27) and (37) we have for any  $m > n$  and  $x \in [-1, 1]$ :

$$\begin{aligned}
 |Q_m(x) - Q_n(x)| &= \left| \sum_{k=n+1}^m \frac{1}{k^2} (k^2 c_k) P_k(x) \right|, \\
 &\leq \epsilon \sum_{k=n+1}^m \frac{1}{k^2} |P_k(x)|, \\
 &\leq \epsilon \sum_{k=n+1}^m \frac{\sqrt{k + \frac{1}{2}}}{k^2} = \epsilon \sum_{k=n+1}^m \frac{1}{k^{3/2}} \sqrt{1 + \frac{1}{2k}}, \\
 &\leq \sqrt{2} \epsilon \sum_{k=n+1}^{\infty} \frac{1}{k^{3/2}}, \\
 &\leq \sqrt{2} \epsilon \int_n^{\infty} \frac{1}{\xi^{3/2}} d\xi, \\
 (38) \quad |Q_m(x) - Q_n(x)| &\leq \frac{2\sqrt{2} \epsilon}{\sqrt{n}}.
 \end{aligned}$$

Here we have used  $k^{-3/2} < \int_{k-1}^k \xi^{-3/2} d\xi$ . It follows from (38) that the least squares polynomials  $\{Q_n(x)\}$  form a Cauchy sequence and converge uniformly on  $[-1, 1]$ .

If we call

$$\lim_{n \rightarrow \infty} Q_n(x) = g(x),$$

then  $g(x)$  is continuous on  $[-1, 1]$  since it is the uniform limit of continuous functions. Again, by the uniformity we may take the limit under the integral sign in (28a) to find, since  $f(x)$  and  $g(x)$  are continuous, that  $f(x) \equiv g(x)$ . Finally, letting  $m \rightarrow \infty$  in (38) and replacing  $\epsilon$  by  $\epsilon/(2\sqrt{2})$  we get the result stated in the theorem.

To complete the proof we must eliminate the requirement that  $f(\pm 1) = f'(\pm 1) = 0$ . To do this, we construct, for any  $f(x)$ , the Hermite interpolation polynomial,  $h_3(x)$ , for which

$$h_3(\pm 1) = f(\pm 1), \quad h_3'(\pm 1) = f'(\pm 1).$$

Then  $g(x) \equiv f(x) - h_3(x)$  satisfies all the requirements of the theorem. However, since  $h_3(x)$  has degree at most 3, it follows from (27b) and the orthogonality of the polynomials  $P_n(x)$ , that the  $c_j$  are unchanged for  $j \geq 4$  if  $f(x)$  is replaced by  $g(x)$ . ■

By using the technique of the above proof, we find

**THEOREM 8.** Let  $f(x)$  have a continuous  $r$ th derivative on  $[-1, 1]$  where  $r \geq 2$ . Then with  $Q_n(x)$  and  $c_n$  defined in (27):  $\lim_{k \rightarrow \infty} k^r c_k = 0$  and

$$|f(x) - Q_n(x)| = \mathcal{O}(n^{-r+\frac{3}{2}}), \quad \text{for all } x \in [-1, 1].$$

*Proof.* See Problem 12. ■

A simple linear change of variable yields the

**COROLLARY.** If  $f(x)$  has  $r \geq 2$  continuous derivatives in  $[a, b]$  then there exists a polynomial approximation,  $q_n(x)$ , of degree at most  $n$  such that  $|f(x) - q_n(x)| = \mathcal{O}(n^{-r+\frac{3}{2}})$ , for all  $x \in [a, b]$ .

*Proof.* See Problem 13. ■

Analogous results can be obtained for various weighted least squares approximations. If the weight function is  $w(x)$  and the interval is  $[a, b]$ , then the approximation to  $f(x)$  is

$$Q_n(x) = \sum_{j=0}^n c_j P_j(x)$$

where the  $c_j$  are defined in (21) and the orthonormal polynomials  $P_n(x)$  satisfy (20). A proof of the pointwise convergence of  $Q_n(x)$  to a sufficiently smooth  $f(x)$  can be given if

(i) the  $P_n(x)$  are the *eigenfunctions* of a *regular second order* differential operator, say

$$\mathcal{L}[P_n(x)] \equiv a(x)P_n''(x) + b(x)P_n'(x) = \lambda_n P_n(x),$$

whose *eigenvalues*,  $\lambda_n$ , satisfy

$$\lim_{n \rightarrow \infty} \lambda_n n^{-2} = \text{const.};$$

(ii) the  $P_n(x)$  are bounded by

$$|P_n(x)| = \mathcal{O}(n^{\frac{1}{2}}) \quad \text{for all } x \in [a, b].$$

In particular, we shall sketch the proof for the case in which the  $P_n(x)$  are related to the *Chebyshev polynomials*. These polynomials are orthogonal over  $[-1, 1]$  with respect to the weight

$$(39) \quad w(x) \equiv \frac{1}{\sqrt{1 - x^2}}.$$

In Problem 9 they are determined as (see Section 5):

$$(40) \quad P_n(x) = \sqrt{\frac{2}{\pi}} \cos(n \cos^{-1} x), \quad n = 1, 2, \dots$$

and they are solutions of

$$(41) \quad (1 - x^2)P_n''(x) - xP_n'(x) = -n^2P_n(x).$$

Thus in this case  $\lambda_n = n^2$ . We require  $f(x)$  to have a continuous second derivative on  $[-1, 1]$ . By the remarks at the end of the proof of Theorem 7, we may assume without loss of generality that  $f(\pm 1) = f'(\pm 1) = 0$  in the present proof. [Since  $(P_i(x), P_j(x)) = \delta_{ij}$ , the  $c_n$  of (21) are unchanged if  $f(x)$  is replaced by  $f(x) - h_3(x)$ , see last paragraph of Theorem 7.]

From (39) and (41) in (21) with  $[a, b] = [-1, 1]$  we have

$$c_n = -\frac{1}{n^2} \int_{-1}^1 \frac{f(x)}{\sqrt{1-x^2}} [(1-x^2)P_n''(x) - xP_n'(x)] dx.$$

Integrating by parts, all derivatives can be removed from  $P_n(x)$  to get

$$(42a) \quad c_n = -\frac{1}{n^2} \int_{-1}^1 [\alpha(x)f(x) + \beta(x)f'(x) + \gamma(x)f''(x)]P_n(x)w(x) dx,$$

where

$$(42b) \quad \alpha(x) \equiv -\frac{1}{1-x^2}, \quad \beta(x) \equiv -3x, \quad \gamma(x) \equiv 1-x^2.$$

Since  $f(\pm 1) = f'(\pm 1) = 0$  and  $f''(x)$  is continuous, we note that  $\alpha(x)f(x)$ ,  $\beta(x)f'(x)$  and  $\gamma(x)f''(x)$  are continuous on  $[-1, 1]$ . Thus the coefficients in the expansion of the sum  $[\alpha f + \beta f' + \gamma f'']$  tend to zero as  $n \rightarrow \infty$  (by the analog of Theorem 3 for weighted polynomials). Using this fact in (42a) implies

$$(43) \quad \lim_{n \rightarrow \infty} n^2 c_n = 0.$$

A sharper bound than that in (ii) is easily obtained for the Chebyshev polynomials. Clearly, from the representation (40),

$$(44) \quad |P_n(x)| \leq \sqrt{\frac{2}{\pi}}.$$

From (43) and (44) we find, as in the proof of Theorem 7, that the  $Q_n(x)$  converge uniformly. However, by using this fact and the mean convergence we easily find that

$$(45) \quad |f(x) - Q_n(x)| = \mathcal{O}\left(\frac{1}{n}\right) \quad \text{for all } x \in [-1, 1].$$

Note that the error estimate here is smaller by  $\mathcal{O}(1/\sqrt{n})$  than that for the Legendre polynomial expansion deduced in Theorem 7.

This argument is easily extended to give

**THEOREM 9.** Let  $f(x)$  have a continuous  $r$ th derivative on  $[-1, 1]$  where  $r \geq 2$ . Then the mean square Chebyshev approximations,  $Q_n(x)$ , with coefficients  $c_i$  defined by (39) and (40) in (21) satisfy

$$|f(x) - Q_n(x)| = \mathcal{O}(n^{1-r}), \quad \text{for all } x \in [-1, 1]. \quad \blacksquare$$

### 3.5. Discrete Least Squares Approximation

For a fixed set  $(x_0, x_1, \dots, x_M)$  of distinct points we might seek to minimize

$$(46) \quad \|f(x) - Q_n(x)\|_D = \sqrt{\sum_{i=0}^M [f(x_i) - Q_n(x_i)]^2}$$

over all polynomials of degree at most  $n$ . Here  $M$  is usually much larger than the degree  $n$  of the class of approximating polynomials. The fact that  $\|\cdot\|_D$  is a norm and essentially strict, is easily shown in Problems 15 and 16. It then follows as in Theorem 0.1, that a minimizing polynomial exists. Further, the minimizing  $Q_n(x)$  is uniquely determined if  $n \leq M$  (see Problem 17).

To actually determine the discrete least squares approximation, we again use the notation

$$(47) \quad \phi(a_0, \dots, a_n) \equiv \|f - Q_n(x)\|_D^2,$$

where

$$Q_n(x) = a_0 + a_1 x + \dots + a_n x^n.$$

This function  $\phi$  is quadratic in the  $a_i$  since

$$(48) \quad \begin{aligned} \phi(a_0, a_1, \dots, a_n) &= \sum_{i=0}^M f^2(x_i) - 2 \sum_{k=0}^n a_k \sum_{i=0}^M x_i^k f(x_i) \\ &\quad + \sum_{k=0}^n \sum_{j=0}^n a_k a_j \sum_{i=0}^M x_i^{k+j}. \end{aligned}$$

At a minimum,  $\{\hat{a}_j\}$  of  $\phi(\mathbf{a})$  we have the necessary conditions

$$\left. \frac{\partial \phi}{\partial a_k} \right|_{\mathbf{a}=\hat{\mathbf{a}}} = 0, \quad k = 0, 1, \dots, n,$$

which yield the *normal system*

$$(49) \quad \sum_{j=0}^n a_j \sum_{i=0}^M x_i^{k+j} = \sum_{i=0}^M x_i^k f(x_i); \quad k = 0, 1, \dots, n.$$

If  $n \leq M$ , the right-hand side of (49) may take on any preassigned values  $(c_0, c_1, \dots, c_n)$  by suitably picking  $[f(x_0), f(x_1), \dots, f(x_M)]$  (e.g., if  $M = n$ , the Vandermonde determinant  $|x_i^k|$  is non-singular and if  $M > n$ , there are fewer restrictions than variables  $f(x_j)$  to determine). Hence, the system (49) is solvable for any right-hand side and therefore non-singular, if  $M \geq n$ .

But, we may avoid the necessity of having to solve the general normal system (49) if, in analogy with (10), we can construct a sequence of polynomials  $P_n(x)$ ,  $n = 0, 1, \dots, M$ , which is orthonormal relative to summation over  $x_i$ ,  $i = 0, 1, \dots, M$ . To this end, we define an *inner product* by

$$(50) \quad (f, g) = (g, f) \equiv \sum_{i=0}^M f(x_i)g(x_i).$$

With this inner product in the Gram-Schmidt process of Subsection 3.1, we can orthonormalize the independent set  $\{x^k\}$  for  $k = 0, 1, 2, \dots, M$ . The result is a set of polynomials  $\{P_k(x)\}$  for which

$$(51) \quad (P_r(x), P_s(x)) = \sum_{i=0}^M P_r(x_i)P_s(x_i) = \delta_{rs}; \quad r, s = 0, 1, \dots, M.$$

Now the unique polynomial  $Q_n(x)$  of degree at most  $n \leq M$  that minimizes (46) can be written as (see Problem 14)

$$(52a) \quad Q_n(x) = \sum_{k=0}^n d_k P_k(x),$$

where

$$(52b) \quad d_k = \sum_{j=0}^M f(x_j)P_k(x_j).$$

We now consider the determination of polynomials which satisfy (51) over various sets of points. As indicated, the Gram-Schmidt procedure could be used, but for the special cases to be treated it is not required. First, we consider uniformly spaced points  $\{x_j\}$  in  $[-1, 1]$ ; say,

$$x_0 = -1, \quad x_j = x_0 + jh, \quad h = \frac{2}{M}; \quad j = 0, 1, \dots$$

The corresponding orthonormal polynomials that satisfy (51) can be written as

$$(53) \quad P_n(x) = C_n \Delta^n (u^{[n]}(x)v^{[n]}(x)), \quad n = 0, 1, \dots, M.$$

Here the forward difference operators  $\Delta^n$  are defined by:

$$(54) \quad \begin{aligned} \Delta^0 f(x) &\equiv f(x) \\ \Delta f(x) &\equiv f(x + h) - f(x), \\ \Delta^2 f(x) &\equiv \Delta(\Delta f(x)) = f(x + 2h) - 2f(x + h) + f(x), \\ &\vdots \\ \Delta^n f(x) &\equiv \Delta[\Delta^{n-1} f(x)] = \sum_{j=0}^n (-1)^j \binom{n}{j} f(x + (n-j)h); \end{aligned}$$

the coefficients  $C_n$  are constants given in (58) and

$$(55) \quad \begin{aligned} u^{[0]}(x) &= v^{[0]}(x) = 1 \\ u^{[n]}(x) &= (x - x_0)(x - x_1) \cdots (x - x_{n-1}) \\ v^{[n]}(x) &= (x - x_{M+1})(x - x_{M+2}) \cdots (x - x_{M+n}). \end{aligned}$$

Formula (53) is the discrete analog of the formula (26) for the Legendre polynomials.

The fact that these polynomials satisfy (51) may be verified by the use of a formula for *summation by parts*, which is analogous to integration by parts. To derive this formula, we note the identity

$$(56a) \quad \Delta(FG) = F\Delta G + G\Delta F + (\Delta F)(\Delta G)$$

which can be written as

$$(56b) \quad G\Delta F = \Delta(FG) - F\Delta G - (\Delta F)(\Delta G).$$

Now assume  $r > s$  and let

$$F(x) \equiv \Delta^{r-1}(u^{[r]}(x)v^{[r]}(x)), \quad G(x) \equiv \Delta^s(u^{[s]}(x)v^{[s]}(x)).$$

We evaluate (56b) at each point  $x_i = x_0 + ih$  and sum over  $0 \leq i \leq M$  to get

$$(57) \quad \begin{aligned} \frac{1}{C_s C_r} \sum_{i=0}^M P_s(x_i) P_r(x_i) &= \sum_{i=0}^M G(x_i) \Delta F(x_i) \\ &= \sum_{i=0}^M \Delta[F(x_i)G(x_i)] - \sum_{i=0}^M F(x_i) \Delta G(x_i) \\ &\quad - \sum_{i=0}^M [\Delta F(x_i)][\Delta G(x_i)], \\ &= F(x)G(x) \Big|_{x_0}^{x_{M+1}} - \sum_{i=0}^M F(x_i) \Delta G(x_i) \\ &\quad - \sum_{i=0}^M [\Delta F(x_i)][\Delta G(x_i)]. \end{aligned}$$

We observe that  $F(x_0) = F(x_{M+1}) = 0$ , whence from (53) and (57) we have converted the sum in (51) into two sums in which  $\Delta G$  appears.

We may continue with this process of summation by parts to successively form sums in which higher order differences of  $G$  appear. The boundary terms vanish since, by Problem 21 and the identity (56a) for  $x = x_0$  or  $x = x_{M+1}$

$$\Delta^k(u^{[r]}(x)v^{[r]}(x)) = 0, \quad k = 0, 1, \dots, r - 1.$$

Now, from the assumption  $r > s$ , it follows that after  $s + 1$  such applications, the term  $\Delta^{2s+1}(u^{[s]}v^{[s]})$  will be a factor in all the resulting sums. But  $u^{[s]}(x)v^{[s]}(x) = q_{2s}(x)$  is a polynomial of degree  $2s$  and hence all the sums vanish identically since for any polynomial  $p_n(x)$  of degree at most  $n$ , the difference operator reduces the degree, i.e.,

$$\begin{aligned} \Delta p_n(x) &\equiv p_{n-1}(x) \\ &\vdots \\ \Delta^m p_n(x) &\equiv 0, \quad \text{for } m > n. \end{aligned}$$

The verification that (51) is valid for  $r = s$  follows when we define

$$(58) \quad C_n = \left[ \sum_{i=0}^M \{\Delta^n[u^{[n]}(x_i)v^{[n]}(x_i)]\}^2 \right]^{-\frac{1}{2}}.$$

The bracket in (58) is not zero if we can show that

$$\Delta^n[u^{[n]}(x_0)v^{[n]}(x_0)] \neq 0.$$

But the polynomial

$$p_{2n}(x) \equiv u^{[n]}(x)v^{[n]}(x)$$

has only the  $2n$  zeros  $(x_0, x_1, \dots, x_{n-1}; x_{M+1}, x_{M+2}, \dots, x_{M+n})$  and  $M \geq n$ . Then from (54) with  $f(x) \equiv p_{2n}(x)$ , only one term in the expression for  $\Delta^n f(x_0)$  is non-zero, i.e.,

$$\Delta^n p_{2n}(x_0) = p_{2n}(x_n) \neq 0.$$

Hence the definition (58) is valid.

The polynomials  $P_n(x)$  of (53) have been called the *Gram polynomials*.

It can be shown that the polynomials  $P_n(x)/\sqrt{2/M}$  converge as  $M \rightarrow \infty$  to the orthonormal polynomials defined in equation (26) (they are related to the Legendre polynomials).

Another interesting set of points for discrete least squares approximation in  $[-1, 1]$  are the zeros,  $(x_0, x_1, \dots, x_M)$  of the  $(M + 1)$ -st Chebyshev polynomial

$$(59) \quad T_{M+1}(x) = 2^{-M} \cos [(M + 1) \cos^{-1} x], \quad M = 0, 1, 2, \dots$$

In Subsection 4.2 we show that these are polynomials of the indicated degrees. Now the points  $\{x_j\}$  are not uniformly spaced, but are given by

$$(60) \quad x_j = \cos \left[ \frac{(2j+1)\pi}{(M+1)2} \right], \quad j = 0, 1, \dots, M.$$

Note that the corresponding points

$$\theta_j = \cos^{-1} x_j = \frac{2j+1}{M+1} \frac{\pi}{2}$$

are uniformly spaced in  $[0, \pi]$ .

We say that these sets  $\{x_j\}$  are interesting, because on the one hand, the discretely orthonormal polynomials are easily found in Theorem 10; and on the other hand, we prove in Subsection 5.1 that various approximation polynomials based on these points converge uniformly to any function  $f(x)$  with two continuous derivatives in  $[-1, 1]$ .

**THEOREM 10.** *For the discrete set of points  $\{x_j\}$  defined in (60), the discretely orthonormal polynomials satisfying (51) are proportional to the Chebyshev polynomials. Specifically they are:*

$$(61) \quad \begin{aligned} P_0(x) &\equiv (M+1)^{-\frac{1}{2}}, \\ P_n(x) &\equiv 2^{\frac{1}{2}}(M+1)^{-\frac{1}{2}} \cos(n \cos^{-1} x), \quad n = 1, 2, \dots, M. \end{aligned}$$

*Proof.* We must verify that  $P_n(x)$  defined in (61) satisfies (51). This follows directly from the discrete orthonormality property of the trigonometric functions expressed in

**LEMMA 1.**

$$(62) \quad \sum_{j=0}^M \cos r\theta_j \cos s\theta_j = \begin{cases} 0 & \text{for } 0 \leq r \neq s \leq M; \\ \frac{M+1}{2} & \text{for } 0 < r = s \leq M; \\ M+1 & \text{for } 0 = r = s; \end{cases}$$

where

$$(63) \quad \theta_j = \frac{2j+1}{M+1} \frac{\pi}{2}, \quad j = 0, 1, \dots, M.$$

We can most readily evaluate the sum in (62) by making use of the well-known formula,

$$(64) \quad e^{ix} \equiv \cos x + i \sin x,$$

where  $i^2 = -1$  and  $x$  is a real number. Then we may write

$$(65) \quad \begin{aligned} \sum_{j=0}^M \cos r\theta_j \cos s\theta_j &= \frac{1}{2} \sum_{j=0}^M [\cos(r+s)\theta_j + \cos(r-s)\theta_j] \\ &= \frac{1}{2} \operatorname{Re} \left( \sum_{j=0}^M e^{i(r+s)\theta_j} \right) + \frac{1}{2} \operatorname{Re} \left( \sum_{j=0}^M e^{i(r-s)\theta_j} \right). \end{aligned}$$

But the right-hand sums may be treated as geometric series. That is, we note that  $\theta_j - \theta_0 = j2\theta_0$ , whence for  $R = 1, 2, \dots, 2M$ ,

$$\begin{aligned} \sum_{j=0}^M e^{iR\theta_j} &= e^{iR\theta_0} \sum_{j=0}^M (e^{iR2\theta_0})^j \\ &= e^{iR\theta_0} \left( \frac{1 - e^{iR2\theta_0(M+1)}}{1 - e^{iR2\theta_0}} \right) \\ &= e^{iR\theta_0} \left( \frac{e^{-iR\theta_0(M+1)} - e^{iR\theta_0(M+1)}}{e^{-iR\theta_0} - e^{iR\theta_0}} \right) \frac{e^{iR\theta_0(M+1)}}{e^{iR\theta_0}} \\ &= \frac{\sin R(M+1)\theta_0}{\sin R\theta_0} e^{iR\theta_0(M+1)}. \end{aligned}$$

By taking the real part, we find

$$\begin{aligned} (66) \quad \operatorname{Re} \left( \sum_{j=0}^M e^{iR\theta_j} \right) &= \frac{\sin R(M+1)\theta_0 \cos R(M+1)\theta_0}{\sin R\theta_0}, \\ &= \frac{\sin 2R(M+1)\theta_0}{2 \sin R\theta_0}, \\ &= \frac{\sin R\pi}{2 \sin \frac{R\pi}{2(M+1)}} = 0, \quad \text{for } R = 1, 2, \dots, 2M. \end{aligned}$$

If we now identify  $R = r \pm s$ , then from (65) with  $r > s$  the first part of (62) follows. The special case  $r = s > 0$ , of (62) results by using (65) and observing that the sum in (66), for  $R = 0$  is simply

$$\operatorname{Re} \sum_{j=0}^M e^{i0\theta_j} = M + 1.$$

Finally, the trivial case  $r = s = 0$ , of (62) is directly verifiable. Thus, Lemma 1 is proven and from it follows Theorem 10. ■

We note the fact that for any set of  $M + 1$  distinct points  $(x_0, x_1, \dots, x_M)$ ,

**THEOREM 11.** *The discrete least squares approximation polynomial  $Q_M(x)$  of degree at most  $M$  which minimizes*

$$\|f(x) - Q_M(x)\|_D^2 = \sum_{i=0}^M [f(x_i) - Q_M(x_i)]^2,$$

*is the interpolation polynomial for  $f(x)$  based on the distinct points  $(x_0, x_1, \dots, x_M)$ .*

*Proof.* Let  $P_M(x)$  be the indicated interpolation polynomial. Then  $\|f(x) - P_M(x)\|_D = 0$  and since the interpolation polynomial is unique we must have  $P_M(x) \equiv Q_M(x)$ . ■

We recall that with least squares approximations (discrete or not) the next higher degree approximation is obtained by simply adding a new term to the previous approximation. Theorem 11 then shows that for the discrete case, as the degree increases for a fixed set of points, the approximations “approach” the interpolation polynomial. However, for the  $n + 1$  unequally spaced points (60) we show in Subsection 5.1 that while the  $n$ th degree interpolation polynomials converge like  $1/\sqrt{n}$  as  $n \rightarrow \infty$  (for a sufficiently smooth function), so do the discrete least square polynomials of degrees  $\geq \sqrt{n}$ .

We observe that one reason for working with the discrete least square method is that sums are readily computable. On the other hand, the integrals of the continuous least square method, say of the form

$$\int_a^b f(x)P_n(x) dx,$$

are generally only determined approximately (frequently by using quadrature formulae, i.e., sums).

The natural extension to *weighted* discrete least squares approximation is omitted. An important application of these approximation methods is to the art of fitting mathematical formulae to empirical data but we shall not treat that here.

### PROBLEMS, SECTION 3

**1.\*** A generalization of the Hilbert segments is furnished by the matrix  $A = (a_{ij})$ ,  $a_{ij} = 1/(\alpha_i + \beta_j)$ ,  $i, j = 1, 2, \dots, n$ , where the  $\alpha_i$  are distinct and the  $\beta_j$  are distinct. Show that the determinant of  $A$  is

$$\det A = \det \left| \frac{1}{\alpha_i + \beta_j} \right| = \frac{\prod_{p=1}^{n-1} \left[ \prod_{q=p+1}^n (\alpha_p - \alpha_q) \right] \cdot \prod_{r=1}^{n-1} \left[ \prod_{s=r+1}^n (\beta_r - \beta_s) \right]}{\prod_{i=1}^n \left[ \prod_{j=1}^n (\alpha_i + \beta_j) \right]},$$

for  $n \geq 2$ .

[Hint: Multiply the  $i$ th row of  $A$  by  $\prod_{j=1}^n (\alpha_i + \beta_j)$  for  $i = 1, 2, \dots, n$  and call the resulting matrix  $C$ . The elements of  $C$  are polynomials in  $\{\alpha_i, \beta_j\}$ , hence,  $\det C$  is a polynomial  $P(\{\alpha_i\}, \{\beta_j\})$  of degree at most  $n(n - 1)$ . Observe that  $P$  is divisible by each of the factors in the numerator of the right-hand side because a determinant vanishes if two columns or two rows are identical. Hence,  $P$  equals the numerator to within a constant factor, since the numerator has degree  $n(n - 1)$ . Therefore,  $\det A$  equals the right-hand side to within a constant factor  $K_n$ . Determine  $K_n$  by induction.]

**2.\*** Notice that the cofactor of any element in the above matrix,  $A$ , is the determinant of a matrix of similar form. Use the cofactor and the determinant of  $A$  to represent the elements of  $A^{-1} \equiv (b_{jk})$ . Express these elements in terms

of the Lagrange interpolation coefficients with respect to the points  $\alpha_i$  and  $\beta_j$ ; the result should be

$$b_{jk} = (\alpha_k + \beta_j) A_k(-\beta_j) B_j(-\alpha_k),$$

where

$$A_k(x) = \prod_{s \neq k} \left( \frac{\alpha_s - x}{\alpha_s - \alpha_k} \right),$$

$$B_k(x) = \prod_{s \neq k} \left( \frac{\beta_s - x}{\beta_s - \beta_k} \right).$$

Verify, by using equation (2.12) that  $AA^{-1} = A^{-1}A = I$ .

3. Show that any polynomial of degree  $m$  which is orthogonal to the first  $m + 1$  orthogonal polynomials (i.e., to all orthogonal polynomials of degree  $m$  or less) is the identically vanishing polynomial.

4. Verify that if  $f(x)$  is continuous in  $[a, b]$ ,  $(x_0, x_1, \dots, x_n)$  are distinct points in  $(a, b)$  and

$$w_M(x) \equiv \begin{cases} \frac{(x - x_j + \epsilon_M)}{\epsilon_M^2} & \text{for } x_j - \epsilon_M \leq x \leq x_j, \\ \frac{-(x - x_j - \epsilon_M)}{\epsilon_M^2} & \text{for } x_j \leq x \leq x_j + \epsilon_M, \quad j = 0, 1, \dots, n, \\ 0 & \text{if } x \text{ is not in any of the above intervals,} \end{cases}$$

where

$$\epsilon_M = \min_{i, j} \frac{|x_i - x_j|}{M},$$

then the associated weighted least squares polynomial approximations  $P_{n, M}(x)$  converge to the Lagrange interpolation polynomial as  $M \rightarrow \infty$ .

5. Given the linearly independent set of functions  $\{g_i(x)\}$  for  $i = 0, 1, 2, \dots, n$ , verify that with the definition (22), the Gram-Schmidt orthogonalization process (15)–(16) produces an orthonormal set  $\{f_i(x)\}$ .

6. If  $w(x) \equiv 1$ ,  $[a, b] \equiv [-1, 1]$ , show that for  $g_k(x) \equiv x^k$ ,  $k = 0, 1, 2, \dots$ , the orthonormal polynomials  $P_n(x)$  resulting from (14)–(16) are

$$P_n(x) \equiv (n + \frac{1}{2})^{\frac{1}{2}} \frac{1}{n! 2^n} \frac{d^n}{dx^n} (x^2 - 1)^n.$$

[Hint: Verify

$$\int_{-1}^1 P_n(x) P_m(x) dx = \delta_{nm}$$

by integration by parts and use uniqueness of Gram-Schmidt process.]

[The polynomials  $p_n(x) \equiv P_n(x)(n + \frac{1}{2})^{-\frac{1}{2}}$  are called the *Legendre polynomials*, and have the properties

$$p_n(1) = 1; \quad p_n(-1) = (-1)^n;$$

(recurrence relation)

$$p_{n+1}(x) = \frac{2n+1}{n+1} x p_n(x) - \frac{n}{n+1} p_{n-1}(x), \quad n \geq 1.$$

7. Verify that the Legendre polynomials also satisfy (*differential equation*)

$$(1 - x^2)p_n'' - 2xp_n' + n(n + 1)p_n = 0.$$

[Hint: Let  $u = (x^2 - 1)^n$  and apply Leibnitz' rule to get  $d^{n+1}/dx^{n+1}$  of both sides of

$$(x^2 - 1)u'(x) = 2nxu.]$$

8. Prove that for the Legendre polynomials

$$|p_n(x)| \leq 1 \quad \text{for } |x| \leq 1.$$

[Hint: Consider

$$n(n + 1)f(x) \equiv n(n + 1)p_n^2(x) + (1 - x^2)[p_n'(x)]^2.$$

Note that  $f(x) = p_n^2(x)$  if  $p_n'(x) = 0$ , or  $x^2 = 1$ . But, by using the differential equation of Problem 7,  $n(n + 1)f'(x) = 2x[p_n'(x)]^2 \geq 0$  if  $x \geq 0$  and hence the value of  $|p_n(x)|$  at a local maximum point,  $|x| < 1$  is  $\leq 1$ .]

9. If  $w(x) \equiv (1 - x^2)^{-\frac{1}{2}}$ ,  $[a, b] \equiv [-1, 1]$ ,  $g_k(x) \equiv x^k$  for  $k = 0, 1, 2, \dots$ , show that the sequence defined by Problem 5 is

$$P_n(x) \equiv \sqrt{\frac{2}{\pi}} \cos(n \cos^{-1} x), \quad n = 1, 2, \dots,$$

$$P_0(x) \equiv \frac{1}{\sqrt{\pi}}.$$

[The polynomials

$$T_n(x) \equiv \frac{1}{2^{n-1}} \cos(n \cos^{-1} x), \quad n = 1, 2, \dots,$$

$$T_0(x) \equiv 2$$

are called *Chebyshev polynomials* of the first kind. They satisfy (*recurrence relation*):

$$T_{n+1}(x) = xT_n(x) - \frac{1}{2}T_{n-1}(x);$$

(*differential equation*):

$$(1 - x^2)T_n'' = xT_n' - n^2T_n.]$$

10. Show that  $\|\cdot\|_2$  is a strict norm [see (2)].

[Hint: (a) The triangle inequality follows from the *Schwarz inequality*:

$$\int_a^b f(x)g(x) dx \leq FG$$

where

$$F = \left\{ \int_a^b [f(x)]^2 dx \right\}^{\frac{1}{2}}, \quad G = \left\{ \int_a^b [g(x)]^2 dx \right\}^{\frac{1}{2}}.$$

Observe

$$0 \leq \int_a^b [\alpha f(x) + \beta g(x)]^2 dx \equiv (\alpha F + \beta G)^2 + 2\alpha\beta \left[ \int_a^b f(x)g(x) dx - FG \right].$$

For  $F \neq 0 \neq G$ ,  $\alpha F + \beta G = 0$  implies  $\alpha\beta < 0$  and the inequality follows.

(b) Now to show strictness, note  $\|f + g\| = \|f\| + \|g\|$  implies  $\int_a^b f(x)g(x) dx = FG$ , whence with non-trivial  $\alpha, \beta$ ,

$$0 = (\alpha F + \beta G)^2 \equiv \int_a^b [\alpha f(x) + \beta g(x)]^2 dx.]$$

**11.** Given  $(c_0, c_1, \dots, c_n)$  find a polynomial  $W_n(x)$  of degree at most  $n$ , such that

$$\int_a^b x^k W_n(x) dx = c_k \quad \text{for } k = 0, 1, \dots, n.$$

[Hint: Use the orthonormal polynomials  $P_j(x)$ ,  $j = 0, 1, \dots, n$ , defined by (15), (16) and (17).]

**12.\*** Prove Theorem 8.

**13.\*** Prove Corollary to Theorem 8.

**14.** Given the discrete orthonormal polynomials  $\{P_n(x)\}$ , for  $0 \leq n \leq M$ , on the set  $\{x_j\}$  for  $0 \leq j \leq M$  [i.e.,  $\sum_{j=0}^M P_r(x_j)P_s(x_j) = \delta_{rs}$ ] verify, if  $n \leq M$ , that

$$Q_n(x) = \sum_{k=0}^n d_k P_k(x) \quad \text{with } d_k = \sum_{j=0}^M f(x_j)P_k(x_j)$$

is the unique polynomial of degree at most  $n$  which minimizes

$$\sum_{j=0}^M |f(x_j) - Q_n(x_j)|^2.$$

[Hint: Show that

$$\sum_{j=0}^M |f(x_j) - W_n(x_j)|^2 = \sum_{j=0}^M |f(x_j) - Q_n(x_j)|^2 + \sum_{k=0}^n e_k^2,$$

where

$$W_n(x) = Q_n(x) + \sum_{k=0}^n e_k P_k(x).$$

**15.** Show that  $\|\cdot\|_D$  is a semi-norm.

[Hint:  $\|f + g\|_D \leq \|f\|_D + \|g\|_D$  is a consequence of the Cauchy-Schwarz inequality (see Chapter 1, Section 4)]

$$\sum_{i=0}^M f_i g_i \leq \left\{ \sum_{i=0}^M (f_i)^2 \right\}^{1/2} \left\{ \sum_{i=0}^M (g_i)^2 \right\}^{1/2}.$$

**16.**  $\|\cdot\|_D$  is *essentially strict* if  $\|f + g\|_D = \|f\|_D + \|g\|_D$  implies there exist non-trivial  $\alpha, \beta$  such that  $\alpha f_i + \beta g_i = 0$  for  $i = 0, 1, \dots, M$ . Prove  $\|\cdot\|_D$  is essentially strict.

**17.** If  $n \leq M$ , show that the polynomial  $Q_n(x)$  which minimizes  $\|f(x) - Q_n(x)\|_D$  is unique.

**18.** Verify that the orthonormal polynomials  $\{P_n(x)\}$  for weight  $w(x)$  and interval  $[a, b]$  may be represented in the form

$$P_n(x) = c_n \frac{1}{w(x)} \frac{d^n}{dx^n} [v_n(x)],$$

with  $c_n$  a normalization constant, in the common classical cases:

- (a)  $[a, b] \equiv [-1, 1]$ ,  $w(x) \equiv (1 - x)^\alpha(1 + x)^\beta$  with  $\alpha > -1$ ,  $\beta > -1$  [i.e.,  $v_n(x) \equiv (1 - x)^{\alpha+n}(1 + x)^{\beta+n}$ ] (*Jacobi polynomials*)  
 (b)  $[a, b] \equiv [0, \infty]$ ,  $w(x) \equiv e^{-\alpha x}$  with  $\alpha > 0$  [i.e.,  $v_n(x) \equiv x^n e^{-\alpha x}$ ] (*Laguerre polynomials*)  
 (c)  $[a, b] \equiv [-\infty, \infty]$ ,  $w(x) \equiv e^{-\alpha^2 x^2}$  [i.e.,  $v_n(x) \equiv e^{-\alpha^2 x^2}$ ] (*Hermite polynomials*)

[Hint:

$$P_n(x) \equiv \frac{1}{w(x)} \frac{d^n}{dx^n} [v_n(x)]$$

is a polynomial of at most degree  $n$ , if  $v_n(x)$  satisfies

$$\frac{d^{n+1}}{dx^{n+1}} \left[ \frac{1}{w(x)} \frac{d^n}{dx^n} (v_n(x)) \right] = 0;$$

furthermore,  $\int_a^b P_n(x)P_m(x)w(x) dx = 0$  for  $n > m$  is implied with the use of integration by parts from

$$\begin{aligned} \int_a^b \frac{d^n}{dx^n} [v_n(x)]P_m(x) dx &= - \int_a^b \frac{d^{n-1}}{dx^{n-1}} [v_n(x)]P_m'(x) dx \\ &\vdots \\ &= (-1)^{m+1} \int_a^b \frac{d^{n-m-1}}{dx^{n-m-1}} [v_n(x)]P_m^{(m+1)}(x) dx \end{aligned}$$

if

$$\left. \frac{d^r}{dx^r} [v_n(x)] \right|_{x=a, b} = 0 \quad \text{for } r = 0, 1, \dots, n-1.$$

**19.** By the use of Problem 18 find another representation for the Chebyshev polynomials.

**20.\*** Prove Theorem 9 for  $r > 2$ .

**21.** Verify that with the definitions (54) and (55),

$$\Delta u^{[n]}(x) = nh u^{[n-1]}(x), \quad n \geq 1.$$

#### 4. POLYNOMIALS OF "BEST" APPROXIMATION

Another measure of the deviation between a function,  $f(x)$ , and an approximating polynomial of degree  $n$ ,

$$(1) \quad P_n(x) = a_0 + a_1 x + \cdots + a_n x^n,$$

is the so-called *maximum norm*:

$$(2) \quad \|f(x) - P_n(x)\|_\infty \equiv \max_{a \leq x \leq b} |f(x) - P_n(x)| \equiv D(f, P_n).$$

A polynomial which minimizes this norm is conventionally called a polynomial of "best" approximation.

Equation (2) defines a function of the  $n + 1$  coefficients  $\{a_i\}$  that is not as explicit as (3.4),

$$(3) \quad d(a_0, a_1, \dots, a_n) \equiv \max_{a \leq x \leq b} |f(x) - P_n(x)|.$$

A polynomial of best approximation is characterized by a point  $\hat{\mathbf{a}}$  in  $(n + 1)$ -space at which  $d(\mathbf{a})$  is a minimum. The existence of such a polynomial is shown by

**THEOREM 1.** *Let  $f(x)$  be a given function continuous in  $[a, b]$ . Then for any integer  $n$  there exists a polynomial  $\hat{P}_n(x)$ , of degree at most  $n$ , that minimizes  $\|f(x) - P_n(x)\|_\infty$ .*

*Proof.* We shall verify the hypotheses of Theorem 0.1 to obtain existence of a minimizing polynomial  $\hat{P}_n(x) \equiv \sum_{i=0}^n \hat{a}_i x^i$ .

Clearly,  $\|\cdot\|_\infty$  is a norm in the space of continuous functions on  $[a, b]$ . We only need to establish (0.5) for this norm. That is, we must show that on the subset of polynomials  $\{P_n(x)\}$  such that

$$\sum_{j=0}^n a_j^2 = 1, \min \left\| P_n(x) \right\|_\infty \equiv m_n > 0.$$

By the argument in the proof of Theorem 0.1,  $\|P_n(x)\|_\infty$  is a continuous function of the variables  $\{a_i\}$ . If  $\Sigma$  is the closed bounded set  $\sum_{j=0}^n a_j^2 = 1$ , we may apply the Weierstrass theorem which assures us that there is a point  $\{\tilde{a}_j\}$  for which

$$m_n = \min_{\Sigma} \|P_n(x)\|_\infty,$$

is attained. But at this point

$$m_n = \left\| \sum_{j=0}^n \tilde{a}_j x^j \right\|_\infty \neq 0,$$

since

$$\sum_{j=0}^n \tilde{a}_j^2 = 1$$

and any non-trivial polynomial, of degree at most  $n$  can have at most  $n$  zeros (i.e., if  $\|\tilde{P}_n(x)\|_\infty = 0$  then  $\tilde{P}_n(x) \equiv 0$ ). ■

At this point we observe that  $\|\cdot\|_\infty$  is not a strict norm, and hence we cannot use Theorem 0.2 to establish uniqueness of  $\hat{P}_n(x)$ . Nevertheless, the “best” approximation polynomial is unique and we will prove this fact

in Theorem 3. It is of interest to note that Theorem 1 remains true if the norm  $\|\cdot\|_\infty$  of (2) is replaced by the semi-norm  $\|\cdot\|_{\infty,S}$  defined as

$$\|f(x)\|_{\infty,S} \equiv \sup_{x \in S} |f(x)|$$

where  $S$  is any subset of  $[a, b]$  containing at least  $n + 1$  points.

From the Weierstrass approximation theorem, it follows that

$$\lim_{n \rightarrow \infty} D(f, \hat{P}_n) = 0.$$

Furthermore, if  $f(x)$  has  $r$  continuous derivatives in  $[a, b]$ , then by the convergence result for expansions in Chebyshev polynomials (see Theorem 9 in Subsection 3.4)

$$D(f, \hat{P}_n) = \mathcal{O}(n^{1-r}) \quad \text{for } r \geq 2.$$

#### 4.1. The Error in the Best Approximation

It is a relatively easy matter to obtain bounds on the deviation of the best approximation polynomial of degree  $n$ . Let us call this quantity

$$(4) \quad d_n(f) \equiv \min_{\{a_0, \dots, a_n\}} d(a_0, \dots, a_n) \equiv \min_{\{P_n(x)\}} D(f, P_n).$$

Then for any polynomial  $P_n(x)$  we have the upper bound

$$d_n(f) \leq D(f, P_n).$$

Lowers bounds can be obtained by means of

**THEOREM 2 (DE LA VALLÉE-POUSSIN).** *Let an  $n$ th degree polynomial  $P_n(x)$  have the deviations from  $f(x)$*

$$(5) \quad f(x_j) - P_n(x_j) = (-1)^j e_j, \quad j = 0, 1, \dots, n + 1,$$

where  $a \leq x_0 < x_1 < \dots < x_{n+1} \leq b$  and all  $e_j > 0$  or else all  $e_j < 0$ . Then

$$(6) \quad \min_j |e_j| \leq d_n(f).$$

*Proof.* Assume that for some polynomial  $Q_n(x)$ ,

$$(7) \quad D(f, Q_n) < \min_j |e_j|.$$

Then the  $n$ th degree polynomial

$$Q_n(x) - P_n(x) = [f(x) - P_n(x)] - [f(x) - Q_n(x)]$$

has the same sign at the points  $x_j$  as does  $f(x) - P_n(x)$ . Thus, there are  $n + 1$  sign changes and consequently, at least  $n + 1$  zeros of this difference.

But then, this  $n$ th degree polynomial identically vanishes and so  $P_n(x) \equiv Q_n(x)$ , which from (7) and (2) is impossible. This contradiction arose from assuming (7); hence  $D(f, Q_n) \geq \min_j |e_j|$  for every polynomial  $Q_n(x)$ , and (6) is established. ■

To employ this theorem we need only construct a polynomial which oscillates about the function being approximated at least  $n + 1$  times. This can usually be done by means of an interpolation polynomial of degree  $n$ .

A necessary and sufficient condition which characterizes a best approximation polynomial and establishes its uniqueness is contained in

**THEOREM 3 (CHEBYSHEV).** *A polynomial of degree at most  $n$ ,  $P_n(x)$ , is a best approximation of degree at most  $n$  to  $f(x)$  in  $[a, b]$  if and only if  $f(x) - P_n(x)$  assumes the values  $\pm D(f, P_n)$ , with alternate changes of sign, at least  $n + 2$  times in  $[a, b]$ . This best approximation polynomial is unique.*

*Proof.* Suppose  $P_n(x)$  has the indicated oscillation property. Then let  $x_j$ , with  $j = 0, 1, \dots, n + 1$  be  $n + 2$  points at which this maximum deviation is attained with alternate sign changes. Using these points in Theorem 2 we see that  $|e_j| = D(f, P_n)$  and hence

$$d_n(f) \geq D(f, P_n).$$

From equation (4), the definition of  $d_n(f)$ , it follows that  $D(f, P_n) = d_n(f)$  and the  $P_n(x)$  in question is a best approximation polynomial. This shows the sufficiency of the *uniform oscillation property*.

To demonstrate the necessity, we will show that if  $f(x) - P_n(x)$  attains the values  $\pm D(f, P_n)$  with alternate sign changes at most  $k$  times where  $2 \leq k \leq n + 1$ , then  $D(f, P_n) > d_n(f)$ . Let us assume, with no loss in generality, that

$$f(x_j) - P_n(x_j) = (-1)^j D(f, P_n), \quad j = 1, 2, \dots, k,$$

where  $a \leq x_1 < x_2 < \dots < x_k \leq b$ . Then, there exist points  $\xi_1, \xi_2, \dots, \xi_{k-1}$ , separating the  $x_j$ , i.e.,

$$x_1 < \xi_1 < x_2 < \xi_2 < \dots < \xi_{k-1} < x_k$$

and an  $\epsilon > 0$  such that  $|f(\xi_j) - P_n(\xi_j)| < D(f, P_n)$  and

$$-D(f, P_n) \leq f(x) - P_n(x) < D(f, P_n) - \epsilon,$$

for  $x$  in the “odd” intervals,  $[a, \xi_1], [\xi_2, \xi_3], [\xi_4, \xi_5], \dots$ ; while

$$-D(f, P_n) + \epsilon < f(x) - P_n(x) \leq D(f, P_n),$$

for  $x$  in the “even” intervals,  $[\xi_1, \xi_2], [\xi_3, \xi_4], \dots$ . For example, we may

define  $\xi_1 = \frac{1}{2}(\eta_1 + \zeta_1)$  where  $\eta_1 = \text{g.l.b. } \{\eta\}$  for  $a \leq \eta \leq x_2$  and  $f(\eta) - P_n(\eta) = D(f, P_n)$ ; and similarly:  $\zeta_1 = \text{l.u.b. } \{\zeta\}$  for  $a \leq \zeta \leq x_2$  and  $f(\zeta) - P_n(\zeta) = -D(f, P_n)$ . Then  $x_1 \leq \zeta_1 < \eta_1 \leq x_2$ ; otherwise, we may insert  $\eta_1$  and  $\zeta_1$  in place of  $x_1$  in the original sequence and find  $k + 1$  alternations of sign. That is, alternately for each of the  $k$  intervals  $[a, \xi_1], \dots, [\xi_{k-1}, b]$ , the deviation  $f(x) - P_n(x)$  takes on only one of the extreme deviations  $\pm D(f, P_n)$  and is bounded away from the extreme of opposite sign. The polynomial

$$r(x) = (x - \xi_1)(x - \xi_2) \cdots (x - \xi_{k-1})$$

has degree  $k - 1$  and is of one sign throughout each of the  $k$  intervals in question. Let the maximum value of  $|r(x)|$  in  $[a, b]$  be  $M$ . Now define  $q(x) \equiv (-1)^k r(x)/2M$  and consider the  $n$ th degree polynomial (since  $k - 1 \leq n$ )

$$Q_n(x) = P_n(x) + \epsilon q(x),$$

for sufficiently small positive  $\epsilon$ . We claim that  $D(f, Q_n) < D(f, P_n)$ , and so  $P_n(x)$  could not be a best approximation. Indeed, in the interior of any of the “odd” intervals  $(a, \xi_1), (\xi_2, \xi_3), \dots$ , we have that  $-\frac{1}{2} \leq q(x) < 0$  and conversely in the “even” intervals  $(\xi_1, \xi_2), (\xi_3, \xi_4), \dots$ , we have that  $0 < q(x) \leq \frac{1}{2}$ . However, recalling the above inequalities,

$$-D(f, P_n) - \epsilon q(x) \leq f(x) - Q_n(x) \leq D(f, P_n) - \epsilon[1 + q(x)], \quad x \text{ in odd intervals};$$

$$-D(f, P_n) + \epsilon[1 - q(x)] \leq f(x) - Q_n(x) \leq D(f, P_n) - \epsilon q(x), \quad x \text{ in even intervals}.$$

From the signs and magnitude of  $q(x)$  in each interval, it easily follows that  $D(f, Q_n) < D(f, P_n)$  and the proof of necessity is completed.

To demonstrate uniqueness we assume that there are two best approximations say,  $P_n(x)$  and  $Q_n(x)$ , both of degree at most  $n$ . Since by assumption  $D(f, P_n) = D(f, Q_n) = d_n(f)$ , we have in  $[a, b]$ ,

$$\begin{aligned} |f(x) - \frac{1}{2}[P_n(x) + Q_n(x)]| &\leq \frac{1}{2}|f(x) - P_n(x)| + \frac{1}{2}|f(x) - Q_n(x)| \\ &\leq d_n(f). \end{aligned}$$

Thus,  $\frac{1}{2}[P_n(x) + Q_n(x)]$  is another best approximation and we must have, by the first part of the theorem,

$$|f(x) - \frac{1}{2}[P_n(x) + Q_n(x)]| = d_n(f)$$

at  $n + 2$  distinct points in  $[a, b]$ . From the inequality, it follows that at these points  $f(x) - P_n(x) = f(x) - Q_n(x) = \pm d_n(f)$ . Thus, the difference  $[f(x) - P_n(x)] - [f(x) - Q_n(x)] = Q_n(x) - P_n(x)$  vanishes at  $n + 2$  distinct points. Since this difference is an  $n$ th degree polynomial, it vanishes identically, i.e.,  $Q_n(x) \equiv P_n(x)$ , and the proof is complete. ■

This theorem can be used to recognize the best approximation polynomial. It is also the basis, along with Theorem 2, of various methods for approximating the best approximation polynomial. There is no finite procedure for constructing the best approximation polynomial for arbitrary continuous functions. However, the best approximation is known in some important special cases; see, for example, the next subsection and Problem 2.

As an obvious consequence of Theorem 3, it follows that the best approximation,  $P_n(x)$ , of degree at most  $n$  is *equal* to  $f(x)$ , the function it approximates, *at*  $n + 1$  *distinct points*, say  $x_0, x_1, \dots, x_n$ . Thus,  $P_n(x)$  is *the* interpolation polynomial for  $f(x)$  with respect to the points  $\{x_i\}$  (since by Lemma 2.1 the interpolation polynomial of degree at most  $n$  is unique). Of course, for an arbitrary continuous function,  $f(x)$ , a corresponding set of interpolation points  $\{x_i\}$  is not known *a priori*. Thus, this observation cannot, in general, be used to determine  $P_n(x)$ . However, if  $f(x)$  has  $n + 1$  continuous derivatives, Theorem 2.1 applies since  $P_n(x)$  is an interpolation polynomial, and we have determined a form for the error in the best approximation of degree at most  $n$ . In summary, these observations can be stated as a

**COROLLARY.** *Let  $f(x)$  have a continuous  $(n + 1)$ st derivative in  $[a, b]$  and let  $P_n(x)$  be the best polynomial approximation to  $f(x)$  of degree at most  $n$  in this interval. Then, there exist  $n + 1$  distinct points  $x_0, x_1, \dots, x_n$  in  $a < x < b$  such that*

$$(8) \quad R_n(x) \equiv f(x) - P_n(x) = \frac{(x - x_0)(x - x_1) \cdots (x - x_n)}{(n + 1)!} f^{(n+1)}(\xi),$$

where  $\xi = \xi(x)$  is in the interval:

$$\min(x, x_0, \dots, x_n) < \xi < \max(x, x_0, \dots, x_n).$$
■

## 4.2. Chebyshev Polynomials

In the expression (8) for  $R_n(x)$ , the error of the best approximation, it will, in general, not be known at what point,  $\xi = \xi(x)$ , the derivative is to be evaluated. Hence, the value of the derivative is not known. An exception to this is the case when  $f^{(n+1)}(x) = \text{constant}$ , which occurs if and only if  $f(x)$  is a polynomial of degree at most  $n + 1$ . In this special case, the error (8) can be minimized by choosing the points  $x_0, x_1, \dots, x_n$  such that the polynomial

$$(9a) \quad (x - x_0)(x - x_1) \cdots (x - x_n)$$

has the smallest possible maximum absolute value in the interval in question (say,  $a \leq x \leq b$ ). In the general case, the choice of these same interpola-

tion points may be expected to yield a reasonable approximation to the best polynomial, i.e., the smaller the *variation* of  $f^{(n+1)}(x)$  in  $[a, b]$ , the better the approximation.

We are thus led to consider the following problem: *Among all polynomials of degree  $n + 1$ , with leading coefficient unity, find the polynomial which deviates least from zero in the interval  $[a, b]$ .* In other words, we are seeking the best approximation to the function  $g(x) \equiv 0$  among polynomials of the form

$$(9b) \quad x^{n+1} - P_n(x),$$

where  $P_n(x)$  is a polynomial of degree at most  $n$ . Alternatively, the problem can then be formulated as: *find the best polynomial approximation of degree at most  $n$  to the function  $x^{n+1}$ .*

For this latter problem, Theorem 1 is applicable and we conclude that such a polynomial exists and it is uniquely characterized in Theorem 3. Thus, we need only construct a polynomial of the form (9) whose maximum absolute value is attained at  $n + 2$  points with alternate sign changes.

Consider, for the present, the interval  $[a, b] \equiv [-1, 1]$ . We introduce the change of variable

$$(10) \quad x = \cos \theta,$$

which takes on each value in  $[-1, 1]$  once and only once when  $\theta$  is restricted to the interval  $[0, \pi]$ . Furthermore, the function  $\cos(n + 1)\theta$  attains its maximum absolute value, unity, at  $n + 2$  successive points with alternate signs for

$$\theta = j\left(\frac{\pi}{n+1}\right), \quad j = 0, 1, \dots, n+1.$$

Therefore, the function

$$(11) \quad T_{n+1}(x) = A_{n+1} \cos(n+1)\theta = A_{n+1} \cos[(n+1)\cos^{-1}x],$$

has the required properties as regards its extrema. To show that  $T_{n+1}(x)$  is also a polynomial in  $x$  of degree  $n + 1$  we consider the standard trigonometric addition formula

$$(12) \quad \cos(n+1)\theta + \cos(n-1)\theta = 2 \cos \theta \cos n\theta, \quad n = 0, 1, \dots$$

Let us define

$$(13a) \quad t_n(x) \equiv \cos(n \cos^{-1} x), \quad n = 0, 1, 2, \dots,$$

in terms of which (12) becomes

$$(13b) \quad t_{n+1}(x) = 2t_1(x)t_n(x) - t_{n-1}(x), \quad n = 1, 2, 3, \dots$$

Clearly, from (13a),  $t_0(x) = 1$ ,  $t_1(x) = x$  and so, by induction, it follows from (13b) that  $t_{n+1}(x)$  is a polynomial in  $x$  of degree  $n + 1$ . It also follows by induction that

$$(13c) \quad t_{n+1}(x) = 2^n x^{n+1} + q_n(x), \quad n = 0, 1, 2, \dots,$$

where  $q_n(x)$  is a polynomial of degree at most  $n$ . Thus, with the choice  $A_{n+1} = 2^{-n}$  in (11), these results imply

$$(14) \quad \begin{aligned} T_{n+1}(x) &= 2^{-n} \cos [(n+1) \cos^{-1} x] = 2^{-n} t_{n+1}(x) \\ &= x^{n+1} + 2^{-n} q_n(x). \end{aligned}$$

At the  $n + 2$  points

$$(15a) \quad \xi_k = \cos \frac{k\pi}{n+1}, \quad k = 0, 1, \dots, n+1,$$

which are in  $[-1, 1]$ , we have from (14)

$$(15b) \quad T_{n+1}(\xi_k) = 2^{-n} \cos k\pi = 2^{-n}(-1)^k.$$

Thus we have proven that  $T_{n+1}(x)$  is the polynomial of form (9) which deviates least from zero in  $[-1, 1]$ ; the maximum deviation is  $2^{-n}$ .

The polynomials in (14) are called the Chebyshev polynomials (of the first kind—see Problem 9 of Section 3). If the zeros of the  $(n+1)$ -st such polynomial are used to construct an interpolation polynomial of degree at most  $n$ , then for  $x$  in  $[-1, 1]$  the coefficient of  $f^{(n+1)}(\xi)$  in the error (8) of this approximation will have the least possible absolute maximum.

If the interval of approximation for the continuous function  $g(y)$  is  $a \leq y \leq b$ , then the transformation

$$(16) \quad x = \frac{a - 2y + b}{a - b} \quad \text{or} \quad y = \frac{1}{2}[(b - a)x + (a + b)]$$

converts the problem of approximating  $g(y)$  into that of approximating  $f(x) \equiv g[y(x)]$  in the  $x$ -interval  $[-1, 1]$ . The zeros of  $T_{n+1}(x)$  are at

$$(17a) \quad x_k = \cos \left( \frac{2k+1}{n+1} \frac{\pi}{2} \right), \quad k = 0, 1, \dots, n;$$

and the corresponding interpolation points in  $[a, b]$  are then at

$$(17b) \quad y_k = \frac{1}{2}[(b - a)x_k + (a + b)], \quad k = 0, 1, \dots, n.$$

The value of the maximum deviation of  $\prod_{j=0}^n (y - y_j)$  from zero in  $[a, b]$  is then, using (16) and (17b):

$$(18) \quad \begin{aligned} \max_{a \leq y \leq b} \prod_{j=0}^n |y - y_j| &= \left| \frac{b-a}{2} \right|^{n+1} \cdot \max_{-1 \leq x \leq 1} \prod_{j=0}^n (x - x_k) \\ &= \frac{1}{2^n} \left| \frac{b-a}{2} \right|^{n+1}. \end{aligned}$$

We stress that when the points in (17b) are employed to determine an interpolation polynomial for  $g(y)$  over  $[a, b]$ , this polynomial will not, in general, be the Chebyshev best approximation polynomial of degree  $n$ . However, these points may be used to get an estimate for the error of the best approximation polynomial. Iterative methods, based on Theorems 2 and 3, have been devised to compute with arbitrary precision the polynomial of best approximation.

## PROBLEMS, SECTION 4

1. Prove that the  $n$ th Chebyshev polynomial can be expressed as:

$$T_n(x) = 2^{-n}[(x + \sqrt{x^2 - 1})^n + (x - \sqrt{x^2 - 1})^n].$$

2. Find the best approximations of degrees 0 and 1 to  $f(x) \in C_2[a, b]$  provided  $f''(x) \neq 0$  in  $[a, b]$  [i.e., calculate the coefficients of these best approximation polynomials in terms of properties of  $f(x)$ ].

## 5. TRIGONOMETRIC APPROXIMATION

We say that  $S_n(x)$  is a *trigonometric sum of order at most n*, if

$$(1a) \quad S_n(x) = \frac{1}{2}a_0 + \sum_{k=1}^n (a_k \cos kx + b_k \sin kx).$$

The coefficients  $a_k$  and  $b_k$  are real numbers. By using the exponential function

$$(1b) \quad e^{i\theta} \equiv \cos \theta + i \sin \theta, \quad \cos \theta = \frac{1}{2}(e^{i\theta} + e^{-i\theta}),$$

$$\text{and} \quad \sin \theta = \frac{-i}{2}(e^{i\theta} - e^{-i\theta})$$

where now  $i^2 = -1$ , it follows that (1a) can be written in a simpler form with simpler coefficients:

$$(1c) \quad S_n(x) = \sum_{k=-n}^n c_k e^{ikx}.$$

Here

$$c_0 = \frac{a_0}{2}, \quad c_k = \frac{1}{2}(a_k - ib_k), \quad c_{-k} = \bar{c}_k = \frac{1}{2}(a_k + ib_k),$$

$$\text{for } k = 1, 2, \dots, n.$$

A basic result on approximation by such trigonometric sums is again due to Weierstrass and can be stated as:

**THEOREM 1 (WEIERSTRASS).** *Let  $f(\theta)$  be continuous on  $[-\pi, \pi]$  and periodic with period  $2\pi$ . Then for any  $\epsilon > 0$  there exists an  $n = n(\epsilon)$  and a trigonometric sum,  $S_n(\theta)$ , such that*

$$|f(\theta) - S_n(\theta)| < \epsilon$$

for all  $\theta$ .

*Proof.* A proof of this theorem can be given by employing the Weierstrass polynomial approximation theorem of Section 1.

We sketch a simpler proof based on the Weierstrass polynomial approximation theorem for a continuous function  $g(x, y)$  in the square,  $-1 \leq x, y \leq 1$  (see Problem 2, Section 1). Define  $g(x, y) \equiv \rho f(\theta)$  for  $x = \rho \cos \theta$ ,  $y = \rho \sin \theta$ ,  $0 \leq \rho \leq 2$ ,  $-\pi \leq \theta \leq \pi$ .

Clearly,  $g(x, y)$  is defined and continuous in the square. Hence, given  $\epsilon > 0$ , there exists a polynomial  $P_n(x, y)$  such that  $|g(x, y) - P_n(x, y)| \leq \epsilon$ . But, then for  $\rho = 1$ , we have  $g(x, y) \equiv f(\theta)$ ,  $x^2 + y^2 = 1$ , and therefore,  $|f(\theta) - P_n(\cos \theta, \sin \theta)| \leq \epsilon$ . We leave as Problem 3, the verification that  $P_n(\cos \theta, \sin \theta)$  may be written as a trigonometric sum,  $S_n(\theta)$ . ■

We proceed to show that the previous methods of polynomial approximation have corresponding trigonometric counterparts.

### 5.1. Trigonometric Interpolation

If the points of interpolation are *equally spaced*, it is relatively easy to determine a trigonometric sum which takes on specified values at the appropriate points. Let  $f(x)$  be continuous and have period  $2\pi$ . For this section only, we introduce the convention

$$\sum'_{j=-n}^n a_j \equiv \sum_{j=-n}^n a_j - \frac{1}{2}(a_n + a_{-n}).$$

With this notation, we define the trigonometric sum

$$(2) \quad U_n(x) = \sum'_{j=-n}^n c_j e^{ijx}.$$

On the interval  $[-\pi, \pi]$  we place the  $2n + 1$  equally spaced points

$$(3) \quad x_k = kh, \quad k = 0, \pm 1, \pm 2, \dots, \pm n, \quad h = \frac{\pi}{n}.$$

The interpolation problem is to find coefficients  $c_j$  such that

$$(4a) \quad U_n(x_k) = f(x_k), \quad k = 0, \pm 1, \dots, \pm n.$$

Later we consider interpolation on a different set of uniformly spaced points. Since  $f(x)$  and  $U_n(x)$  have period  $2\pi$ ,  $f(x_n) = f(x_{-n})$  and  $U_n(x_n) =$

$U_n(x_{-n})$ , so that there are only  $2n$  independent conditions in (4a) to determine the  $2n + 1$  coefficients,  $c_k$ . We require that

$$(4b) \quad c_n = c_{-n},$$

as the extra condition, and it will be shown to be consistent with the conditions (4a).

A simple calculation, based on summing a geometric series (see Sub-section 3.5), reveals that

$$(5) \quad \sum_{k=-n}^{n'} e^{ijx_k} e^{-imx_k} = \begin{cases} 0 & \text{if } j \not\equiv m \pmod{2n}, \\ 2n & \text{if } j \equiv m \pmod{2n}. \end{cases}$$

In direct analogy with orthogonal functions over an interval, we see that the quantities  $\{e^{ikx_j}\}$  are orthogonal with respect to the summation  $\sum_{j=-n}^{n'}$ .

Hence, we set  $x = x_k$  in (2), multiply both sides of (2) by  $e^{-imx_k}$  and sum with respect to  $k$  to find upon the use of (5):

$$\begin{aligned} \sum_{k=-n}^{n'} e^{-imx_k} U_n(x_k) &= \sum_{k=-n}^{n'} e^{-imx_k} \sum_{j=-n}^{n'} c_j e^{ijx_k}, \\ &= \sum_{j=-n}^{n'} c_j \sum_{k=-n}^{n'} e^{ijx_k} e^{-imx_k} \\ &= \begin{cases} 2nc_m & \text{if } |m| < n, \\ 2n\left(\frac{c_n + c_{-n}}{2}\right) & \text{if } |m| = n. \end{cases} \end{aligned}$$

By applying the conditions (4), then

$$(6) \quad c_j = \frac{1}{2n} \sum_{k=-n}^{n'} f(x_k) e^{-ijx_k}, \quad j = 0, \pm 1, \dots, \pm n.$$

It is now easy to check, by using (5), that the trigonometric sum (2) with coefficients given by (6) satisfies the conditions (4); i.e., the required interpolatory trigonometric sum is determined.

If we define new coefficients  $\alpha_j$  and  $\beta_j$  by

$$\alpha_j = c_j + c_{-j}, \quad \beta_j = i(c_j - c_{-j}), \quad j = 0, 1, 2, \dots, n,$$

then the sum (2) becomes, upon recalling (4b) and (1b),

$$(7) \quad U_n(x) = \frac{1}{2}\alpha_0 + \sum_{j=1}^n \alpha_j \cos jx + \sum_{j=1}^{n-1} \beta_j \sin jx.$$

† If  $j - m$  is an integral multiple of  $2n$ , we say that  $j$  and  $m$  are congruent modulo  $2n$ , or we write  $j \equiv m \pmod{2n}$ . If not, we write  $j \not\equiv m \pmod{2n}$ .

From (6) it follows that these coefficients are real numbers given by

$$(8) \quad \begin{aligned} \alpha_j &= \frac{1}{n} \sum_{k=-n}^n' f(x_k) \cos jx_k, \quad j = 0, 1, \dots, n, \\ \beta_j &= \frac{1}{n} \sum_{k=-n}^n' f(x_k) \sin jx_k, \quad j = 1, 2, \dots, n-1. \end{aligned}$$

Equations (7) and (8) are the real form for the trigonometric interpolation sum. This form is suitable for computations without complex arithmetic.

It can be shown that the trigonometric sum (7) satisfying the conditions (4) is unique. This follows from Lemma 2 in Subsection 5.3.

We may also determine unique interpolatory trigonometric sums of order  $n$  that take on specified values at  $2n + 1$  distinct points *arbitrarily spaced* in, say,  $-\pi \leq x < \pi$  (not including both endpoints). The coefficients of such a sum are the solutions of a non-singular linear system. The non-singularity of this system and the interpolating trigonometric sum are treated in Problems 1 and 2.

However, another trigonometric interpolation scheme for equally spaced points can be based on the orthogonality property expressed in Lemma 1 of Subsection 3.5. That is, using  $\theta$  as the independent variable, we consider the  $n + 1$  points,

$$(9) \quad \theta_j = \theta_0 + j\Delta\theta, \quad \theta_0 \equiv \frac{\Delta\theta}{2}, \quad \Delta\theta \equiv \frac{\pi}{n+1}, \quad j = 0, 1, \dots, n.$$

These  $\theta_j$  are equally spaced in  $[0, \pi]$ , there may be an odd or even number of them, and they do not include the endpoints [in contrast to those points in (3)]. Now we seek a special trigonometric sum of order  $n$  in the form

$$(10a) \quad C_n(\theta) = \frac{1}{2}\gamma_0 + \sum_{r=1}^n \gamma_r \cos r\theta,$$

such that for some function  $g(\theta)$ , continuous in  $[0, \pi]$ ,

$$(11) \quad C_n(\theta_j) = g(\theta_j), \quad j = 0, 1, \dots, n.$$

That is, we seek to interpolate  $g(\theta)$  at the points (9) with a sum of the form (10). Using the form (10) in (11) we multiply by  $\cos s\theta_j$  and sum over  $j$  to get by (3.62)

$$(10b) \quad \gamma_s = \frac{2}{n+1} \sum_{j=0}^n g(\theta_j) \cos s\theta_j, \quad s = 0, 1, \dots, n.$$

Thus, the interpolation problem is solved with the coefficients (10b) in the trigonometric sum (10a). [Compare the formulae (8) and (10b).] We note

that the sum (10a) is an even function of  $\theta$ . Hence, if the same is true of  $g(\theta)$ , then  $C_n(\theta)$  is the interpolation sum at  $2(n + 1)$  equally spaced points in  $[-\pi, \pi]$ . Again, uniqueness of the  $n$ th order sum easily follows in this case from Lemma 2 of Subsection 5.3. If  $g(\theta)$  is not even, then the approximation (10) should be used only over  $[0, \pi]$ .

An important convergence property of this approximation procedure is contained in

**THEOREM 2.** *Let  $g(\theta)$  be an even function with period  $2\pi$  and a continuous second derivative on  $[-\pi, \pi]$ . Then the trigonometric interpolation sums  $C_n(\theta)$  given by (10a and b), which satisfy (11) on the equally spaced points (9), converge uniformly as  $n \rightarrow \infty$  to  $g(\theta)$  on  $[-\pi, \pi]$ . In fact,*

$$|g(\theta) - C_n(\theta)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right), \quad \text{for all } |\theta| \leq \pi.$$

*Proof.* We first estimate the rate of decay of the coefficients  $\gamma_s$  for large  $s$ . Clearly from (10b), since  $|\cos s\theta| \leq 1$ ,

$$\begin{aligned} (12) \quad |\gamma_s| &\leq \frac{2}{n+1} \sum_{j=0}^n |g(\theta_j)|, \\ &\leq 2 \max_{0 \leq \theta \leq \pi} |g(\theta)|. \end{aligned}$$

Such a bound holds for the coefficients in any sum of the form (10).

With the spacing  $\Delta\theta = \pi/(n + 1)$  of (9), we define the function

$$(13) \quad G(\theta) \equiv \frac{g(\theta + \Delta\theta) - 2g(\theta) + g(\theta - \Delta\theta)}{\Delta\theta^2}.$$

This function satisfies the same smoothness and periodicity conditions as  $g(\theta)$ . If we set

$$(14) \quad B_n(\theta) \equiv \left(\frac{1}{\Delta\theta}\right)^2 [C_n(\theta + \Delta\theta) - 2C_n(\theta) + C_n(\theta - \Delta\theta)],$$

then from (11) and (13) it follows that

$$B_n(\theta_j) = G(\theta_j), \quad j = 0, 1, \dots, n,$$

and so,  $B_n(\theta)$  is the unique  $n$ th order trigonometric interpolation sum for  $G(\theta)$  with respect to the points  $\theta_j$  in (9). If we use (10a) and the identities

$$\cos(\phi + h) - 2\cos\phi + \cos(\phi - h) = \cos\phi(2\cos h - 2),$$

$$= -4\sin^2\frac{h}{2}\cos\phi,$$

in (14), we obtain

$$B_n(\theta) = -\left(\frac{2}{\Delta\theta}\right)^2 \sum_{r=1}^n \gamma_r \sin^2 \frac{r\Delta\theta}{2} \cos r\theta.$$

Thus,  $B_n(\theta)$  has the form (10a) and by the remark after (12) it now follows that

$$(15) \quad \left| \gamma_r r^2 \left( \frac{2}{r\Delta\theta} \sin \frac{r\Delta\theta}{2} \right)^2 \right| \leq 2 \max_{0 \leq \theta \leq \pi} |G(\theta)|, \quad r = 1, 2, \dots, n.$$

From Taylor's theorem

$$g(\theta \pm \Delta\theta) = g(\theta) \pm \Delta\theta g'(\theta) + \frac{(\Delta\theta)^2}{2} g''(\theta \pm \phi_{\pm} \Delta\theta), \quad 0 < \phi_{\pm} < 1.$$

By using this and the continuity of  $g''(\theta)$  in (13), we find

$$|G(\theta)| \leq \left| \max_{0 \leq \varphi \leq \pi} |g''(\varphi)| \right|, \quad \theta \in [0, \pi].$$

This bound and the inequality (see Problem 5)

$$\left| \frac{h}{\sin h} \right| \leq \frac{\pi}{2}, \quad \text{for } 0 < h \leq \frac{\pi}{2}.$$

in (15) yield finally

$$(16) \quad \begin{aligned} |\gamma_r| &\leq \frac{\pi^2}{2r^2} \max_{0 \leq \theta \leq \pi} |g''(\theta)|, \\ &= \mathcal{O}\left(\frac{1}{r^2}\right), \quad r = 1, 2, \dots, n. \end{aligned}$$

This is the required estimate of the coefficients in (10).

In Subsection 5.2 we define Fourier series, and for  $g(\theta)$  as above it follows from Theorem 3 that

$$(17) \quad |g(\theta) - S_m(\theta)| = \mathcal{O}\left(\frac{1}{m}\right), \quad a_m = \mathcal{O}\left(\frac{1}{m^2}\right), \quad m = 1, 2, \dots$$

where the partial sums,  $S_m(\theta)$ , and coefficients,  $a_m$ , are defined by

$$(18a) \quad S_m(\theta) \equiv \frac{a_0}{2} + \sum_{k=1}^m a_k \cos k\theta, \quad m = 1, 2, \dots$$

$$(18b) \quad a_k \equiv \frac{1}{\pi} \int_{-\pi}^{\pi} g(\theta) \cos k\theta d\theta, \quad k = 0, 1, 2, \dots$$

[The sine terms are absent since  $g(\theta)$  is even, and hence the  $b_k \equiv 0$ .] For any  $m \leq n$ , we define the truncated trigonometric interpolation sum

$$(19) \quad C_{n,m}(\theta) \equiv \frac{1}{2} \gamma_0 + \sum_{r=1}^m \gamma_r \cos r\theta,$$

where the  $\gamma_r$  are given in (10b).

We now consider

$$\begin{aligned}
 |g(\theta) - C_n(\theta)| &= |g(\theta) - S_m(\theta) + S_m(\theta) \\
 &\quad - C_{n,m}(\theta) + C_{n,m}(\theta) - C_n(\theta)|, \\
 (20) \qquad \qquad \qquad &\leq |g(\theta) - S_m(\theta)| + |S_m(\theta) - C_{n,m}(\theta)| \\
 &\quad + |C_{n,m}(\theta) - C_n(\theta)|.
 \end{aligned}$$

From (19), (10), and the estimate (16) we have

$$\begin{aligned}
 (21) \qquad |C_{n,m}(\theta) - C_n(\theta)| &= \left| \sum_{r=m+1}^n \gamma_r \cos r\theta \right|, \\
 &\leq \sum_{r=m+1}^n |\gamma_r|, \\
 &\leq \frac{\pi^2}{2} \max |g''(\theta)| \sum_{r=m+1}^{\infty} \frac{1}{r^2}, \\
 &= \mathcal{O}\left(\frac{1}{m}\right).
 \end{aligned}$$

To estimate the middle term in (20) we note from (9), (10b), (18b), and the evenness of  $g(\theta)$  that

$$\begin{aligned}
 a_k - \gamma_k &= \frac{2}{\pi} \left[ \int_0^\pi g(\theta) \cos k\theta d\theta - \sum_{j=0}^n g(\theta_j) \cos k\theta_j \Delta\theta \right], \\
 k &= 0, 1, \dots
 \end{aligned}$$

This sum is clearly an approximation to the integral. It is, in fact, the midpoint quadrature formula (see Chapter 7) and since the integrand has a continuous second derivative it is easily shown that (see Problem 6)

$$\begin{aligned}
 |a_k - \gamma_k| &\leq \frac{\pi^2}{12(n+1)^2} \max_{0 \leq \theta \leq \pi} \left| \frac{d^2}{d\theta^2} [g(\theta) \cos k\theta] \right| \\
 &= \begin{cases} \mathcal{O}(k^2/n^2), & k \geq 1, \\ \mathcal{O}(1/n^2), & k = 0. \end{cases}
 \end{aligned}$$

Using this estimate we have

$$\begin{aligned}
 (22) \qquad |S_m(\theta) - C_{n,m}(\theta)| &= \left| \frac{1}{2}(a_0 - \gamma_0) + \sum_{k=1}^m (a_k - \gamma_k) \cos k\theta \right|, \\
 &\leq \frac{1}{n^2} \left| \mathcal{O}(1) + \sum_{k=1}^m \mathcal{O}(k^2) \right|, \\
 &= \mathcal{O}\left(\frac{m^3}{n^2}\right).
 \end{aligned}$$

Thus (17), (21), and (22) in (20) imply

$$|g(\theta) - C_n(\theta)| = \mathcal{O}\left(\frac{1}{m} + \frac{m^3}{n^2}\right).$$

Finally, we set  $m = \sqrt{n}$ . ■

Several interesting corollaries easily follow. First, we have

**COROLLARY 1.** *Under the hypothesis of Theorem 2 let  $\sqrt{n} \leq m \leq n$  and  $C_{n,m}(\theta)$  be the  $m$ th order trigonometric sum defined by (19) and (10b). Then,*

$$(23) \quad |g(\theta) - C_{n,m}(\theta)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

*Proof.* It follows as in the theorem from (17), (22), and as in (21) that  $|C_{n,\sqrt{n}}(\theta) - C_{n,m}(\theta)| = \mathcal{O}(1/\sqrt{n})$ . ■

Thus various truncated trigonometric sums may furnish as good an approximation as the entire  $n$ th order sum. Next, by changing variables we obtain a result on the convergence of interpolation polynomials for special unequally spaced points. We state this as

**COROLLARY 2.** *Let  $f(x)$  have a continuous second derivative on  $[-1, 1]$ . If  $P_n(x)$  is the interpolation polynomial of degree at most  $n$  for  $f(x)$ , based on the  $n + 1$  points*

$$(24) \quad x_j = \cos\left(\frac{2j+1}{n+1}\frac{\pi}{2}\right), \quad j = 0, 1, 2, \dots, n,$$

*then  $P_n(x)$  converges to  $f(x)$  on  $[-1, 1]$  as  $n \rightarrow \infty$ . In fact,*

$$|f(x) - P_n(x)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

*Proof.* We introduce the new variable  $\theta$  in  $[0, \pi]$  by

$$\theta = \cos^{-1} x,$$

and then define

$$g(\theta) = f(\cos \theta).$$

We make  $g(\theta)$  even and continuous, and  $2\pi$  periodic by setting  $g(-\theta) = g(\theta)$ . Thus, the points (24) become the points  $\theta_j$  of (9) and  $g(\theta)$  satisfies the hypothesis of Theorem 2. Now the  $n$ th order interpolatory trigonometric sum (10) for  $g(\theta)$  becomes the interpolation polynomial of degree at most  $n$  in  $x$  (represented in terms of the Chebyshev polynomials) upon using the indicated variable change. So we have

$$|f(x) - P_n(x)| = |g(\theta) - C_n(\theta)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). ■$$

If we use the variable change and function  $f(x)$ , we find from Corollary 1 that for *some polynomials*, call them  $P_{n,m}(x)$ , of *degree at most m*, where  $\sqrt{n} \leq m \leq n$ ,

$$|f(x) - P_{n,m}(x)| = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right).$$

[Note that the  $P_{n,m}(x)$  are not obtained by simply truncating the interpolation polynomial,  $P_n(x)$ .] It is not difficult to show however, that  $P_{n,m}(x)$  is the *mth degree discrete least squares approximation to f(x) with respect to the n + 1 points* (24), see Problem 7. Thus we have established uniform convergence of the discrete least squares polynomials when the  $n + 1$  points are as in (24) and the degree is at least  $\sqrt{n}$ . Our present estimates of convergence indicate that no improvement occurs by interpolating at these  $n + 1$  points with a polynomial of degree  $n$ . This is another indication that high order interpolation polynomials should be avoided.

## 5.2. Least Squares Trigonometric Approximation. Fourier Series

If  $f(x)$  is periodic of period†  $2\pi$  and square integrable on  $[-\pi, \pi]$ , we can seek a trigonometric sum of form (1a) for which

$$(25) \quad \|f - S_n\|_2 = \left( \int_{-\pi}^{\pi} [f(x) - S_n(x)]^2 dx \right)^{\frac{1}{2}}$$

is a minimum with respect to all such sums. This norm now defines a quadratic function of  $2n + 1$  variables

$$J(a_0, a_1, \dots, a_n, b_1, \dots, b_n) = \|f - S_n\|_2^2$$

which can be minimized as was (3.8). The trigonometric functions satisfy the orthogonality relations

$$(26) \quad \begin{aligned} \int_{-\pi}^{\pi} \cos jx \cos kx dx &= \begin{cases} 0, & j \neq k, \\ \pi, & j = k \neq 0, \end{cases} \\ \int_{-\pi}^{\pi} \sin jx \sin kx dx &= \begin{cases} 0, & j \neq k, \\ \pi, & j = k \neq 0, \end{cases} \\ \int_{-\pi}^{\pi} \sin jx \cos kx dx &= 0. \end{aligned}$$

By using these results in the normal system obtained by minimizing (25), we find in analogy with (3.11),

$$(27) \quad a_k = \frac{1}{\pi} \int_{-\pi}^{\pi} \cos kx f(x) dx, \quad b_k = \frac{1}{\pi} \int_{-\pi}^{\pi} \sin kx f(x) dx.$$

† If the period of  $f(x)$  is some number  $p$ , then the change of variable  $\xi = 2\pi x/p$  results in a function  $g(\xi) \equiv f(p\xi/2\pi)$  which has period  $2\pi$ .

The trigonometric sum (1a) with coefficients given in (27) determines the best least squares approximation of order  $n$  to  $f(x)$  by such sums.

We deduce as in Section 3 the corresponding Bessel's inequality [by using (26) and (27) in  $\|f - S_n\|_2^2 \geq 0$ ]:

$$(28) \quad \frac{1}{2}a_0^2 + \sum_{k=1}^n (a_k^2 + b_k^2) \leq \frac{1}{\pi} \int_{-\pi}^{\pi} f^2(x) dx.$$

Since the right-hand side is independent of  $n$ , we also conclude that  $\frac{1}{2}a_0^2 + \sum_{k=1}^{\infty} (a_k^2 + b_k^2)$  converges and that

**LEMMA 1.**

$$\lim_{k \rightarrow \infty} a_k = \lim_{k \rightarrow \infty} b_k = 0. \quad \blacksquare$$

The trigonometric sum (1a) is, of course, the  $n$ th partial sum of the infinite series

$$(29) \quad \frac{1}{2}a_0 + \sum_{k=1}^{\infty} (a_k \cos kx + b_k \sin kx).$$

With coefficients given by (27), this is the *Fourier series* associated with the function  $f(x)$ .

We can now state

**THEOREM 3.** *Let  $f(x)$  be continuous and  $2\pi$  periodic. Then the partial sums  $S_n(x)$  of the Fourier series, with coefficients defined in (27), converge in the mean to  $f(x)$  and Parseval's equality holds. That is,*

$$(30a) \quad \lim_{n \rightarrow \infty} \int_{-\pi}^{\pi} [f(x) - S_n(x)]^2 dx = 0$$

$$(30b) \quad \frac{a_0^2}{2} + \sum_{k=1}^{\infty} a_k^2 + b_k^2 = \frac{1}{\pi} \int_{-\pi}^{\pi} [f(x)]^2 dx.$$

*Proof.* Simply modify the proof of Theorem 3, Section 3. ■

**THEOREM 4.** *Let  $f(x)$  have two continuous derivatives and be  $2\pi$  periodic. Then*

$$|a_k| = \mathcal{O}\left(\frac{1}{k^2}\right), \quad |b_k| = \mathcal{O}\left(\frac{1}{k^2}\right)$$

and

$$|f(x) - S_n(x)| = \mathcal{O}\left(\frac{1}{n}\right) \quad \text{for } -\pi \leq x \leq \pi.$$

*Proof.* Let  $a_k''$ ,  $b_k''$  be the Fourier coefficients corresponding to the  $2\pi$

periodic and continuous function  $f''(x)$ . Now, by repeated use of integration by parts,

$$\begin{aligned} a_k'' &= \frac{1}{\pi} \int_{-\pi}^{\pi} \cos kx f''(x) dx = \frac{k}{\pi} \int_{-\pi}^{\pi} \sin kx f'(x) dx \\ &= -\frac{k^2}{\pi} \int_{-\pi}^{\pi} \cos kx f(x) dx = -k^2 a_k \end{aligned}$$

and similarly  $b_k'' = -k^2 b_k$ .

But by Lemma 1,  $b_k'' \rightarrow 0$ , and  $a_k'' \rightarrow 0$ , whence

$$|a_k| = \mathcal{O}\left(\frac{1}{k^2}\right) \quad \text{and} \quad b_k = \mathcal{O}\left(\frac{1}{k^2}\right).$$

We therefore know that the Fourier series converges uniformly to a continuous function  $g(x)$ . Hence, we may let  $n \rightarrow \infty$  under the integral in the statement (30a) of mean convergence, thus proving that  $f(x) \equiv \lim_{n \rightarrow \infty} S_n(x)$ . The error estimate  $|f(x) - S_n(x)| = \mathcal{O}(1/n)$  follows from the boundedness of  $\{\cos kx, \sin kx\}$  and the relation

$$\sum_n^{\infty} \frac{1}{k^2} = \mathcal{O}\left(\frac{1}{n}\right). \quad \blacksquare$$

The theory of approximation by orthogonal functions owes much to J. Fourier, who employed trigonometric series of the form (29). In fact, least squares approximations of the form (3.7) with orthogonal polynomials or other orthogonal sets of functions are generally called Fourier series (assuming  $n \rightarrow \infty$ ) and the coefficients given by (3.11), (3.21), or (27) are called the Fourier coefficients.

Finally, we observe a close connection between the trigonometric interpolation coefficients (8) and the Fourier coefficients (27). Recalling the definitions of  $\sum'$  and the points  $x_k$  in (3) we can write (8) as

$$\begin{aligned} \alpha_j &= \frac{1}{\pi} \sum_{k=1-n}^n \left[ \frac{\cos jx_{k-1} f(x_{k-1}) + \cos jx_k f(x_k)}{2} \right] (x_k - x_{k-1}), \\ \beta_j &= \frac{1}{\pi} \sum_{k=1-n}^n \left[ \frac{\sin jx_{k-1} f(x_{k-1}) + \sin jx_k f(x_k)}{2} \right] (x_k - x_{k-1}). \end{aligned}$$

As  $n \rightarrow \infty$  we have  $x_k - x_{k-1} = \pi/n \rightarrow 0$  and [say for piecewise continuous  $f(x)$ ] these sums converge to the corresponding integrals in (27). Thus,  $(\alpha_j, \beta_j) \rightarrow (a_j, b_j)$  and the trigonometric interpolating sum (7) converges, formally, to the Fourier series (29).

These sums correspond to the trapezoidal rule of numerical integration. On the other hand, the coefficients  $\gamma_j$  in (10b) for trigonometric interpolation with respect to the points  $\theta_j$  in (9) approximate the coefficients  $a_j$ ,

by the midpoint rule for evaluating the integrals in (27). In the proof of Theorem 2, it is shown that for fixed  $j$ :  $|\gamma_j - a_j| = \mathcal{O}(1/n^2)$ . [For the case that the function  $g(\theta)$  is not necessarily even, the corresponding trigonometric interpolatory coefficients are defined in Problem 4, and similarly converge to the Fourier coefficients.]

### 5.3. “Best” Trigonometric Approximation

If  $f(x)$  is continuous on  $[-\pi, \pi]$  we can seek a trigonometric sum (1), of order  $n$ , which minimizes the maximum norm

$$\|f(x) - S_n(x)\|_\infty = \max_{-\pi \leq x \leq \pi} |f(x) - S_n(x)|.$$

The existence of such a best trigonometric approximation could be demonstrated by using an analogue of Theorem 0.1. Another proof is given in Problem 9.

Results analogous to those in Theorem 4.2 and Theorem 4.3 are also valid for the best trigonometric approximation of order  $n$ . A careful glance at the proofs of these theorems reveals that the only property of polynomials employed is the fact that if a polynomial of degree  $n$  vanishes at  $n + 1$  points, then it vanishes identically. Such a property is also true of trigonometric sums. In fact, best approximation by other sets of functions is possible and the property they must possess to insure a unique best approximation is called the *Haar property*, defined as follows:

*A sequence of functions  $\{f_0(x), f_1(x), \dots\}$  has the Haar property if for every  $m$  the only linear combination*

$$P_m(x) \equiv a_0 f_0(x) + a_1 f_1(x) + \dots + a_m f_m(x)$$

*with  $m + 1$  distinct zeros† is the identically vanishing combination  $P_m(x) \equiv 0$ .* It was proven by Haar that these conditions are necessary and sufficient for uniqueness. However, we shall be concerned only with the trigonometric case. Thus, we consider

**LEMMA 2.** *The sequence of trigonometric functions*

$$\{1, \cos x, \sin x, \cos 2x, \sin 2x, \dots, \cos nx, \sin nx, \dots\}$$

*has the Haar property.*

*Proof.* We need only show that every non-trivial trigonometric sum, (1), of order  $n$  has at most  $2n$  roots in  $-\pi \leq x < \pi$ . Let us define  $\xi = e^{ix}$  and note that  $|\xi| = 1$ . Then we have from (1c),

$$S_n(x) = \xi^{-n} \sum_{k=-n}^n c_k \xi^{n+k}.$$

† If the  $f_k(x)$  are periodic, then the zeros must all lie in a period.

The sum on the right-hand side is a polynomial in  $\xi$  of degree at most  $2n$ . Thus, this sum has at most  $2n$  roots if it has any non-zero coefficients. ■

This lemma can be used to prove uniqueness of the trigonometric interpolation problems solved in Subsection 5.1 and Problem 2. We apply the lemma to prove the analog of Theorem 4.2, namely

**THEOREM 5.** *Let  $f(x) \in C[-\pi, \pi]$  and let an  $n$ th order trigonometric sum,  $S_n(x)$ , have the deviations from  $f(x)$*

$$f(x_j) - S_n(x_j) = (-1)^j e_j, \quad j = 0, 1, \dots, 2n + 1,$$

where  $-\pi \leq x_0 < x_1 < \dots < x_{2n+1} \leq \pi$  and all  $e_j > 0$  or all  $e_j < 0$ . Then

$$\min_j |e_j| \leq \min_{\{S_n\}} \|f(x) - S_n(x)\|_\infty.$$

*Proof.* Assume that for some trigonometric sum of order  $n$ , say  $S_n^*(x)$ ,

$$\|f(x) - S_n^*(x)\|_\infty < \min_j |e_j|.$$

Then, the  $n$ th order sum

$$S_n^*(x) - S_n(x) = [f(x) - S_n(x)] - [f(x) - S_n^*(x)]$$

has the same sign at the points  $x_j$  as does  $f(x) - S_n(x)$ . Thus there are  $2n + 1$  sign changes and at least  $2n + 1$  zeros of this difference in  $(-\pi, \pi)$ . But then, by the above lemma, this trigonometric sum vanishes identically and so  $S_n(x) \equiv S_n^*(x)$  which leads to a contradiction. ■

Continuing the analogy, we have in place of Theorem 4.3

**THEOREM 6.** *A trigonometric sum of order  $n$ ,  $S_n(x)$ , is the best trigonometric approximation of order  $n$  to  $f(x) \in C[-\pi, \pi]$  if and only if  $f(x) - S_n(x)$  assumes the values  $\pm \|f(x) - S_n(x)\|_\infty$ , with alternate changes of sign, at least  $2n + 2$  times in  $[-\pi, \pi]$ . This best approximation is unique.*

*Proof.* The proof is exactly analogous to that of Theorem 4.4. To show sufficiency we employ Theorem 5. To demonstrate necessity we must construct a trigonometric sum of order  $2k$ , say, and which has specified zeros at distinct points  $\xi_1, \xi_2, \dots, \xi_{2k}$  in  $(-\pi, \pi)$ . This is done by forming the determinant

$$(31) \quad t(x) = \begin{vmatrix} 1 & \cos x & \sin x & \cdots & \cos kx & \sin kx \\ 1 & \cos \xi_1 & \sin \xi_1 & \cdots & \cos k\xi_1 & \sin k\xi_1 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & \cos \xi_{2k} & \sin \xi_{2k} & \cdots & \cos k\xi_{2k} & \sin k\xi_{2k} \end{vmatrix}.$$

The expression  $t(x)$  is used in place of the polynomial  $r(x)$  to obtain a contradiction [Note the relation of  $t(x)$  to the determinant in Problem 1.] The uniqueness follows by using the Haar property of the trigonometric functions. The details are left to the reader. ■

### PROBLEMS, SECTION 5

- 1.** Let  $-\pi \leq x_0 < x_1 < \dots < x_{2n} < \pi$ , and define the determinant of order  $2n + 1$ ,

$$\Delta = \begin{vmatrix} 1 & \cos x_0 & \sin x_0 & \cdots & \cos nx_0 & \sin nx_0 \\ 1 & \cos x_1 & \sin x_1 & \cdots & \cos nx_1 & \sin nx_1 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & \cos x_{2n} & \sin x_{2n} & \cdots & \cos nx_{2n} & \sin nx_{2n} \end{vmatrix}.$$

Show that  $\Delta \neq 0$  and in fact, that

$$\Delta = (-1)^{n(n-1)/2} 2^{2n^2} \prod_{j=1}^{2n} \left[ \prod_{k=0}^{j-1} \sin \left( \frac{x_j - x_k}{2} \right) \right].$$

[Hint: Express  $\sin \theta$  and  $\cos \theta$  in exponential form, form linear combinations of successive pairs of columns so that only one exponential appears in each element, rearrange columns so that each row forms a geometric progression. The result is a Vandermonde determinant.]

- 2.** Show that  $S_n(x)$  is a trigonometric sum which satisfies  $S_n(x_j) = f(x_j)$ ,  $j = 0, 1, \dots, 2n$ , where  $\{x_k\}$  is given as in Problem 1 and

$$S_n(x) = \sum_{j=0}^{2n} f(x_j) \prod_{\substack{k=0 \\ (k \neq j)}}^{2n} \left( \frac{\sin \frac{x-x_k}{2}}{\sin \frac{x_j-x_k}{2}} \right).$$

This is the general interpolatory trigonometric sum of order  $n$ .

[Hint: Use Problem 3 and  $\sin a \sin b = \frac{1}{2} [\cos(a-b) - \cos(a+b)]$ .]

- 3.** Verify that a trigonometric polynomial of degree at most  $n$ ,

$$P_n(\cos \theta, \sin \theta) \equiv \sum_{i+j \leq n} c_{i,j} (\cos \theta)^i (\sin \theta)^j,$$

may be written as a trigonometric sum of order at most  $n$ ,

$$P_n(\cos \theta, \sin \theta) \equiv S_n(\theta) \equiv \sum_{k=0}^n a_k \cos k\theta + b_k \sin k\theta,$$

and vice versa.

[Hint: Use (1b).]

- 4.** Given  $g(x)$  is  $2\pi$  periodic, find the trigonometric sum

$$S_n(x) = \frac{a_0}{2} + \sum_{k=1}^{n-1} (a_k \cos kx + b_k \sin kx) + \frac{b_n}{2} \sin nx$$

such that

$$S_n(x_j) = g(x_j) \quad \text{for } x_j = \frac{(2j+1)\pi}{2n}, \quad -n \leq j \leq n-1.$$

[Hint: Establish  $\sum_{j=-n}^{n-1} \cos rx_j \cos sx_j = \begin{cases} 0 & r \neq s \\ n & n > r = s > 0 \\ 2n & r = s = 0 \\ 0 & r = s = n \end{cases}$

$$\sum_{j=-n}^{n-1} \cos rx_j \sin sx_j = 0$$

$\sum_{j=-n}^{n-1} \sin rx_j \sin sx_j = \begin{cases} 0 & r \neq s \\ n & n > r = s > 0 \\ 0 & r = s = 0 \\ 2n & r = s = n, \end{cases}$

whence

$$a_r = \frac{1}{n} \sum_{j=-n}^{n-1} g(x_j) \cos rx_j,$$

$$b_r = \frac{1}{n} \sum_{j=-n}^{n-1} g(x_j) \sin rx_j.]$$

5. Verify that

$$\left| \frac{\theta}{\sin \theta} \right| \leq \frac{\pi}{2}$$

provided  $|\theta| \leq \pi/2$ .

[Hint: Consider the chord joining  $(0, 0)$  and  $(\pi/2, 1)$  on the graph of  $y = \sin x$ .]

6. Let  $f''(x)$  be continuous in  $[a, b]$  and

$$x_j = x_0 + jh, \quad x_0 = a + \frac{h}{2}, \quad h = \frac{b-a}{n+1}, \quad j = 0, 1, \dots, n.$$

Show that

$$|E| \equiv \left| \int_a^b f(x) dx - \sum_{j=0}^n f(x_j)h \right| \leq \frac{|b-a|}{24} h^2 \max_{a < x < b} |f''(x)|.$$

[Hint: Write

$$E = \sum_{j=0}^n \int_{x_j-h/2}^{x_j+h/2} [f(x) - f(x_j)] dx$$

and then use Taylor's theorem in each integrand to get

$$f(x) = f(x_j) + (x - x_j)f'(x_j) + \frac{(x - x_j)^2}{2} f''\left(x_j + \theta \frac{h}{2}\right), \quad -1 < \theta < 1.]$$

7. Show that the  $m$ th degree discrete least squares polynomial approximation to  $f(x)$  with respect to the  $n + 1$  points (24) is obtained by truncating the  $n$ th order trigonometric interpolation sum (10) for  $g(x)$  on the points (9) after the  $m$ th term and using the variable change  $\theta = \cos^{-1} x$ ,  $g(\theta) = f(\cos \theta)$ .

[Hint: Simply change variables in the representation (3.52) by using the discrete orthonormal polynomials (3.61). The uniqueness of the discrete least square polynomials is used.]

8. With the notation of equations (2), (3), (4), and (6), verify the *discrete* analogue of the *Parseval equality*,

$$\frac{1}{2n} \sum_{k=-n}^n |f(x_k)|^2 = \frac{1}{2n} \sum_{k=-n}^n |U_n(x_k)|^2 = \sum_{j=-n}^n |c_j|^2.$$

[Hint: Use equation (5).]

9. Prove the existence of a best trigonometric approximation in the form (1) for the given function  $f(x)$  on the interval  $[-\pi, \pi]$ .

[Hint: In Problem 2, if  $|f(x)| < M$  and  $\{x_i\}$  are distinct, then the trigonometric sum  $S_n(x)$  has bounded coefficients. For fixed  $n$ , consider the non-empty set,  $C$ , of trigonometric sums  $\{S_n(x)\}$  such that

$$\|f(x) - S_n(x)\|_\infty \leq \frac{M}{2}.$$

Note that by the above remark, the coefficients of all of the sums in  $C$  are bounded. Let  $S_n^\nu(x)$ ,  $\nu = 1, 2, \dots$ , be a *minimizing sequence* of trigonometric sums in  $C$ , that is,

$$\|f - S_n^\nu\|_\infty \rightarrow \text{g.l.b. } \|f - S_n\|_\infty.$$

Pick a subsequence of  $S_n^\nu$  such that their coefficients converge, i.e.,

$$a_k^\nu \rightarrow \hat{a}_k, b_k^\nu \rightarrow \hat{b}_k.$$

The sum

$$\hat{S}_n(x) \equiv \frac{\hat{a}_0}{2} + \sum_{k=1}^n \hat{a}_k \cos kx + \hat{b}_k \sin kx$$

is a best approximation.]

# 6

## Differences, Interpolation Polynomials, and Approximate Differentiation

### 0. INTRODUCTION

Interpolation polynomials are of particular importance in numerical analysis, and so we devote special attention to them in this chapter. We are led naturally to the study of differences; both divided differences for arbitrarily spaced points and ordinary differences for equally spaced points. Not only can differences be neatly arranged in tables, for convenient hand computation (presumably an affair of the past), but they permit one to easily estimate the error in the approximation. Hence, methods based on differences are useful in this age of digital computers as they suggest very efficient computing techniques and can be used for checking the accuracy of a calculation.

We examine the error in interpolation, when the polynomial passes through equally spaced points, in some detail. This error is generally much less near the center of the interval of interpolation points and grows rapidly outside this interval, i.e., for what is termed extrapolation. Therefore, we construct special forms of the interpolation formulae which are convenient for evaluation near the center of the interpolation interval.

We use interpolation polynomials to determine formulae for the numerical approximation of derivatives of the interpolated function.

## 1. NEWTON'S INTERPOLATION POLYNOMIAL AND DIVIDED DIFFERENCES

We have shown in Chapter 5, Section 2, that the interpolation polynomial exists and is unique. Furthermore, for a fixed set of interpolation points, it is easily constructed using the Lagrange interpolation coefficients. The Lagrange representation has the defect that if another point of interpolation were added, then the new higher degree interpolation polynomial could not be obtained by easily modifying the previous one. (This is in contrast, say, to Taylor's series expansion or to least squares expansion in orthogonal functions where the next order approximation is obtained by simply adjoining a term to the present approximation.) We seek then a representation of the interpolation polynomial which has the property that the next higher degree interpolation polynomial is found by simply adding a new term.

Specifically, let  $Q_k(x)$  be the interpolation polynomial for  $f(x)$ , of degree at most  $k$ , with respect to the  $k + 1$  distinct points  $x_0, x_1, \dots, x_k$ . We seek the successive interpolation polynomials,  $\{Q_k(x)\}$ , of degree at most  $k$  in the form  $Q_0(x) \equiv f(x_0)$  and

$$(1a) \quad Q_k(x) = Q_{k-1}(x) + q_k(x), \quad \text{for } k = 1, 2, \dots, n,$$

where  $q_k(x)$  has at most degree  $k$ . Since we require

$$Q_k(x_j) = f(x_j) = Q_{k-1}(x_j), \quad j = 0, 1, \dots, k - 1$$

it follows that  $q_k(x_j) = 0$  at these  $k$  points. Thus, we may write

$$(1b) \quad q_k(x) = a_k \prod_{j=0}^{k-1} (x - x_j),$$

which represents the most general polynomial of degree at most  $k$  that vanishes at the indicated  $k$  points. The constant  $a_k$  remains to be determined. But, in order that  $Q_k(x_k) = f(x_k)$ , it follows from (1a and b) that

$$(1c) \quad a_k = \frac{f(x_k) - Q_{k-1}(x_k)}{\prod_{j=0}^{k-1} (x_k - x_j)}, \quad \text{for } k = 1, 2, \dots, n.$$

The zero degree interpolation polynomial for the initial point  $x_0$  is, trivially,  $Q_0(x) \equiv f(x_0)$ . Thus, with  $a_0 = f(x_0)$ , we obtain by using (1a and b) recursively, for  $k = 1, 2, \dots, n$ ,

$$(2) \quad Q_n(x) = a_0 + (x - x_0)a_1 + \cdots + (x - x_0) \cdots (x - x_{n-1})a_n.$$

The  $k$ th coefficient is called the  $k$ th order divided difference and is usually expressed in the notation

$$(3) \quad \begin{aligned} a_0 &= f[x_0], \\ a_k &= f[x_0, x_1, \dots, x_k], \quad k = 1, 2, \dots \end{aligned}$$

The values of  $f(x)$  which enter into the determination of  $a_k$  are those at the arguments of  $f[x_0, x_1, \dots, x_k]$ . We now obtain a representation for these coefficients which is more explicit than the recursive form given in (1). Since  $Q_n(x)$  in (2) is *the unique* interpolation polynomial of degree  $n$ , we may equate the leading coefficient,  $a_n$ , in this form with that obtained by using the Lagrange form, see (2.5) in Chapter 5. That is, from

$$Q_n(x) = \sum_{j=0}^n f(x_j) \prod_{\substack{k=0 \\ k \neq j}}^n \frac{x - x_k}{x_j - x_k}$$

the coefficient of  $x^n$  is

$$(4) \quad a_n = f[x_0, x_1, \dots, x_n] = \sum_{j=0}^n \frac{f(x_j)}{\prod_{\substack{k=0 \\ (k \neq j)}}^n (x_j - x_k)}.$$

This form could also be deduced directly from (1); see Problem 1.

From the representation (4) it follows that *the divided differences are symmetric functions of their arguments*. That is, if we adopt the additional notation

$$f_{i, j, k, \dots} \equiv f[x_i, x_j, x_k, \dots]$$

then this symmetry is expressed by

$$(5) \quad f_{0, 1, \dots, n} = f_{j_0, j_1, \dots, j_n}$$

where  $(j_0, j_1, \dots, j_n)$  is any permutation of the integers  $(0, 1, \dots, n)$ .

We may derive a more convenient form than (4) for computing the divided differences by again making use of the *uniqueness* of the interpolation polynomial. That is, we may construct the polynomial  $Q_n(x)$  by matching the values of  $f(x_j)$  in the reverse order  $j = n, n-1, \dots, 1, 0$ . In this way we would obtain, say,

$$(2') \quad Q_n(x) \equiv b_0 + (x - x_n)b_1 + \dots + (x - x_n)(x - x_{n-1}) \cdots (x - x_1)b_n$$

where

$$b_k = f[x_n, x_{n-1}, \dots, x_{n-k}] \quad \text{and} \quad b_0 = f[x_n] = f(x_n).$$

But  $a_n = b_n$  since they are the coefficients of  $x^n$  in the unique polynomial  $Q_n(x)$  of (2) and (2').

Now if we subtract equation (2') from equation (2) but display only the terms which contribute to the coefficients of  $x^n$  and  $x^{n-1}$  we obtain, using  $a_n = b_n$ ,

$$0 \equiv [(x - x_0) - (x - x_n)](x - x_1) \cdots (x - x_{n-1})a_n \\ + (a_{n-1} - b_{n-1})x^{n-1} + p_{n-2}(x)$$

where  $p_{n-2}(x)$  is a polynomial of degree at most  $n - 2$  in  $x$ . Since this expression vanishes identically the coefficient of  $x^{n-1}$  vanishes and hence  $0 = [x_n - x_0]a_n + (a_{n-1} - b_{n-1})$ . Now the symmetry of the divided differences, proven above, implies that

$$b_{n-1} = f[x_n, x_{n-1}, \dots, x_1] = f[x_1, x_2, \dots, x_n],$$

whence from  $a_n = (a_{n-1} - b_{n-1})/(x_0 - x_n)$ , we have

$$(6) \quad f[x_0, x_1, \dots, x_n] = \frac{f[x_0, x_1, \dots, x_{n-1}] - f[x_1, x_2, \dots, x_n]}{x_0 - x_n}, \\ n = 1, 2, \dots$$

We leave it to the reader to verify (6) directly from (4) in Problem 9. This recursion formula justifies the use of the name divided difference. Of course, we then define for completeness

$$f[x_0] = f(x_0).$$

The interpolation polynomial (2) may now be written as

$$(7) \quad Q_n(x) = f[x_0] + (x - x_0)f[x_0, x_1] + \cdots \\ + (x - x_0) \cdots (x - x_{n-1})f[x_0, x_1, \dots, x_n].$$

This form is known as *Newton's divided difference interpolation formula*. Note that to obtain the next higher degree such polynomial we need only add a term similar to the last term but involving a new divided difference of one higher order.

In fact, let us set  $k = n + 1$  in (1b and c) and (3) and then replace  $x_{n+1}$  by  $x$ . We obtain

$$(8) \quad f(x) - Q_n(x) = \left[ \prod_{j=0}^n (x - x_j) \right] f[x_0, x_1, \dots, x_n, x],$$

which for  $x$  distinct from  $\{x_j\}$  defines the indicated divided difference. On the other hand, this identity gives another representation of the *error* in polynomial interpolation.

By means of formula (6) applied to  $(x_j, x_{j+1}, \dots, x_{j+n})$  we can construct a table of divided differences in a symmetric manner based on

$$(6), \quad f_{j, j+1, \dots, j+n} = \frac{f_{j+1, j+2, \dots, j+n} - f_{j, j+1, \dots, j+n-1}}{x_{j+n} - x_j}.$$

**Table 1** *Divided Differences*

$x$	$f(x)$	$f[x, x]$	$f[x, x, x]$	$f[x, x, x, x]$	...
$x_0$	$f_0$	$\frac{f_1 - f_0}{x_1 - x_0} \equiv f_{01}$			
$x_1$	$f_1$		$\frac{f_{12} - f_{01}}{x_2 - x_0} \equiv f_{012}$		
		$\frac{f_2 - f_1}{x_2 - x_1} \equiv f_{12}$		$\frac{f_{123} - f_{012}}{x_3 - x_0} \equiv f_{0123}$	
$x_2$	$f_2$		$\frac{f_{23} - f_{12}}{x_3 - x_1} \equiv f_{123}$		...
		$\frac{f_3 - f_2}{x_3 - x_2} \equiv f_{23}$		$\frac{f_{234} - f_{123}}{x_4 - x_1} \equiv f_{1234}$	
$x_3$	$f_3$		$\frac{f_{34} - f_{23}}{x_4 - x_2} \equiv f_{234}$		:
		$\frac{f_4 - f_3}{x_4 - x_3} \equiv f_{34}$	:		
$x_4$	$f_4$	⋮			
⋮	⋮				

See Table 1. The divided difference required to determine  $Q_{n+1}(x)$  from  $Q_n(x)$  is easily obtained from Table 1 by just completing another "diagonal" line of differences. This simple property is not shared by the Lagrange form of the interpolation polynomial.

Another representation of the divided differences, which is quite useful for estimating their magnitude as well as for many theoretical purposes, is contained in

**THEOREM 1.** *Let  $x, x_0, x_1, \dots, x_{k-1}$  be  $k + 1$  distinct points and let  $f(y)$  have a continuous derivative of order  $k$  in the interval*

$$\min(x, x_0, \dots, x_{k-1}) < y < \max(x, x_0, \dots, x_{k-1}).$$

*Then for some point  $\xi = \xi(x)$  in this interval*

$$(9) \quad f[x_0, \dots, x_{k-1}, x] = \frac{f^{(k)}(\xi)}{(k)!}.$$

*Proof.* From equation (8) with  $n$  replaced by  $k - 1$  we write

$$f(x) - Q_{k-1}(x) = (x - x_0)(x - x_1) \cdots (x - x_{k-1})f[x_0, \dots, x_{k-1}, x].$$

However, since  $Q_{k-1}(x)$  is an interpolation polynomial which is equal to

$f(x)$  at the  $k$  points  $x_0, x_1, \dots, x_{k-1}$  it follows from Theorem 2.1 of Chapter 5 that

$$\begin{aligned} f(x) - Q_{k-1}(x) &\equiv R_{k-1}(x) \\ &= (x - x_0)(x - x_1) \cdots (x - x_{k-1}) \frac{f^{(k)}(\xi)}{(k)!}. \end{aligned}$$

But by the hypothesis  $(x - x_0)(x - x_1) \cdots (x - x_{k-1}) \neq 0$ , and the theorem follows by equating the above right-hand sides. ■

A generalization, permitting coincident values, is established as Corollary 2 of Theorem 2 that follows.

As an immediate consequence of Theorem 1, we can obtain some information on the divided differences of polynomials. These results may be stated as the

**COROLLARY.** *Let*

$$P_n(x) = \alpha_0 + \alpha_1 x + \cdots + \alpha_n x^n, \quad \alpha_n \neq 0,$$

*be any polynomial of degree  $n$  and let  $x_0, x_1, \dots, x_k$  be any  $k+1$  distinct points. Then*

$$P_n[x_0, x_1, \dots, x_k] = \begin{cases} \alpha_n & \text{if } k = n, \\ 0 & \text{if } k > n. \end{cases}$$

*Proof.* The corollary follows from Theorem 1 since  $d^n P_n(x)/dx^n = n! \alpha_n$ ; and higher derivatives vanish. ■

We shall require some continuity and differentiability properties of divided differences for our later discussion of the error in numerical differentiation and integration. Most of these results can be derived from still another representation of the divided differences which we state as

**THEOREM 2.** *Let  $f(x)$  have a continuous  $n$ th derivative in the interval  $\min(x_0, x_1, \dots, x_n) \leq x \leq \max(x_0, x_1, \dots, x_n)$ . Then if the points  $x_0, x_1, \dots, x_n$  are distinct,*

$$(10)_n \quad f[x_0, x_1, \dots, x_n] = \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \times f^{(n)}(t_n[x_n - x_{n-1}] + \cdots + t_1[x_1 - x_0] + x_0),$$

where  $n \geq 1$ ,  $t_0 = 1$ .

*Proof.* For an inductive proof, we first show that

$$f[x_0, x_1] = \int_0^1 dt_1 f'(t_1[x_1 - x_0] + x_0).$$

Let a new variable of integration,  $\xi$ , be introduced by (since  $x_1 \neq x_0$ ):

$$\xi = t_1[x_1 - x_0] + x_0, \quad dt_1 = d\xi/[x_1 - x_0].$$

The integration limits become

$$(t_1 = 0) \rightarrow (\xi = x_0); \quad (t_1 = 1) \rightarrow (\xi = x_1).$$

Therefore, we have

$$\begin{aligned} \int_0^1 dt_1 f'(t_1[x_1 - x_0] + x_0) &= \frac{1}{x_1 - x_0} \int_{x_0}^{x_1} d\xi f'(\xi) \\ &= \frac{f(x_1) - f(x_0)}{x_1 - x_0} = f[x_0, x_1]. \end{aligned}$$

Now we make the inductive hypothesis that

$$\begin{aligned} f[x_0, \dots, x_{n-1}] &= \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-2}} dt_{n-1} \\ &\times f^{(n-1)}(t_{n-1}[x_{n-1} - x_{n-2}] + \cdots + t_1[x_1 - x_0] + x_0). \end{aligned}$$

In the integral in  $(10)_n$  we replace the integration variable  $t_n$  by

$$\begin{aligned} \xi &= t_n[x_n - x_{n-1}] + \cdots + t_1[x_1 - x_0] + x_0, \\ dt_n &= d\xi/[x_n - x_{n-1}]. \end{aligned}$$

The corresponding limits become

$$\begin{aligned} (t_n = 0) \rightarrow (\xi = \xi_0 &\equiv t_{n-1}[x_{n-1} - x_{n-2}] + \cdots + t_1[x_1 - x_0] + x_0), \\ (t_n = t_{n-1}) \rightarrow (\xi = \xi_1 &\equiv t_{n-1}[x_n - x_{n-2}] \\ &+ t_{n-2}[x_{n-2} - x_{n-3}] + \cdots + t_1[x_1 - x_0] + x_0). \end{aligned}$$

Now the innermost integral in  $(10)_n$  is, since  $x_n \neq x_{n-1}$ ,

$$\begin{aligned} \int_0^{t_{n-1}} dt_n f^{(n)}(t_n[x_n - x_{n-1}] + \cdots + t_1[x_1 - x_0] + x_0) \\ &= \int_{\xi_0}^{\xi_1} f^{(n)}(\xi) \frac{d\xi}{x_n - x_{n-1}} \\ &= \frac{f^{(n-1)}(\xi_1) - f^{(n-1)}(\xi_0)}{x_n - x_{n-1}}. \end{aligned}$$

However, by applying the inductive hypothesis we have

$$\begin{aligned} \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-2}} dt_{n-1} &\frac{f^{(n-1)}(\xi_1) - f^{(n-1)}(\xi_0)}{x_n - x_{n-1}} \\ &= \frac{f[x_0, \dots, x_{n-2}, x_n] - f[x_0, \dots, x_{n-2}, x_{n-1}]}{x_n - x_{n-1}} \\ &= f[x_0, \dots, x_n]. \end{aligned}$$
■

Notice that the integrand on the right-hand side of (10) is a continuous function of the  $n + 1$  variables  $x_0, x_1, \dots, x_n$ , and hence the right-hand side is a continuous function of these variables. Thus (10) defines uniquely the continuous extension of  $f[x_0, x_1, \dots, x_n]$  when the arguments lie in any interval of continuity of the  $n$ th derivative of  $f(x)$ . That is, since the divided differences have only been defined for distinct sets of arguments [see (1) and the discussion leading to it], we are at liberty to define them when some of the arguments are not distinct. Naturally, we do this in a manner which maintains, if possible, the above observed continuity. If  $f^{(n)}(x)$  is continuous, then Theorem 2 shows how this can be done for all differences of  $f(x)$  of orders  $0, 1, \dots, n$ . These remarks can be summarized as

**COROLLARY 1.** *Let  $f^{(n)}(x)$  be continuous in  $[a, b]$ . For any set of points  $x_0, x_1, \dots, x_k$  in  $[a, b]$  with  $k \leq n$  let  $f[x_0, x_1, \dots, x_k]$  be given by (10) <sub>$k$</sub> . The divided difference thus defined is a continuous function of its  $k + 1$  arguments in  $[a, b]$  and coincides with that defined by (4), or (6), when the arguments are distinct.* ■

In fact, as in the proof of the First Mean Value Theorem for integrals, (10) <sub>$n$</sub>  yields

$$\begin{aligned} m \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n &\leq f[x_0, \dots, x_n] \\ &\leq M \int_0^1 dt_1 \int_0^{t_1} dt_2 \cdots \int_0^{t_{n-1}} dt_n \end{aligned}$$

where  $m \equiv \min f^{(n)}(x)$  and  $M \equiv \max f^{(n)}(x)$  for  $x$  in

$$\min (x_0, \dots, x_n) \leq x \leq \max (x_0, \dots, x_n).$$

Then by the continuity of  $f^{(n)}$  there is a point  $\mu_n$  in this interval such that

$$f[x_0, \dots, x_n] = \frac{f^{(n)}(\mu_n)}{n!}.$$

Hence, we have established a generalization of Theorem 1, since the points  $x_i$  need not be distinct, in

**COROLLARY 2.** *If  $f^{(n)}(x)$  is continuous in  $[a, b]$  and  $x_0, x_1, \dots, x_n$  are in  $[a, b]$ , then*

$$(11) \quad f[x_0, x_1, \dots, x_n] = \frac{f^{(n)}(\xi)}{n!},$$

where

$$\min (x_0, x_1, \dots, x_n) \leq \xi \leq \max (x_0, x_1, \dots, x_n). \quad \blacksquare$$

A particular case is

**COROLLARY 3.** *If  $f^{(n)}(x)$  is continuous in a neighborhood of  $x$ , then*

$$(12) \quad f[\underbrace{x, x, \dots, x}_{n+1 \text{ terms}}] = \frac{f^{(n)}(x)}{n!}.$$

■

Now, we can deduce yet another representation of the divided difference, when several multiplicities† occur.

**COROLLARY 4.** *If  $f^{(n)}(x)$  is continuous in  $[a, b]$ ,  $y_0, y_1, \dots, y_n$  are in  $[a, b]$ , and  $x$  is distinct from any  $y_i$ , then*

$$(13) \quad f[x, y_0, y_1, \dots, y_n] = \frac{f[x, y_1, \dots, y_n] - f[y_0, y_1, \dots, y_n]}{x - y_0},$$

*gives the unique continuous extension of the definition of divided difference.*

*Proof.* By Theorem 2, the right-hand side of (13) is uniquely defined and continuous in  $(y_0, y_1, \dots, y_n)$  for  $y_0 \neq x$ . Hence, the left-hand side is the unique continuous extension of the definition of divided difference no matter what multiplicities occur in  $(y_0, y_1, \dots, y_n)$ . Observe that only the continuity of the  $n$ th derivative is used. ■

**COROLLARY 5.** *If  $x_i \neq y_j$ , for  $0 \leq i \leq p$ ,  $0 \leq j \leq q$ ;  $f^{(m)}(x)$  continuous in  $[a, b]$ ;  $\{x_i\}$ ,  $\{y_i\}$  in  $[a, b]$ ;  $0 \leq p, q \leq m$ , then*

$$(14) \quad \begin{aligned} f[x_0, \dots, x_p, y_0, \dots, y_q] &= g[x_0, \dots, x_p] \\ &= h[y_0, \dots, y_q] \end{aligned}$$

where

$$g(x) \equiv f[x, y_0, \dots, y_q], \quad h(y) \equiv f[x_0, \dots, x_p, y],$$

*provides the unique continuous extension of definition of divided difference.*

*Proof.* By (13),  $g(x)$  has  $m$  continuous derivatives for  $x \neq y_i$ ,  $0 \leq i \leq q$ . Therefore, by the theorem,  $g[x_0, \dots, x_p]$  is defined and continuous in  $x_0, \dots, x_p$  if  $x_i \neq y_j$ , as postulated. Furthermore,  $g(x)$  is continuous as a function of the parameters  $(y_0, \dots, y_q)$  if  $x \neq y_i$ , by Corollary 4. Hence, the representation (10)<sub>p</sub> of  $g[x_0, \dots, x_p]$  yields the continuity of  $g[x_0, \dots, x_p]$  with respect to all variables  $(x_0, \dots, x_p, y_0, \dots, y_q)$  provided merely that  $x_i \neq y_j$ .

Now the function  $f[x_0, \dots, x_p, y_0, \dots, y_q]$  as defined in (14) is continuous (if  $x_i \neq y_j$ ) in its variables; hence we conclude that (14) is the unique continuous extension, since (14) is valid when the arguments are all distinct. ■

† The conclusions of Corollaries 4–7 that follow, concern continuity properties of and representations for divided differences that are easily established when no multiplicities occur among the arguments. When multiplicities do occur, the corollaries establish the same representations under the hypothesis of minimal differentiability of  $f(x)$ .

**COROLLARY 6.** *If  $f(x)$  has a continuous derivative of order  $m$  in  $[a, b]$ ;  $x_0, \dots, x_p, y_0, \dots, y_q, z_0, \dots, z_r$  are in  $[a, b]$ ;  $x_i \neq y_j, x_i \neq z_k, y_j \neq z_k$  for all  $i, j, k; 0 \leq p, q, r \leq m$ ; then*

$$(15) \quad f[x_0, \dots, x_p, y_0, \dots, y_q, z_0, \dots, z_r]$$

$$= \frac{1}{p! q! r!} \frac{\partial^p}{\partial x^p} \frac{\partial^q}{\partial y^q} \frac{\partial^r}{\partial z^r} f[x, y, z] \Big|_{(\xi, \eta, \zeta)}$$

where

$$\min(x_0, \dots, x_p) \leq \xi \leq \max(x_0, \dots, x_p),$$

$$\min(y_0, \dots, y_q) \leq \eta \leq \max(y_0, \dots, y_q),$$

$$\min(z_0, \dots, z_r) \leq \zeta \leq \max(z_0, \dots, z_r).$$

*Proof.* Let

$$(16) \quad \begin{aligned} g(x) &\equiv f[x, y_0, \dots, y_q, z_0, \dots, z_r] \\ h(y) &\equiv f[x, y, z_0, \dots, z_r] \\ k(z) &\equiv f[x, y, z]. \end{aligned}$$

By Corollaries 2 and 5 [appropriately generalized for sets of variables ( $\{x_i\}$ ,  $\{y_j\}$ ,  $\{z_k\}$ )], we have

$$(17) \quad f[x_0, \dots, x_p, y_0, \dots, y_q, z_0, \dots, z_r] = g[x_0, \dots, x_p]$$

$$(18) \quad g[x_0, \dots, x_p] = \frac{1}{p!} \frac{\partial^p}{\partial x^p} g(x) \Big|_{x=\xi}$$

$$(19) \quad g(x) = h[y_0, \dots, y_q] = \frac{1}{q!} \frac{\partial^q}{\partial y^q} h(y) \Big|_{y=\eta}$$

$$(20) \quad h(y) = k[z_0, \dots, z_r] = \frac{1}{r!} \frac{\partial^r}{\partial z^r} k(z) \Big|_{z=\zeta}.$$

The conclusion (15) follows from (17), (18), (19), and (20). ■

A special case is contained in

**COROLLARY 7.** *If  $f^{(m)}(x)$  is continuous in  $[a, b]$ ;  $x, y, z$  are distinct points in  $[a, b]$ ;  $0 \leq p, q, r \leq m$ ; then*

$$(21) \quad f[\underbrace{x, \dots, x}_{p+1}, \underbrace{y, \dots, y}_{q+1}, \underbrace{z, \dots, z}_{r+1}] = \frac{1}{p! q! r!} \frac{\partial^p}{\partial x^p} \frac{\partial^q}{\partial y^q} \frac{\partial^r}{\partial z^r} f[x, y, z]. \quad ■$$

We leave to Problems 3, 4, 5, 6, and 7, the independent proof of some simple differentiability properties of divided differences, which are needed later.

## PROBLEMS, SECTION 1

1. Deduce the representation (4) for divided differences directly from the expression (1).

[Hint: Use induction and the form (2) for  $Q_{k+1}(x_k)$ .]

2. If  $P_n(x)$  is a polynomial of degree  $n$ , show that  $P_n[x_0, x]$  for  $x \neq x_0$  is a polynomial of degree at most  $n - 1$  in  $x$ .

3. Prove the following

**LEMMA 1.** *If  $x, x_0, \dots, x_n$  are  $n + 2$  distinct points then*

$$f[x_0, \dots, x_n, x] = \sum_{j=0}^n \frac{f[x, x_j]}{\prod_{\substack{k=0 \\ (k \neq j)}}^n (x_j - x_k)}.$$

[Hint: Use equation (8) and the Lagrange form of the interpolation polynomial.] This is another representation of the divided differences which is very useful in deriving their continuity and differentiability properties.

4. Prove directly the following:

**THEOREM 3.** *If  $f(x) \in C[a, b]$  and  $f'(x_j)$  exists for some fixed  $x_j \in [a, b]$ , then  $f[x, x_j]$  is a continuous function of  $x$  in  $[a, b]$  if we assign to it the value at  $x = x_j$ :  $f[x_j, x_j] = f'(x_j)$ .*

5. Prove the following:

**THEOREM 4.** *Let  $f'(x) \in C[a, b]$  and  $f''(x)$  be continuous in an (arbitrarily small) interval about some fixed  $x_j \in [a, b]$ . Then  $df[x, x_j]/dx$  is a continuous function of  $x$  in  $[a, b]$ .*

[Hint: Form  $df[x, x_j]/dx$  for  $x \neq x_j$ ; use the Taylor expansion about  $x_j$  and take limits as  $x = x_j \pm h \rightarrow x_j$ .]

6. Use the results of Problems 3, 4, and 5 to state and prove, if  $(x_0, x_1, \dots, x_n)$  are distinct and in  $[a, b]$ ,

- (i) a theorem on the continuity of  $f[x_0, \dots, x_n, x]$  for  $x \in [a, b]$ ;
- (ii) a theorem on the continuity of  $(d/dx)f[x_0, \dots, x_n, x]$  for  $x \in [a, b]$ .

7. By using the theorem under (i) of Problem 6, note that  $f[x_0, \dots, x_n, x_n] = \lim_{h \rightarrow 0} f[x_0, \dots, x_n, x_n + h]$ . Therefore, show that

$$f[x_0, \dots, x_n, x_n] = \frac{d}{dx_n} f[x_0, \dots, x_n].$$

Prove that this representation is valid under the conditions:  $f'(x_n)$  is defined and  $x_0, x_1, \dots, x_n$  are distinct.

[Hint: use the lemma of Problem 3 and the formula (6).]

8. Prove the symmetry of the divided difference by constructing the  $Q_n(x)$  in (2) using the points  $(x_0, x_1, \dots, x_n)$  in an arbitrary permuted order  $(x_{j_0}, x_{j_1}, \dots, x_{j_n})$ . [This is a generalization of what was done in deriving (2') and proving  $a_n = b_n$ .]

9. Verify equation (6) (the divided difference property) directly from equation (4).

- 10.\* (*Osculatory interpolation*). If  $f(x)$  and its derivatives of order  $r_0 - 1, r_1 - 1, \dots, r_n - 1$  are defined respectively at the distinct points  $(x_0, x_1, \dots, x_n)$

in  $[a, b]$ , then there exists a unique polynomial  $Q_N(x)$  of degree at most  $N$ , where  $N = \sum_{k=0}^n r_k - 1$ , such that  $Q_N^{(k)}(x_j) = f^{(k)}(x_j)$  for  $k = 0, 1, \dots, r_j - 1$ ,  $j = 0, 1, \dots, n$ . The special case  $r_0 = r_1 = \dots = r_n = 2$  has been studied in Chapter 5, Subsection 2.2.

[Hint: Show that the Newton form of the polynomial may be derived by satisfying successively all of the  $r_0$  conditions at  $x_0$  before proceeding to satisfy successively the  $r_1$  conditions at  $x_1$ , etc.]

Arrive at the scheme

$$Q_0(x) \equiv f(x_0)$$

$$Q_s(x) \equiv Q_{s-1}(x) + b_s(x - x_0)^{s_0}, \quad \text{for } 1 \leq s < s_0 \text{ where } s_0 = r_0$$

$$Q_s(x) \equiv Q_{s-1}(x) + b_s \left[ \prod_{k=0}^{j-1} (x - x_k)^{r_k} \right] (x - x_j)^{s - s_{j-1}},$$

$$\text{for } s_{j-1} \leq s < s_j, \quad \text{with } j = 1, 2, \dots, n, \quad \text{where } s_j = \sum_{k=0}^j r_k$$

Show that the  $b_s$  may be recursively defined. The proof of uniqueness may be based on the fact that if a polynomial has degree at most  $N$  and at least  $N + 1$  zeroes (counting multiplicities), then the polynomial is identically zero.]

**11.\*** If  $f^{(p)}(x)$  is continuous in  $[a, b]$ ,  $0 \leq r_i - 1 \leq p$ ,  $\{x_i\}$  in  $[a, b]$ , then show that the coefficients  $b_s$  of Problem 10 are divided differences of  $f(x)$  of order  $s$ , based on the first  $s + 1$  arguments in the sequence  $x_0, x_0, \dots, x_0; x_1, x_1, \dots$ , where each  $x_i$  appears  $r_i$  times (the divided differences have been defined in Theorem 2, Corollary 5).

**12.\*** Verify that the error in osculatory interpolation (see Problem 10) is

$$R_N(x) \equiv f(x) - Q_N(x) = \frac{\prod_{i=0}^n (x - x_i)^{r_i}}{(N + 1)!} f^{(N+1)}(\xi),$$

if  $f$  has  $N + 1$  continuous derivatives in  $[a, b]$ , where  $\xi$  is in  $[a, b]$ .

**13.** Given the values

$$\begin{array}{ll} \sin(1.6) = .9995736030 & \cos(1.6) = -.0291995223 \\ \sin(1.7) = .9916648105 & \cos(1.7) = -.1288444943 \end{array}$$

approximate  $\sin(1.65)$  to seven decimal places by evaluating Taylor's series (about  $x = 1.6$ ) including the third derivative term. Estimate the error by examining the remainder term in the formula. Calculate  $\sin(1.65)$  correct to 9 decimal places and verify that the above estimate of error is correct.

**14.** Use the table in Problem 13 and calculate  $\sin(1.65)$  by linear interpolation. Verify that the magnitude of the error is consistent with the remainder term as given by (8) and (9), or equivalently by equation (2.9) of Chapter 5.

**15.** Construct a table of divided differences from the values given in Problem 13 for the function  $\sin x$  with the repeated arguments  $(1.6)(1.6)(1.7)(1.7)$ ; and find the Newton form of the osculating polynomial of degree 3. Calculate  $\sin(1.65)$  by evaluating the osculating polynomial, and verify that the magnitude of the error is explained by the formula in Problem 12.

Table for Problem 15

	$f(x)$	$f[x, x]$	$f[x, x, x]$	$f[x, x, x, x]$
(1.6)	$\sin(1.6)$	$\cos(1.6)$		
(1.6)	$\sin(1.6)$		$10(\sin 1.7 - \sin 1.6)$	
(1.7)	$\sin(1.7)$		$\cos(1.7)$	
(1.7)	$\sin(1.7)$			

16. Given the repeated arguments  $x_0, x_0, x_0, x_1, x_1$  and the values  $f^{(p)}(x_i)$  in the accompanying divided difference table. Complete the table and verify that the fourth order difference has the value given by (21) if  $x_1 = x_0 + h$ .

Table for Problem 16

	$f[x]$	$f[x, x]$	$f[x, x, x]$	$f[x, x, x, x]$	$f[x, x, x, x, x]$
$x_0$	$f_0$				
		$f_0'$			
$x_0$	$f_0$		$\frac{f_0''}{2}$		
		$f_0'$			
$x_0$	$f_0$		$\frac{f_{01} - f_0'}{h}$		
		$f_{01}$			
$x_1$	$f_1$		$\frac{f_1' - f_{01}}{h}$		
		$f_1'$			
$x_1$	$f_1$				

17. Given the  $m + n + p + 3$  points  $(x_0, x_1, \dots, x_m, y_0, y_1, \dots, y_n, z_0, z_1, \dots, z_p)$  and  $f(x)$  which has derivatives of order  $(m, n, p)$  respectively, in a neighborhood of each of the distinct points  $(x, y, z)$ . Show that  $f[x_0, \dots, z_p] = g^{(m)}(x) + h^{(n)}(y) + k^{(p)}(z)$  if  $x_i = x, y_j = y, z_r = z$  for all  $i, j, r$ , where

$$g(x) \equiv \frac{f(x)}{\prod_{j=0}^n (x - y_j) \prod_{r=0}^p (x - z_r)},$$

$$h(y) \equiv \frac{f(y)}{\prod_{i=0}^m (y - x_i) \prod_{r=0}^p (y - z_r)},$$

$$k(z) \equiv \frac{f(z)}{\prod_{i=0}^m (z - x_i) \prod_{j=0}^n (z - y_j)}.$$

[Hint:  $f[x_0, x_1, \dots, z_{p-1}, z_p] = g[x_0, x_1, \dots, x_m] + h[y_0, y_1, \dots, y_n] + k[z_0, z_1, \dots, z_p]$  for distinct points  $(x_0, \dots, z_p)$ . Therefore,

$$f[x_0, \dots, x_m, y_0, \dots, y_n, z_0, \dots, z_p] = g^{(m)}(\xi) + h^{(n)}(\eta) + k^{(p)}(\zeta).$$

Now let  $x_i \rightarrow x$ ,  $y_j \rightarrow y$ ,  $z_r \rightarrow z$ , all  $i, j, r.$ ]

## 2. ITERATIVE LINEAR INTERPOLATION

The Newton form of the interpolation polynomial permits one to increase easily the accuracy (actually the degree) of the approximating polynomial. It has many important applications and is indeed well suited for computations when the data are available in the appropriate tabular form. However, it can be viewed as one of a class of methods for generating successively higher order interpolation polynomials which we consider briefly. These procedures are iterative and can be very effectively employed on modern digital computers since they are based on successive linear interpolations.

The lemma underlying the *iterative linear interpolation* schemes can be stated as

**LEMMA 1.** *Let  $x_{i_1}, x_{i_2}, \dots, x_{i_n}$  be  $n$  distinct points and denote by  $P_{i_1, i_2, \dots, i_n}(x)$  the interpolation polynomial of degree  $n - 1$  such that*

$$P_{i_1, i_2, \dots, i_n}(x_{i_v}) = f(x_{i_v}), \quad v = 1, 2, \dots, n.$$

*Then if  $x_j$ ,  $x_k$ , and  $x_{i_v}$ ,  $v = 1, 2, \dots, m$  are any  $m + 2$  distinct points*

$$(1) \quad P_{i_1, i_2, \dots, i_m, j, k}(x)$$

$$\equiv \frac{(x - x_k)P_{i_1, i_2, \dots, i_m, j}(x) - (x - x_j)P_{i_1, i_2, \dots, i_m, k}(x)}{x_j - x_k},$$

for  $m = 0, 1, 2, \dots$

*Proof.* We establish (1) by observing that the right-hand side defines a polynomial of degree at most  $m + 1$  which takes on the values  $f(x_{i_v})$  at  $x_{i_v}$  for  $v = 1, \dots, m$ ,  $f(x_j)$  at  $x_j$ , and  $f(x_k)$  at  $x_k$ . Hence, the polynomial on the right-hand side of (1) is the unique interpolation polynomial which appears on the left-hand side of (1). ■

The variety of schemes which employ Lemma 1 to determine successively higher order interpolation polynomials differ in the order in which the pairs of values  $(x_j, f(x_j))$  are used. For many applications, particularly on digital computers, the function values are generated sequentially and it may not be known in advance how many values are to be generated.

Table 1

$x_0$	$f(x_0)$					
$x_1$	$f(x_1)$	$P_{0, 1}(x)$				
$x_2$	$f(x_2)$	$P_{1, 2}(x)$	$P_{0, 1, 2}(x)$	.	.	.
:	:	:	:	.	.	.
$x_k$	$f(x_k)$	$P_{k-1, k}(x)$	$P_{k-2, k-1, k}(x)$	$\dots$	$P_{0, 1, \dots, k}(x)$	

For such cases we may employ always the latest pair of values and compute each row sequentially to find the array in Table 1. Any  $P_{\dots}(x)$  in this scheme is obtained by using the two quantities to its left and diagonally above. Thus, to determine, say, the  $(k + 1)$ st row, only the  $k$ th row need be retained. Of course, as more points are generated, rows of greater length must be saved. If it is known in advance that a fixed number, say  $k + 1$ , of function values is to be generated, then a different order of computing is appropriate. That is, we may compute by columns and when a particular column has been evaluated the preceding column may be discarded. The schemes based on Table 1 are known as *Neville's iterated interpolation*.

Another sequence of interpolants are used in *Aitken's iterative interpolation* as is indicated in Table 2. Again, computation by columns is

Table 2

$x_0$	$f(x_0)$					
$x_1$	$f(x_1)$	$P_{0, 1}(x)$				
$x_2$	$f(x_2)$	$P_{0, 2}(x)$	$P_{0, 1, 2}(x)$	.	.	.
:	:	:	:	.	.	.
$x_k$	$f(x_k)$	$P_{0, k}(x)$	$P_{0, 1, k}(x)$	$\dots$	$P_{0, 1, \dots, k}(x)$	

appropriate for a known fixed value of  $k$ . Note that the  $(k + 1)$ st row can be computed if we save only the "diagonal" elements  $f(x_0), P_{01}(x), \dots, P_{0, 1, \dots, k}(x)$ . In brief, the basic difference between these two procedures is that in Aitken's, the interpolants on the row with  $x_k$  use points with subscripts nearest 0, while in Neville's they use points with subscripts nearest  $k$ , as we read the entries from left to right.

A particularly important application of Neville's method is to what is called *iterative inverse interpolation*. Given  $y = f(x)$  we define the inverse function,  $x = g(y)$ , such that  $y = f(g(y))$  and  $x = g(f(x))$ . Then it is desired to find a particular value of  $x$ , say  $x = \bar{x}$ , such that  $f(\bar{x}) = \bar{y}$ .

Thus, we need only find the value of  $g(\bar{y})$ . Given pairs of values  $(x_i, f(x_i)) = (g(y_i), y_i)$ , interpolation may be used to approximate  $g(\bar{y})$ . As additional pairs  $(x_j, f(x_j))$  become available, better or rather higher order interpolation can be obtained by using Neville's scheme. Now, however, in Table 1, the first two columns are interchanged and the argument of the polynomials is  $f$  (or  $y$ ). The computation should be done by rows. After several steps it is advisable to use only a few of the row elements and to discard those rows involving values  $x_i$  which were "early" iterates and thus presumably not sufficiently close to the desired value,  $\bar{x}$ . It should be noted that this procedure is essentially one of the generalizations of the method of false position or of Aitken's  $\delta^2$ -method (see Chapter 3, Subsections 2.3 and 2.4). When the evaluation of the function  $f(x)$  is not a difficult task, most workers prefer to use only a single linear interpolation at each stage, where the values  $[x_k, f(x_k)]$  and  $[x_{k+1}, f(x_{k+1})]$  are the most recent, i.e., *regula falsi*. Of course, inverse interpolation, in general, is meaningful only if  $f^{-1}(x)$  is defined as a single-valued function over the interval in question.

## PROBLEMS, SECTION 2

- Given the accompanying table for  $\sin(x)$ , interpolate for  $\sin(1.65)$  by determining the value  $P_3(1.65)$  of the Lagrange interpolation polynomial with the use of both the Neville and the Aitken schemes of successive linear interpolations. (Make up Tables corresponding to Tables 1 and 2.)

**Table for Problem 1**

---

$\sin(1.5) = .9974949866$
$\sin(1.6) = .9995736030$
$\sin(1.7) = .9916648105$
$\sin(1.8) = .9738476309$

---

- Evaluate  $\sin(1.65)$  by finding the Newton form of  $P_3(1.65)$ . Compare the amount of work in Problems 1 and 2, when done by hand.
- FORWARD DIFFERENCES AND EQUALLY SPACED INTERPOLATION POINTS**

Many of the results in Section 1 are simplified and additional important consequences are obtained if the points of interpolation are equally spaced. Very many, if not most, of the practical applications are with

such sets of points. Thus, we take  $x_0$  to be an arbitrary fixed point and let  $h > 0$  be the spacing between adjacent points. Then the points to be considered are

$$(1) \quad x_j = x_0 + jh; \quad j = 0, \pm 1, \pm 2, \dots$$

Note that  $x_j - x_k = (j - k)h$  and, since  $j$  may be negative,  $x_0$  need not be an endpoint of the interval under consideration.

Associated with equally spaced points is the (first order) *forward difference* which is defined by

$$(2) \quad \Delta f(x) \equiv f(x + h) - f(x).$$

Higher order differences are defined in the obvious way as

$$(3) \quad \Delta^{n+1}f(x) = \Delta^n[\Delta f(x)] = \Delta^n f(x + h) - \Delta^n f(x); \quad n = 1, 2, \dots$$

In analogy with the divided difference table, we can easily construct the higher order forward differences for the points (1) by means of Table 1.

Table 1

$x$	$f(x)$	$\Delta$	$\Delta^2$	$\Delta^3$	$\dots$
$x_0$	$f(x_0)$				
$x_1$	$f(x_1)$	$f(x_1) - f(x_0) = \Delta f(x_0)$	$\Delta f(x_1) - \Delta f(x_0) = \Delta^2 f(x_0)$	$\Delta^2 f(x_1) - \Delta^2 f(x_0) = \Delta^3 f(x_0)$	$\dots$
$x_2$	$f(x_2)$	$f(x_2) - f(x_1) = \Delta f(x_1)$	$\Delta f(x_2) - \Delta f(x_1) = \Delta^2 f(x_1)$	$\Delta^2 f(x_2) - \Delta^2 f(x_1) = \Delta^3 f(x_1)$	$\dots$
$x_3$	$f(x_3)$	$f(x_3) - f(x_2) = \Delta f(x_2)$	$\Delta f(x_3) - \Delta f(x_2) = \Delta^2 f(x_2)$	$\vdots$	
$x_4$	$f(x_4)$	$f(x_4) - f(x_3) = \Delta f(x_3)$			
$\vdots$	$\vdots$				

A relation between divided differences with equally spaced arguments and forward differences is easily obtained from the representation (1.6) and the definition (3). Thus by taking  $x$  to be any one of the points  $x_j$  of (1), say  $x_0$ , we have

$$\Delta f(x_0) = f(x_1) - f(x_0) = (x_1 - x_0) \frac{f(x_1) - f(x_0)}{x_1 - x_0} = h f[x_0, x_1].$$

Now to proceed by induction on  $n$ , the order of the difference, we assume that

$$(4) \quad \Delta^n f(x_0) = n! h^n f[x_0, x_1, \dots, x_n],$$

and obtain

$$\begin{aligned}
 \Delta^{n+1}f(x_0) &= \Delta^n f(x_1) - \Delta^n f(x_0) \\
 &= n! h^n f[x_1, x_2, \dots, x_{n+1}] - n! h^n f[x_0, x_1, \dots, x_n] \\
 &= n! h^n (x_{n+1} - x_0) \frac{f[x_1, x_2, \dots, x_{n+1}] - f[x_0, x_1, \dots, x_n]}{(x_{n+1} - x_0)} \\
 &= (n + 1)! h^{n+1} f[x_0, x_1, \dots, x_{n+1}].
 \end{aligned}$$

Thus (4) is established for all  $n \geq 1$ .

Another representation of the forward differences can now be obtained by specializing the representation (1.4) to equally spaced points. We note that, by (1)

$$\begin{aligned}
 \prod_{\substack{j=0 \\ (j \neq i)}}^n (x_i - x_j) &= \prod_{\substack{j=0 \\ (j \neq i)}}^n (i - j)h \\
 &= h^n \prod_{j=0}^{i-1} (i - j) \prod_{l=i+1}^n (i - l) \\
 &= (-1)^{n-i} h^n (i)! (n - i)!
 \end{aligned}$$

By using this result in (1.4) we obtain

$$(5) \quad f[x_0, x_1, \dots, x_n] = \frac{1}{n! h^n} \sum_{i=0}^n (-1)^{n-i} \binom{n}{i} f(x_i),$$

where

$$\binom{n}{i} = \frac{n!}{i! (n - i)!}$$

are the usual binomial coefficients and  $0! = 1$ . From (5) in (4) there results

$$(6) \quad \Delta^n f(x_0) = \sum_{i=0}^n (-1)^{n-i} \binom{n}{i} f(x_i).$$

A final expression for the forward differences is obtained by using (4) in (1.9) with  $x = x_k$  to get

$$(7) \quad \Delta^n f(x_0) = h^n f^{(n)}(\xi); \quad x_0 < \xi < x_n.$$

Of course, (7) is valid assuming that  $f(x)$  has an  $n$ th derivative in the indicated interval.

It is clear, from (7), that the  $n$ th forward difference of a polynomial of degree  $n$  is a constant and that higher order differences vanish. More generally, if  $f(x)$  has all derivatives bounded, say  $|f^{(n)}(x)| \leq M$  for all  $n$ , then (7) implies that

$$|\Delta^n f| \leq Mh^n.$$

Thus if  $h < 1$  the magnitude of  $n$ th order differences of  $f(x)$  will in general decrease as  $n$  increases. On the other hand, if the  $n$ th derivative of  $f(x)$  grows with  $n$ , the  $n$ th difference will decrease only if  $h$  is “sufficiently small.” This may be illustrated by the function  $f(x) = e^{\alpha x}$  where  $\alpha > 0$ . Clearly,  $f^{(n)}(x) = \alpha^n e^{\alpha x}$  and using (7)  $\Delta^n f(x_0) = h^n \alpha^n e^{\alpha \xi}$ . If the interpolation points are to be confined to the interval  $x_0 \leq x \leq x_0 + L$ , then  $0 \leq nh \leq L$  and the differences will generally decrease only if  $h < 1/\alpha$  (we here neglect the variation in  $e^{\alpha \xi}$  which may occur). Finally, if  $f^{(n)}(x)$  is not bounded for all  $n$  we can expect  $|\Delta^n f|$  to decrease, for sufficiently small  $h$ , only for the first several values of  $n$ . This heuristic observation is the basis of a method for detecting isolated errors in forward difference tables of supposedly smooth functions.

To describe this method we first observe that

$$\Delta^n[f(x) + g(x)] = \Delta^n f(x) + \Delta^n g(x).$$

Now suppose that  $f(x)$  is a smooth function, say for simplicity with all derivatives bounded, and that in tabulating this function an error of amount  $\delta$  is made in the single entry  $f(x_j)$ . Thus the function actually tabulated can be written as  $f(x) + g(x)$  where

$$g(x_i) = \begin{cases} 0, & i \neq j; \\ \delta, & i = j. \end{cases}$$

Applying the representation (6) we see that the column of  $n$ th differences of  $g$  will contain quantities of the form

$$(-1)^{n-k} \binom{n}{k} \delta.$$

Thus in examining the higher differences of the tabulated data  $f(x) + g(x)$ , since  $\Delta^n f$  decreases with  $n$ , an error will be apparent if the entries, from some column on, alternate in sign and the magnitudes of these varying entries are proportional to the appropriate binomial coefficients. The power of this method is illustrated in Problem 2.

This procedure will not be practical if the isolated error  $\delta$  is of the same order of magnitude as the roundoff errors generally present in all of the tabular data. In fact, we shall see that if roundoff errors are present, differences of a sufficiently high order may have no significance. Let the tabular entries be,  $f(x) + \rho(x)$ , where  $\rho(x)$  is the rounding error.

Let  $\epsilon$  be a bound on the rounding error, i.e.,  $|\rho(x)| \leq \epsilon$ . The worst possibility for the distribution of these errors is

$$(8) \quad \rho(x_i) \equiv (-1)^i \epsilon.$$

Table 2

$\rho(x)$	$\Delta$	$\Delta^2$	$\Delta^3$	$\Delta^4$
$\epsilon$				
	$-2\epsilon$			
$-\epsilon$		$4\epsilon$		
	$2\epsilon$		$-8\epsilon$	
$\epsilon$		$-4\epsilon$		$16\epsilon$
	$-2\epsilon$		$8\epsilon$	
$-\epsilon$		$4\epsilon$		
	$2\epsilon$			

This is made clear by the table of differences for such a distribution (Table 2). Any other distribution leads to some entries which would be less in absolute value than the corresponding entries above. From Table 2 we see that

$$|\Delta^n \rho(x_j)| = 2^n \epsilon.$$

[This result may be easily proved for  $j = 0$  by using (8) in (6).] Thus the roundoff error present in the  $n$ th difference is at most  $2^n \epsilon$ . If there are  $s$  decimals retained with a roundoff error of at most one-half unit in the last place then  $\epsilon = \frac{1}{2} 10^{-s}$ , and the bound on the roundoff error in the  $n$ th difference becomes  $2^{n-1} 10^{-s}$ .

### 3.1. Interpolation Polynomials and Remainder Terms for Equidistant Points

The Lagrange and Newton forms of the interpolation polynomials become simplified when the interpolation points are equally spaced. To be consistent with the notation (1) we introduce a new independent variable,  $t$ , by setting

$$(9) \quad x = x_0 + th.$$

Thus,  $t$  measures  $x - x_0$  in units of  $h$  and is an integer only at the points  $x_j$  of (1).

Now the Lagrange interpolation coefficients, (2.6) of Chapter 5, can be written as

$$\begin{aligned} \phi_{n,j}(x) &= \phi_{n,j}(x_0 + th) = \prod_{\substack{k=0 \\ (k \neq j)}}^n \frac{x - x_k}{x_j - x_k} \\ &= \prod_{\substack{k=0 \\ (k \neq j)}}^n \frac{t - k}{j - k} \\ &= \frac{(-1)^{n-j}}{n!} \binom{n}{j} \prod_{\substack{k=0 \\ (k \neq j)}}^n (t - k). \end{aligned}$$

It is convenient to introduce

$$(10) \quad \begin{aligned} \pi_0(t) &\equiv t, \\ \pi_n(t) &\equiv t(t - 1)\cdots(t - n), \quad n = 1, 2, \dots. \end{aligned}$$

The function  $\pi_n(t)$  is a polynomial in  $t$  of degree  $n + 1$  and is frequently called the  $(n + 1)$ st *factorial polynomial*. In terms of this polynomial, the Lagrange interpolation coefficients become

$$\phi_{n,j}(x_0 + th) = \frac{\pi_n(t)}{n!} \binom{n}{j} \frac{(-1)^{n-j}}{t - j}.$$

Using this formula in (2.5) of Chapter 5, the Lagrange form of the interpolation polynomial simplifies to

$$(11) \quad P_n(x_0 + th) = \frac{\pi_n(t)}{n!} \sum_{j=0}^n (-1)^{n-j} \binom{n}{j} \frac{f(x_j)}{t - j}.$$

Newton's form of the interpolation polynomial is simplified by using (4), (9), and (10) in (1.7) to get

$$(12) \quad \begin{aligned} Q_n(x_0 + th) &= f(x_0) + \frac{\pi_0(t)}{1!} \Delta f(x_0) + \frac{\pi_1(t)}{2!} \Delta^2 f(x_0) + \cdots \\ &\quad + \frac{\pi_{n-1}(t)}{n!} \Delta^n f(x_0). \end{aligned}$$

The error in these interpolation polynomials is, by (1.8), (9), (10), and Theorem 1.1,

$$(13) \quad \begin{aligned} R_n(x) &= R_n(x_0 + th) = \pi_n(t) h^{n+1} f[x_0, \dots, x_n, x] \\ &= \pi_n(t) h^{n+1} \frac{f^{(n+1)}(\xi)}{(n+1)!}. \end{aligned}$$

$R_n(x)$  may also be called the *remainder* for the interpolation formula. Of course, the final form is valid only if  $f(x)$  has  $n + 1$  derivatives in the interval containing  $x$ ,  $x_0$  and  $x_0 + nh$ . To obtain a clear idea of the behavior of this error, we shall study some properties of the factorial polynomials  $\pi_n(t)$ .

From the definition (10) it is clear that  $\pi_n(t)$  has  $n + 1$  real roots and they are at  $t = 0, 1, \dots, n$ . These polynomials have certain symmetries about the point  $t = n/2$  which is the midpoint of the zeros of  $\pi_n(t)$ .

**LEMMA 1.** *For  $n$  odd,*

$$\pi_n\left(\frac{n}{2} - \tau\right) = \pi_n\left(\frac{n}{2} + \tau\right),$$

[i.e.,  $\pi_n(t)$  is symmetric about  $t = n/2$ ]; for  $n$  even:

$$\pi_n\left(\frac{n}{2} - \tau\right) = -\pi_n\left(\frac{n}{2} + \tau\right)$$

[i.e.,  $\pi_n(t)$  is anti-symmetric about  $t = n/2$ ].

*Proof.* Clearly,  $\pi_n(n/2 - \tau)$  and  $\pi_n(n/2 + \tau)$  are polynomials of degree  $n + 1$  in  $\tau$  and have the same  $n + 1$  roots:

$$\tau = \frac{n}{2}, \frac{n}{2} - 1, \frac{n}{2} - 2, \dots, -\frac{n}{2}.$$

Thus, these polynomials differ by at most a constant factor. Comparing coefficients of the leading terms in each then yields

$$\pi_n\left(\frac{n}{2} + \tau\right) = (-1)^{n+1}\pi_n\left(\frac{n}{2} - \tau\right). \quad \blacksquare$$

A further result which contains a comparison of the magnitudes of  $\pi_n(t)$  at various points is contained in

**LEMMA 2.** (a) Let  $t + 1$  be any non-integral point in  $0 < t + 1 \leq n/2$ . Then

$$|\pi_n(t + 1)| < |\pi_n(t)|.$$

(b) Let  $t$  be any non-integral point in  $n/2 \leq t < n$ . Then

$$|\pi_n(t)| < |\pi_n(t + 1)|.$$

*Proof.* Since, in part (a),  $t + 1$  is non-integral,  $t < n$  is also non-integral and we may form

$$\begin{aligned} \left| \frac{\pi_n(t + 1)}{\pi_n(t)} \right| &= \left| \frac{(t + 1)(t)(t - 1) \cdots (t - n + 1)}{(t)(t - 1) \cdots (t - n + 1)(t - n)} \right| \\ &= \left| \frac{t + 1}{t - n} \right| = \left| \frac{t + 1}{n - t} \right| = \frac{t + 1}{(n + 1) - (t + 1)} \\ &\leq \frac{n/2}{(n + 1) - (n/2)} = \frac{1}{1 + 2/n} < 1. \end{aligned}$$

Thus, part (a) is proved. Part (b) follows from part (a) by using the symmetry properties of Lemma 1.  $\blacksquare$

The properties of  $\pi_n(t)$  proven in Lemmas 1 and 2 are illustrated in Figure 1 where  $\pi_n(t)$  is plotted for  $n = 5$  and  $n = 6$ . It easily follows from Lemma 2 that the maximum of  $|\pi_n(t)|$  in  $[0, n]$  occurs in the interval

$(0, 1)$ , or equivalently in  $(n - 1, n)$ . A *lower bound* on this magnitude is furnished by

$$\begin{aligned} \max_{0 \leq t \leq n} |\pi_n(t)| &\geq |\pi_n(\frac{1}{2})| = \prod_{j=0}^n |\frac{1}{2} - j| \\ &= \frac{(2n)!}{2^{2n+1} n!}. \end{aligned}$$

Using this bound we may compare the quantity  $h^{n+1}\pi_n(t)$  for equally spaced interpolation points with the corresponding *error factor*,  $\prod_{j=0}^n (x - x_j)$ , for the Chebyshev interpolation points (i.e., with that factor which deviates least from zero determined in Chapter 5, Subsection 4.2). If the interpolation is to be employed over the interval  $[a, b]$ , then for the Chebyshev points we have by (4.17) of Chapter 5

$$M_{\text{Ch}} \equiv \max_{a \leq x \leq b} \left| \prod_{j=0}^n (x - x_j) \right| = \frac{1}{2^n} \left( \frac{b-a}{2} \right)^{n+1};$$

and this value is attained at least  $n + 2$  times in the interval. For equal spacing in  $[a, b]$ , we have  $h = (b-a)/n$  and so

$$M_{\text{Eq}} \equiv \max_{0 \leq t \leq n} |h^{n+1}\pi_n(t)| > \frac{(2n)!}{2^{2n+1} n!} \left( \frac{b-a}{n} \right)^{n+1}.$$

Thus, using Stirling's formula we find that for  $n$  large

$$\frac{M_{\text{Ch}}}{M_{\text{Eq}}} < \frac{n}{\sqrt{2}} \left( \frac{e}{4} \right)^n = \frac{n}{\sqrt{2}} (0.6796\dots)^n.$$

So if interpolation is to be employed over the entire range  $[a, b]$ , we find that the ratio of the maximum error factors essentially decreases, at least exponentially, for large  $n$ . Thus, the Chebyshev fit is better in the above comparison. However, we may only want to employ the interpolation polynomial near the center of the interval  $[a, b]$ . Specifically for  $n$  odd, say that  $n = 2m + 1$ , and that we are interested in the error over

$$\left[ \frac{a+b}{2} - \frac{h}{2}, \frac{a+b}{2} + \frac{h}{2} \right],$$

i.e., an interval of length  $h$  centered in  $[a, b]$ . Now we find that

$$\begin{aligned} M_{\text{Eq}}^* &\equiv \max_{m \leq t \leq m+1} |h^{n+1}\pi_n(t)| = h^{n+1} |\pi_n(m + \frac{1}{2})| \\ &= \left( \frac{b-a}{n} \right)^{n+1} \left( \frac{n!}{2^n m!} \right)^2. \end{aligned}$$

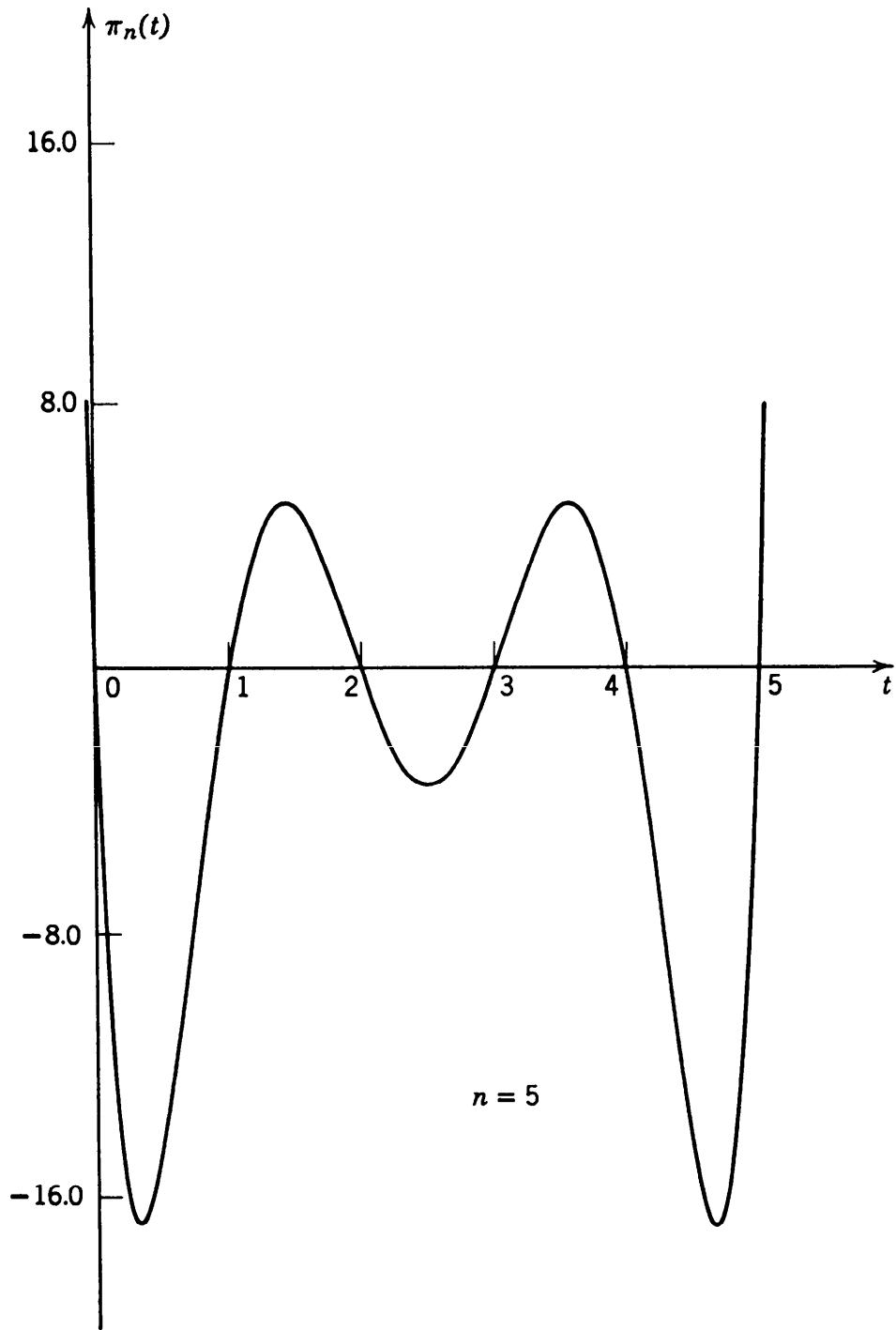


Figure 1a. Error factor for interpolation with 6 equidistant points.

Hence it follows that for large  $n = 2m + 1$ ,

$$\frac{M_{\text{Ch}}}{M_{\text{Eq}}^*} \approx \frac{1}{2} \left(\frac{e}{2}\right)^n = \frac{1}{2}(1.3591\dots)^n.$$

We thus find that for interpolation near the center of the interval of interpolation points the error factor for equally spaced points is exponentially smaller than the maximum for the Chebyshev fit over the entire interval.

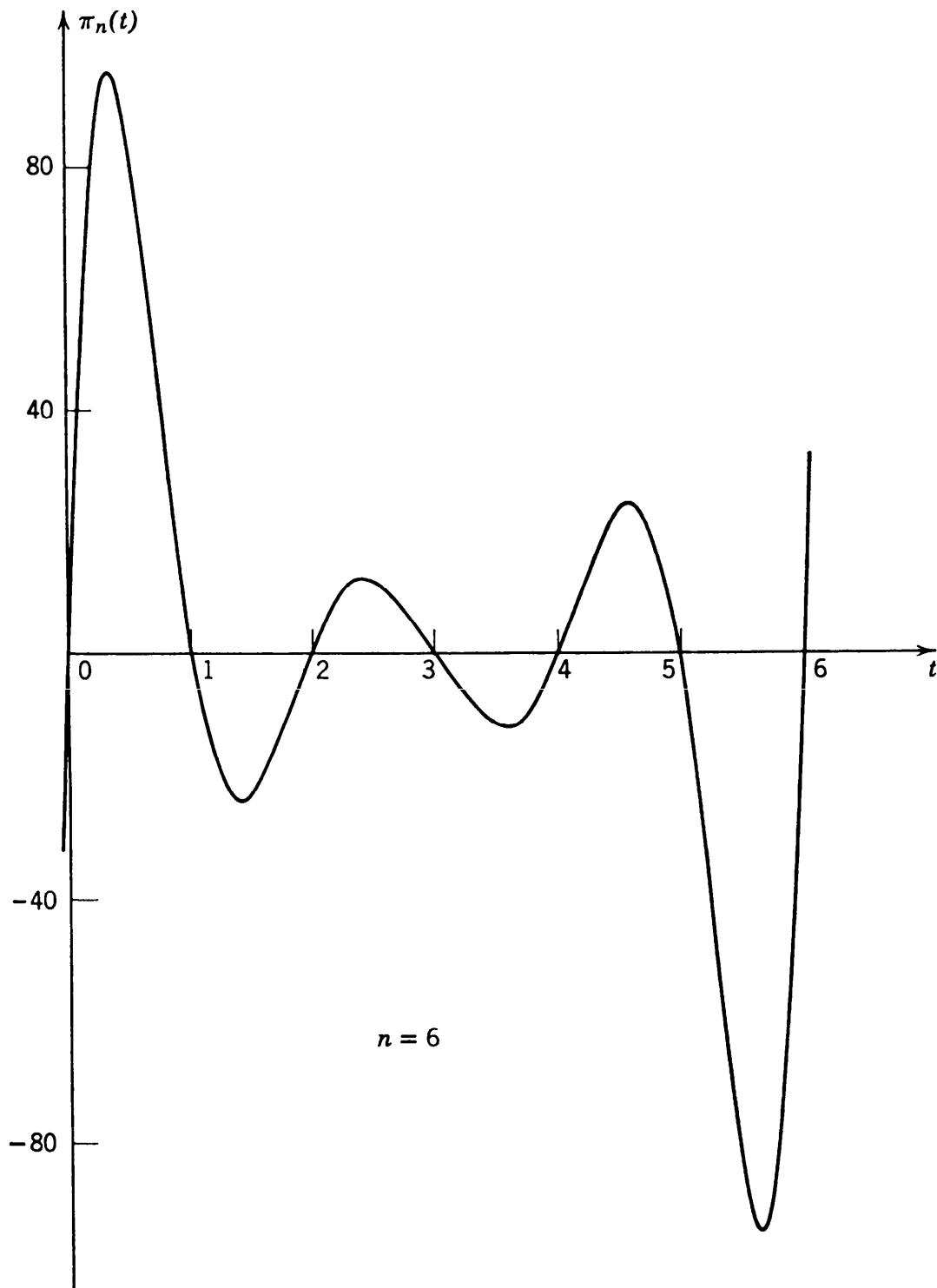


Figure 1b. Error factor for interpolation with 7 equidistant points.

It should also be observed (see Figure 1) that the error factor  $|\pi_n(t)|$  grows rapidly for  $t < 0$  and  $t > n$ . Thus if the “equally spaced” interpolation polynomial is used to extrapolate a function (i.e., to estimate values outside the interval of the interpolation points) we may expect the error to be much larger, in general, than it is for interpolation. Of course, the same is also true of the extrapolation error using the Chebyshev

points. In fact, for any distribution of interpolation points, the growth of the magnitude of the error factor can be bounded, outside the interval of interpolation points, by the error factor with the Chebyshev points. More precisely, for any choice of interpolation points  $x_0, x_1, \dots, x_n$  in  $[-1, 1]$  the error factor is, say,  $\prod_{j=0}^n (x - x_j) \equiv p_{n+1}(x)$ . If the Chebyshev points are used, then

$$\prod_{j=0}^n (x - x_j) = T_{n+1}(x).$$

Now if  $M \equiv \max_{-1 \leq x \leq 1} |p_{n+1}(x)|$ , it can be shown (see Problem 5) that for all  $x$  such that  $|x| > 1$

$$|p_{n+1}(x)| \leq M 2^n |T_{n+1}(x)|.$$

The equality can only hold if  $p_{n+1}(x) \equiv T_{n+1}(x)$ .

### 3.2. Centered Interpolation Formulae

Consider the error (13) in the interpolation polynomial for equal spacing. This error may be estimated if  $f^{(n+1)}(\xi)$  does not vary “too much” in the interval  $\min(x_0, x) < \xi < \max(x_n, x)$ . [An idea of this variation may be obtained, as a result of (7), by examining the differences  $\Delta^{n+1}f$ .] If the variation is not too large, then as an *estimate* of the error we may use

$$R_n(x) \equiv R_n(x_0 + th) \cong \frac{\pi_n(t)}{(n+1)!} \Delta^{n+1}f(x_0).$$

An approximate *bound* on the error is obtained if  $\pi_n(t)$  is replaced by its maximum absolute value in the interval in question.

Since, in general, we do not know  $f^{(n+1)}(\xi)$ , the best that can be done in order to obtain the smallest possible error is to use the interpolation polynomials only for that range of  $t$  where  $\pi_n(t)$  has its least absolute value, i.e., by Theorem 2, for  $t$  near  $n/2$  or equivalently for  $x$  near the midpoint of  $[x_0, x_n]$ . So if there are tabular points equally distributed about the interval of interpolation, then the interpolation polynomial to be employed should use tabular points centered as nearly as possible about the interval of interpolation. It is clear (see Figure 1) that when  $x$  is outside the interval  $[x_0, x_n]$ , or near the endpoints,  $|\pi_n(t)|$  may be relatively large and, if so, may cause the extrapolation or interpolation error to be relatively large.

It is rather inconvenient, in general, to locate in the Difference Table 1 those differences which must be employed in (12) when  $t$  is at  $n/2$ . For this purpose we derive special formulae which simplify the task. Let us assume that the interval in which the interpolation is to be done is

$x_0 < x < x_1$ , and that we have arbitrarily many tabular points  $x_j$ ,  $j = 0, \pm 1, \pm 2, \dots$ , about this interval. Then the ordinary Newton interpolation polynomial, using successively the points  $x_0, x_1, x_{-1}, x_2, x_{-2}, \dots$ , will have the desired features with regard to the interval  $[x_0, x_1]$ . This polynomial is of the form

$$(14) \quad Q_n(x) = f_0 + (x - x_0)f_{0,1} + (x - x_0)(x - x_1)f_{0,1,-1} \\ + (x - x_0)(x - x_1)(x - x_{-1})f_{0,1,-1,2} + \dots,$$

the form of the final term depending upon the oddness or evenness of  $n$ .

However, since the divided differences are symmetric functions of their arguments we may write

$$\begin{aligned} f_{0,1,-1} &= f_{-1,0,1}, \\ f_{0,1,-1,2} &= f_{-1,0,1,2}, \\ &\vdots \\ f_{0,1,-1,\dots,m,-m} &= f_{-m,\dots,-1,0,1,\dots,m}. \end{aligned}$$

Then using (4) with the appropriate shifts in the subscripts,

$$\begin{aligned} f_{-1,0,1} &= \frac{1}{2! h^2} \Delta^2 f_{-1}, \\ f_{-1,0,1,2} &= \frac{1}{3! h^3} \Delta^3 f_{-1}; \\ f_{-m,\dots,-1,0,1,\dots,m} &= \frac{1}{(2m)! h^{2m}} \Delta^{2m} f_{-m}, \\ f_{-m,\dots,-1,0,1,\dots,m,m+1} &= \frac{1}{(2m+1)! h^{2m+1}} \Delta^{2m+1} f_{-m}. \end{aligned}$$

The interpolation polynomial (14) can now be written as, for even  $n = 2m$ :

$$(15a) \quad Q_n(x_0 + th) = f_0 + t\Delta f_0 + \frac{t(t-1)}{2!} \Delta^2 f_{-1} \\ + \frac{t(t-1)(t+1)}{3!} \Delta^3 f_{-1} + \dots \\ + \frac{t(t-1)(t+1)\dots(t-m)}{(2m)!} \Delta^{2m} f_{-m};$$

and for odd  $n = 2m + 1$ :

$$(15b) \quad Q_n(x_0 + th) = f_0 + t\Delta f_0 + \dots \\ + \frac{t(t-1)(t+1)\dots(t-m)(t+m)}{(2m+1)!} \Delta^{2m+1} f_{-m}.$$

This is the *Gaussian (forward) form* for the interpolation polynomials.

The differences used in forming these polynomials are on the line containing  $x_0$  and the line between  $x_0$  and  $x_1$  (see Table 3).

**Table 3** *Differences*

$x$	$f(x)$	$\Delta$	$\Delta^2$	...	$\Delta^{2m}$	$\Delta^{2m+1}$
$x_{-m}$	$f_{-m}$					
		$\Delta f_{-m}$				
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
$x_{-1}$	$f_{-1}$				$\vdots$	$\vdots$
		$\Delta f_{-1}$				
$x_0$	$f_0$		$\Delta^2 f_{-1}$	...	$\Delta^{2m} f_{-m}$	
		$\Delta f_1$		...		$\Delta^{2m+1} f_{-m}$
$x_1$	$f_1$				$\vdots$	$\vdots$
$\vdots$	$\vdots$	$\vdots$	$\vdots$			
		$\Delta f_{m-1}$				
$x_m$	$f_m$					

A more symmetric form of the interpolation polynomial can be obtained when  $n = 2m + 1$  (i.e., of odd degree). For this purpose we must introduce the centered difference notation, for even order differences,

$$(16a) \quad \begin{aligned} \delta^2 f_r &\equiv \Delta f_r - \Delta f_{r-1} = \Delta^2 f_{r-1}; \\ \delta^{2k} f_r &= \delta^2(\delta^{2(k-1)} f)_r = \Delta^{2k} f_{r-k}, \quad k = 2, 3, \dots \end{aligned}$$

The point  $x_r$  is always the midpoint about which these differences are centered. With this notation, all odd order ordinary differences, higher than the first, can be written as the difference of two centered differences:

$$(16b) \quad \Delta^{2m+1} f_{-m} = \delta^{2m} f_1 - \delta^{2m} f_0; \quad m = 1, 2, \dots$$

Using (16) in (15b) yields:

$$\begin{aligned} Q_n(x_0 + th) &= f_0 + t\Delta f_0 + \frac{t(t-1)}{2!} \Delta^2 f_{-1} + \frac{t(t-1)(t+1)}{3!} \Delta^3 f_{-1} + \dots \\ &\quad + \frac{t(t-1)(t+1)\dots(t-m)}{(2m)!} \Delta^{2m} f_{-m} \\ &\quad + \frac{t(t-1)(t+1)\dots(t-m)(t+m)}{(2m+1)!} \Delta^{2m+1} f_{-m} \end{aligned}$$

$$\begin{aligned}
&= f_0 + t(f_1 - f_0) + \frac{t(t-1)}{2!} \delta^2 f_0 \\
&\quad + \frac{t(t-1)(t+1)}{3!} (\delta^2 f_1 - \delta^2 f_0) + \dots \\
&\quad + \frac{t(t-1)(t+1)\dots(t-m)}{(2m)!} \delta^{2m} f_0 \\
&\quad + \frac{t(t-1)(t+1)\dots(t-m)(t+m)}{(2m+1)!} (\delta^{2m} f_1 - \delta^{2m} f_0) \\
&= tf_1 + \frac{t(t-1)(t+1)}{3!} \delta^2 f_1 + \dots \\
&\quad + \frac{t(t-1)(t+1)\dots(t-m)(t+m)}{(2m+1)!} \delta^{2m} f_1 \\
&\quad + (1-t)f_0 + \frac{t(t-1)}{3!} (2-t) \delta^2 f_0 + \dots \\
&\quad + \frac{t(t-1)(t+1)\dots(t-m)}{(2m+1)!} (m+1-t) \delta^{2m} f_0.
\end{aligned}$$

By introducing  $s \equiv 1 - t$  we may simplify the coefficients of  $\delta^{2k} f_0$  and the above finally takes on the symmetric form:

$$\begin{aligned}
(17) \quad Q_n(x_0 + th) = & \quad sf_0 + \frac{s(s^2 - 1^2)}{3!} \delta^2 f_0 + \dots \\
& + \frac{s(s^2 - 1^2)\dots(s^2 - m^2)}{(2m+1)!} \delta^{2m} f_0 \\
& + tf_1 + \frac{t(t^2 - 1^2)}{3!} \delta^2 f_1 + \dots \\
& + \frac{t(t^2 - 1^2)\dots(t^2 - m^2)}{(2m+1)!} \delta^{2m} f_1.
\end{aligned}$$

This is known as *Everett's form* of the interpolation polynomial.

### 3.3. Practical Observations on Interpolation

In this subsection, we gather and comment on some of the “rules of thumb” which are used by the practitioners of interpolation.

(a) A convenient “rule” to determine approximately the magnitude of the error in linear interpolation is

$$|f(x_0 + th) - tf(x_0 + h) - (1-t)f(x_0)| \leq \left| \frac{\Delta^2 f(x_0)}{8} \right|.$$

The factor  $\frac{1}{8}$  is an upper bound for  $t(t - 1)/2$ , if  $0 \leq t \leq 1$ , while by (7)  $h^2 f''(\xi) \cong \Delta^2 f(x_0)$  in the remainder term (13).

(b) A “rule” for estimating the magnitude of the error in general polynomial interpolation is to use the magnitude of the first neglected term in the Newton form. That is, the error in using (12), (15a), or (15b) is approximately the next term in the series. By (13) this estimate is seen to be good if the ratio  $f^{(n+1)}(\xi)/f^{(n+1)}(\eta)$  is near to 1 for all  $\xi$  and  $\eta$  in an interval containing  $x$  and all the indicated  $x_i$ .

(c) In a table of differences, we may compute the “average value,”  $\tilde{\Delta}^p$ , of a column of  $p$ th order differences from the definition:

$$\tilde{\Delta}^p = \frac{1}{n} \sum_{i=0}^{n-1} \Delta^p f_i.$$

It is easy to show (see Problem 7), that if an isolated error is made in  $f_k$ , for some  $k$  satisfying  $p \leq k \leq n - p$ , then  $\tilde{\Delta}^p$  is unaffected. This observation provides a simple way of locating  $k$  and estimating the  $p + 1$  errors that arise in the column of  $p$ th order differences from an isolated error in  $f_k$ ; and hence yields the error in  $f_k$  approximately.

A table user could difference a printed table (whose accuracy has not been established) in order to weed out isolated typographical errors and to decide upon the order of interpolation that may be necessary.

(d) In the construction of a mathematical table, one tries to present a listing that provides a reasonable number of decimal places (or significant figures) and also permits a simple interpolation process to attain almost the full accuracy of the table, without its being too voluminous. To this end, some table makers list not only  $f(x_i)$  but  $\delta^2 f(x_i)$ , where  $\delta^2 f(x_i)$  is called a *modified second difference* (only the significant figures of  $\delta^2$  are listed). The modification is based on the use of the Everett form of the interpolation formula (17) with  $m = 2$ :

$$\begin{aligned} f(x_0 + th) &\cong tf(x_1) + \frac{t(t^2 - 1)}{6} \left[ \delta^2 f(x_1) + \frac{t^2 - 4}{20} \delta^4 f(x_1) \right] \\ &\quad + sf(x_0) + \frac{s(s^2 - 1)}{6} \left[ \delta^2 f(x_0) + \frac{s^2 - 4}{20} \delta^4 f(x_0) \right], \end{aligned}$$

where  $s = 1 - t$ . In order to incorporate most of the “effect” of the fourth difference into the second difference, one uses an “average value” for the coefficient  $(t^2 - 4)/20$ . Very simply, since

$$\int_0^1 \frac{p^2 - 4}{20} dp = -\frac{1}{60} \cong -0.18333,$$

we may define the modified second difference by

$$\hat{\delta}^2 f(x) = \delta^2 f(x) - \frac{1}{60} \delta^4 f(x),$$

and then use, for interpolation in the table,

$$f(x_0 + th) \cong tf(x_1) + \frac{t(t^2 - 1)}{6} \delta^2 f(x_1) + sf(x_0) + \frac{s(s^2 - 1)}{6} \delta^2 f(x_0).$$

Other, more sophisticated arguments have been given to justify the use of other “average values” for the coefficient  $(t^2 - 4)/20$ , e.g.,  $-0.18393$  (see Problem 6 for justification).

### 3.4. Divergence of Sequences of Interpolation Polynomials

It is not generally true that higher degree interpolation polynomials yield more accurate approximations. In fact, for equidistant points of interpolation one should use polynomials of relatively low order. We shall illustrate this by examining the interpolation error,  $R_n(x)$ , as a function of  $n$  and  $x$  for a particular function.

Specifically we take a function considered by Runge:

$$(18a) \quad f(x) \equiv \frac{1}{1 + x^2},$$

and consider, in  $[-5, 5]$ , the equally spaced points

$$(18b) \quad x_j = -5 + j\Delta x, \quad j = 0, 1, 2, \dots, n, \quad \Delta x = \frac{10}{n}.$$

For each  $n$  there is a unique polynomial  $P_n(x)$  of degree at most  $n$  such that  $P_n(x_j) = f(x_j)$ . This is the interpolation polynomial for (18a) using the points (18b). We shall show that  $|f(x) - P_n(x)|$  will become arbitrarily large at points in  $[-5, 5]$  if  $n$  is sufficiently large. This occurs even though the interpolation points  $\{x_j\}$  become dense in  $[-5, 5]$  as  $n \rightarrow \infty$ .

The remainder in interpolation is, by (1.8)

$$(19) \quad R_n(x) = f(x) - P_n(x),$$

$$= \prod_{j=0}^n (x - x_j) f[x_0, \dots, x_n, x].$$

However, with the function and points in (18) we claim that

$$(20) \quad f[x_0, \dots, x_n, x] = f(x) \cdot \frac{(-1)^{r+1}}{\prod_{j=0}^r (1 + x_j^2)} \cdot \begin{cases} 1, & \text{if } n = 2r + 1; \\ x, & \text{if } n = 2r. \end{cases}$$

We first prove this for the odd case,  $n = 2r + 1$ , by induction on  $r$ . Note that in this case there are an even number of interpolation points in (18b) and they satisfy  $x_j = -x_{n-j}$ . For  $r = 0$  we have  $n = 1$  and  $x_0 = -x_1$ .

Then a direct calculation from the divided difference representation (1.4) using (18a) yields

$$f[x_0, x_1, x] = - \left( \frac{1}{1+x^2} \right) \left( \frac{1}{1+x_0^{-2}} \right) = f(x) \frac{(-1)}{1+x_0^{-2}},$$

and the first step of the induction is established. Now assume (20) to be valid for  $n = 2r + 1$ , with any  $x_0, \dots, x_n$  that are pairwise symmetric (i.e.,  $x_j = -x_{n-j}$ ), and let  $m = n + 2 = 2(r + 1) + 1$ . We define the function  $g(x)$  by

$$g(x) \equiv f[x_1, x_2, \dots, x_{m-1}, x],$$

where now  $x_j = -x_{m-j}$ , and use (1.6) to write

$$f[x_0, x_1, \dots, x_m, x] = g[x_0, x_m, x].$$

However, by the inductive hypothesis it follows that

$$g(x) = f(x) \cdot A_r, \quad A_r \equiv \frac{(-1)^{r+1}}{\prod_{j=1}^r (1+x_j^2)}.$$

Also from (18b) we note that  $x_0 = -x_m$  and hence by the previous calculation

$$g[x_0, x_m, x] = f(x) \frac{(-1)}{1+x_0^2} A_r,$$

which upon substitution for  $A_r$  concludes the induction.

The verification of (20) for  $n = 2r$  is similar to the above and is left to the reader. Another proof of (20) is given as Problem 9.

Since  $(x - x_j)(x - x_{n-j}) = x^2 - x_j^2$  by (18b) and recalling that  $x_r = 0$  for  $n = 2r$  we have

$$(21) \quad \prod_{j=0}^n (x - x_j) = \prod_{j=0}^r (x^2 - x_j^2) \cdot \begin{cases} 1 & \text{if } n = 2r + 1; \\ 1/x & \text{if } n = 2r. \end{cases}$$

From (21) and (20) in (19) the error can be written as

$$(22) \quad R_n(x) = (-1)^{r+1} f(x) g_n(x), \quad g_n(x) \equiv \prod_{j=0}^r \frac{x^2 - x_j^2}{1+x_j^2}.$$

Note that  $\frac{1}{2\delta} \leq f(x) \leq 1$  for  $x$  in  $[-5, 5]$  and so the convergence or divergence properties are determined by  $g_n(x)$ . Further, since  $g_n(x) = g_n(-x)$ , we need only consider the interval  $[0, 5]$ . [It is also of interest to note that  $R_n(x)$  is, in fact, an even function of  $x$  for all  $n$ .]

To examine  $|g_n(x)|$  for large  $n$ , or equivalently for large  $r$ , we write

$$(23a) \quad |g_n(x)| = [e^{\Delta x \ln |g_n(x)|}]^{1/\Delta x},$$

where from the definition in (22)

$$(23b) \quad \Delta x \ln |g_n(x)| = \sum_{j=0}^r \ln \left| \frac{x^2 - x_j^2}{1 + x_j^2} \right| \Delta x.$$

In Problem 8 we show that for appropriate  $x \in [1, 5]$  and for all  $x_j$ :

$$(24) \quad |x + x_j| > C |\Delta x|^m.$$

For these values of  $x$  the sum in (23b) converges uniformly as  $n \rightarrow \infty$ ; that is, explicitly,

$$\begin{aligned} \lim_{n \rightarrow \infty} \Delta x \ln |g_n(x)| &= \lim_{r \rightarrow \infty} \sum_{j=0}^r \ln \left| \frac{x^2 - x_j^2}{1 + x_j^2} \right| \Delta x, \\ (25) \quad &= \int_{-5}^0 \ln \left| \frac{x^2 - \xi^2}{1 + \xi^2} \right| d\xi, \\ &\equiv q(x). \end{aligned}$$

To demonstrate this convergence, we note that

$$\ln \left| \frac{x^2 - x_j^2}{1 + x_j^2} \right| = \ln |x + x_j| + \ln |x - x_j| - \ln |1 + x_j^2|,$$

and similarly, with  $x_j$  replaced by  $\xi$ . Next we show that each of the three sums converges to the corresponding integrals. Those sums corresponding to the last two terms converge to their corresponding integrals by the definition of Riemann integrals since the corresponding integrands are continuous functions of  $\xi$  [recall that  $x \geq 1$  by (24) and  $x_j \leq 0$  for  $j \leq r$  by (18b)]. Hence, we need only show that

$$(26) \quad \lim_{r \rightarrow \infty} \sum_{j=0}^r \ln |x + x_j| \Delta x = \int_{-5}^0 \ln |x + \xi| d\xi,$$

provided  $x$  satisfies (24). Let  $\delta < 1$  be an arbitrarily small fixed positive number. Then

$$\begin{aligned} (27a) \quad &\lim_{r \rightarrow \infty} \sum_{|x+x_j| > \delta} \ln |x + x_j| \Delta x \\ &= \int_{-5}^{-x-\delta} \ln |x + \xi| d\xi + \int_{-x+\delta}^0 \ln |x + \xi| d\xi \end{aligned}$$

since  $\ln |x + \xi|$  is a continuous function of  $\xi$  over the indicated intervals of integration. The missing part of the integral in (26) is

$$(27b) \quad \int_{-x-\delta}^{-x+\delta} \ln |x + \xi| d\xi = 2 \int_0^\delta \ln \eta d\eta = 2(\delta \ln \delta - \delta).$$

The remaining part of the sum in (26) can be bounded if we take  $r$  so large that  $\Delta x < \delta$ . Then recalling (24) we have

$$\begin{aligned} \left| \sum_{|x+x_j| \leq \delta} \ln |x + x_j| \Delta x \right| &\leq 2\Delta x \left| \ln c(\Delta x)^m \right| + \left| \sum_{\Delta x \leq |x+x_j| \leq \delta} \ln |x + x_j| \Delta x \right| \\ (27c) \quad &\leq 2\Delta x \left| \ln c(\Delta x)^m \right| + 2 \left| \int_0^\delta \ln \eta d\eta \right|, \\ &\leq 2(\Delta x |m \ln \Delta x + \ln c| + \delta |\ln \delta - 1|) \\ &= \mathcal{O}(\delta |\ln \delta|). \end{aligned}$$

The first term on the right in the first line is obtained from the two terms say,  $x_k$  and  $x_{k+1}$  which are nearest to  $-x$ . The remaining sum has been bounded by the integral by means of the monotonicity of the function  $\ln x$ . That is, since  $\delta < 1$ , we use

$$0 > \Delta x \ln |x + x_j| > \int_{|x+x_{j-1}|}^{|x+x_j|} \ln \eta d\eta,$$

if  $|x+x_{j-1}| < |x+x_j|$  (otherwise limits of integration are  $|x+x_j|$ ,  $|x+x_{j+1}|$ ). Letting  $r \rightarrow \infty$  in (27c) and using (27a and b) we get (26) since  $\delta$  is arbitrarily small. This concludes the proof of (25).

In Problem 10 we indicate how  $q(x)$  can be explicitly evaluated and it is required to show that

$$(28a) \quad q(x) = 0 \quad \text{at } x = 3.63\dots;$$

$$(28b) \quad q(x) < 0 \quad \text{for } |x| < 3.63\dots;$$

$$(28c) \quad q(x) > 0 \quad \text{for } 3.63\dots < |x| \leq 5.$$

Now let  $x$  satisfy (24) and  $x > 3.63\dots$  as  $n \rightarrow \infty$ . Then by (25) and (28c) in (23a) we have, recalling that  $\Delta x = 10/n$ ,

$$\lim_{n \rightarrow \infty} |g_n(x)| = \infty.$$

That is, from (22),  $|R_n(x)| \rightarrow \infty$  as  $n \rightarrow \infty$  for  $x$  as above. Also, since  $R_n(x) = R_n(-x)$  the points of divergence are symmetrically located on the axis.

This example illustrates part of the general convergence theory for sequences of interpolation polynomials based on uniformly spaced points

in an interval  $[a, b]$ . According to this theory, if  $f(z)$  is “analytic” in a domain of the complex plane containing  $[a, b]$ , then the sequence of interpolation polynomials for  $f(z)$  converges inside the largest “lemniscate” whose interior is in the domain of analyticity of  $f(z)$ . The “lemniscate” passing through  $z = \pm 3.63\dots$  also passes through the points  $z = \pm\sqrt{-1}$  at which  $f(z) = 1/(1 + z^2)$  is singular.

The “lemniscates” associated with an interval  $[a, b]$  are simple closed curves which are analogous to the circles that characterize the domain of convergence of a power series expansion about a given point. That is, the sequence

$$S_n(z) = \sum_{k=0}^n \frac{f^{(k)}(\alpha)}{k!} (x - \alpha)^k$$

converges to the function  $f(z)$  for all  $z$  inside the largest circle  $|z - \alpha| = r$  about  $\alpha$  in which  $f(z)$  is analytic. For  $f(z) = 1/(1 + z^2)$  we obtain the sequence

$$L_{2n}(z) \equiv \sum_{k=0}^n (-1)^k z^{2k},$$

which converges for  $|z| < 1$ . This is the largest circle about the origin not containing the singular points  $z = \pm\sqrt{-1}$ .

### PROBLEMS, SECTION 3

1. Without using  $\Delta^n f(x) = h^n f^{(n)}(\xi)$ , derive the value of  $\Delta^n P_n(x)$  where  $P_n(x) = a_0 + a_1 x + \dots + a_n x^n$ .
2. Find the errors in the following table of function values taken at evenly spaced arguments: 50173, 53503, 56837, 60197, 63522, 66871, 70226, 73566, 76950, 80320, 83695, 87084, 90459, 93849, 97244, 100634, 104049, 107460. In this example it suffices to examine the column of second differences.

[Hint: See Problem 7.]

3. Compare error factors,  $\prod_{j=0}^n (x - x_j)$ , in equally spaced and Chebyshev interpolation over  $[a, b]$ . (That is, use Stirling’s approximation to  $n!$  and verify the estimates of  $M_{Ch}/M_{Eq}$  and  $M_{Ch}/M_{Eq}^*$ , following Lemma 2.)

4. Derive the result that  $\Delta \pi_n(t) = (n + 1)\pi_{n-1}(t)$ ,  $n = 1, 2, \dots$ ; where the spacing in  $\Delta$  is  $h = 1$ .

5.\* Prove the following

**THEOREM.** *If  $p_n(x)$  is a polynomial of degree at most  $n$  and  $\max_{|x| \leq 1} |p_n(x)| = M$  then for all  $x$  in  $|x| > 1$ ,*

$$|p_n(x)| \leq M 2^{n-1} T_n(x) \equiv M \cos(n \cos^{-1} x).$$

[Hint: For a proof by contradiction, consider the polynomial

$$q_n(x) = p_n(\xi) \frac{T_n(x)}{T_n(\xi)} - p_n(x),$$

with  $\xi$  a point where the conclusion is invalid show  $q_n(x)$  has  $n + 1$  zeros.]

6.\* The technique of *L. J. Comrie* for modifying the second difference uses the idea of selecting a constant,  $-c$ , to replace  $(t^2 - 4)/20$  and  $(s^2 - 4)/20$  in the Everett formula so that the maximum error in the resulting interpolation formula is a minimum. Supply the missing details in the following sketch:

Let the Everett form of the interpolation polynomial be

$$P_5(x_0 + ph) = qu_0 + pu_1 + e(q)\delta^2 u_0 + e(p)\delta^2 u_1 + d(q)\delta^4 u_0 + d(p)\delta^4 u_1,$$

where

$$q = 1 - p, \quad e(p) = \frac{p(p^2 - 1)}{6}, \quad d(p) = \frac{p(p^2 - 1)(p^2 - 4)}{120}.$$

Then

$$P_5(x_0 + ph) = qu_0 + pu_1 + e(p)(\delta^2 u_1 - c\delta^4 u_1) + e(q)(\delta^2 u_0 - c\delta^4 u_0) + R,$$

with

$$R = [d(p) + ce(p)]\delta^4 u_1 + [d(q) + ce(q)]\delta^4 u_0.$$

If we try to pick  $c$  so as to minimize  $\max_{0 \leq p \leq 1} |R|$ , we see that  $c$  must depend on  $u$ .

Hence, we simplify the problem by noting that  $\delta^4 u_1 = \delta^4 u_0 + \Delta \delta^4 u_0$ . Now if  $\Delta \delta^4 u_0$  is much smaller than  $\delta^4 u_0$  we may neglect the fifth difference and minimize

$$\begin{aligned} \max_{0 \leq p \leq 1} & |d(p) + d(q) + c[e(p) + e(q)]| \\ &= \max_{0 \leq p \leq 1} \left| \frac{p(p^2 - 1)(p - 2)}{24} + c \frac{p(p - 1)}{2} \right|. \end{aligned}$$

If we let the polynomial inside the absolute value sign be  $g(p, c)$ , then the maximum occurs when

$$\frac{\partial g}{\partial p} \equiv \frac{1}{6}(p - \frac{1}{2})(p^2 - p - 1 + 6c) = 0$$

or  $p = \frac{1}{2}, \frac{1}{2} \pm \sqrt{\frac{5}{4} - 6c}$ . Since  $p(p - 1)$  is of one sign,  $|g(p, c)|$  should have equal values at its maxima, in order that they be minimum. Set

$$|g(\frac{1}{2})| = |g(\frac{1}{2} \pm \sqrt{\frac{5}{4} - 6c})|$$

which yields

$$\frac{3 - 16c}{128} = \frac{(1 - 6c)^2}{24}.$$

Only the larger one of the roots,  $c$ , is appropriate:  $c \cong 0.18393$ .

7. Given a table of  $f(x)$  for  $x_j = x_0 + jh$ ,  $0 \leq j \leq n$ . Show that if  $f_k$  is replaced by  $f_k + \delta$ , for any  $k$  with  $p \leq k \leq n - p$  then  $\sum_{j=0}^{n-q} \Delta^q f_j / (n - q + 1)$  is unaffected, for  $1 \leq q \leq p$ .

8. For  $x$  a fixed positive irrational algebraic number of degree  $m$ , Liouville's theorem states that for all positive integers  $(p, q)$   $\exists$  a constant  $K(x) \exists$ .

$$|x - p/q| > Kq^{-m}.$$

Show that this implies (24) for some constant  $c(x)$  and all  $x_j$ .

**9.** Verify that if  $f(x) \equiv 1/(x + c)$

$$f[x_0, x_1, \dots, x_n] = (-1)^n \frac{1}{\prod_{j=0}^n (x_j + c)}.$$

Hence, establish (20) by writing

$$\frac{1}{1+x^2} = \frac{1}{2i} \left( \frac{1}{x-i} - \frac{1}{x+i} \right),$$

where  $i^2 = -1$ .

**10.** Verify for the function

$$q(x) \equiv \int_0^5 \ln \left| \frac{x^2 - \xi^2}{1 + \xi^2} \right| d\xi$$

that

- (a)  $q(x) = 0$  at  $x \approx 3.63\dots$ ;
- (b)  $q(x) < 0$  for  $|x| < 3.63\dots$ ;
- (c)  $q(x) > 0$  for  $3.63\dots < |x| \leq 5$ .

[Hint: Derive for  $0 \leq x \leq 5$ ,

$$\begin{aligned} q(x) &\equiv \int_0^x \ln(x - \xi) d\xi + \int_x^5 \ln(\xi - x) d\xi \\ &\quad + \int_0^5 \{\ln(\xi + x) - \ln(1 + \xi^2)\} d\xi \\ &\equiv (5 + x) \ln(5 + x) + (5 - x) \ln(5 - x) \\ &\quad - 5 \ln 26 - 2 \arctan 5. ] \end{aligned}$$

#### 4. CALCULUS OF DIFFERENCE OPERATORS

When dealing with equally spaced data there is a very useful operator method available to suggest new formulae and to aid in recalling the fundamental ones. The basic operators are:

- |     |   |
|-----|---|
| (1) | (a) Identity $If(x) \equiv f(x);$<br>(b) Displacement $Ef(x) \equiv f(x + h);$<br>(c) Difference $\Delta f(x) \equiv f(x + h) - f(x);$<br>(d) Derivative $Df(x) \equiv \frac{df(x)}{dx}.$ |
|-----|---|

Note that the displacement and difference operators imply a fixed spacing,  $h$ , by which the argument is to be shifted. We assume that  $E$  and  $\Delta$  use the same such value unless otherwise specified. To employ  $D$ , the function on which it operates must be differentiable. In fact, the classes of functions to which all symbolic formulae apply must generally be restricted. We

shall consider only the class of polynomials (but more general extensions are possible). Two operators,  $A$  and  $B$ , are said to be equal if  $Af(x) = Bf(x)$  for every function  $f(x)$  of the class under consideration, i.e., for every polynomial.

From the definitions (1), it is clear that the four operators are linear, i.e., if  $A$  is any one of them then

$$(2) \quad A[\alpha f(x) + \beta g(x)] = \alpha Af(x) + \beta Ag(x),$$

for arbitrary numbers  $\alpha, \beta$  and functions  $f(x), g(x)$ . The product,  $AB$ , and the sum,  $A + B$ , of two operators  $A$  and  $B$  are defined by

$$(3a) \quad (AB)f(x) \equiv A[Bf(x)],$$

$$(3b) \quad (A + B)f(x) \equiv Af(x) + Bf(x).$$

From the definition (3a) the integral powers,  $A^n$ , of any operator  $A$  may be defined inductively as

$$(4) \quad \begin{aligned} A^0 &\equiv I, \\ A^n &\equiv AA^{n-1}, \quad n = 1, 2, \dots \end{aligned}$$

In addition, we define non-integral powers of the displacement (or shift) operator,  $E^s$ , by

$$(5) \quad E^s f(x) \equiv f(x + sh),$$

where  $s$  is any real number, and observe that  $E^s E^r = E^{s+r}$ .

Using the definitions (1a), (1b), and (3b) we have

$$\begin{aligned} Ef(x) &= f(x + h) \\ &= f(x) + f(x + h) - f(x) \\ &= (I + \Delta)f(x). \end{aligned}$$

Thus, we conclude, from the definition of equality of operators, that

$$(6) \quad E = I + \Delta.$$

Equivalently, we have  $\Delta = E - I$  and from the definition of powers of operators it easily follows that

$$(7) \quad \Delta^n = (E - I)^n,$$

$$= \sum_{i=0}^n (-1)^i \binom{n}{i} E^{(n-i)}.$$

This result may be proved by induction, just as is the usual binomial expansion. However, by applying (7) to  $f(x_0)$  we obtain (3.6) which has

previously been derived for general functions. Thus (3.6) yields an independent proof of (7).

From the extended definition (5) we may write

$$f(x + sh) = E^s f(x).$$

On the other hand, the Newton form of the interpolation polynomial gives

$$f(x + sh) = \left[ 1 + \pi_0(s)\Delta + \frac{\pi_1(s)}{2!} \Delta^2 + \cdots + \frac{\pi_k(s)}{(k+1)!} \Delta^{k+1} + \cdots \right] f(x),$$

where we note that the series terminates with  $\Delta^p$  if  $f(x)$  is a polynomial of degree  $p$ .

But the formal binomial series expansion of  $(I + \Delta)^s$ , for  $s$  arbitrary, is identical with the series on the right-hand side, and hence we adopt it as the definition for  $s$  non-integral. That is, with this convention,

$$(8) \quad E^s = (I + \Delta)^s \quad \text{for } s \text{ arbitrary.}$$

Thus, it is clear that the steps leading to (8) are not a derivation of Newton's formula but can now be used to recall that formula when required.

Similarly, such manipulations can be employed to suggest new formulae which can then be verified independently. For example, we define the backward difference operator,  $\nabla$ , by

$$\nabla f(x) \equiv f(x) - f(x - h).$$

Then as in deriving (6) we find that

$$E^{-1} = (I - \nabla),$$

and proceeding as in (8) we obtain, formally

$$(9) \quad f(x - sh) = \left[ 1 - \pi_0(s)\nabla + \frac{\pi_1(s)}{2!} \nabla^2 + \cdots + (-1)^{k+1} \frac{\pi_k(s)}{(k+1)!} \nabla^{k+1} + \cdots \right] f(x).$$

The formula suggested is, in fact, well known as *Newton's backward difference formula*. By introducing centered difference operators,

$$\delta f(x) \equiv f\left(x + \frac{h}{2}\right) - f\left(x - \frac{h}{2}\right),$$

we could derive other formulae.

To relate  $D$  to the other operators we write the formal Taylor's series expansion

$$f(x + h) = f(x) + \frac{h}{1!} f'(x) + \frac{h^2}{2!} f''(x) + \cdots + \frac{h^n}{n!} f^{(n)}(x) + \cdots$$

in the symbolic form

$$Ef(x) = \left[ 1 + \frac{hD}{1!} + \frac{h^2 D^2}{2!} + \cdots + \frac{h^n D^n}{n!} + \cdots \right] f(x).$$

Since the series in the brackets is the expansion of  $e^{hD}$  and the equality is valid for all polynomials, we have the interesting result:

$$(10) \quad e^{hD} = E = I + \Delta.$$

[We recall that this merely says that the operator  $E$  and the operator  $I + hD + \cdots + h^n D^n/n!$  are equivalent when applied to a polynomial of degree  $n$ , for all positive integers  $n$ .] Formally by taking logarithms in equation (10) we find that

$$(11) \quad hD = \ln(I + \Delta)$$

$$= \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \cdots + (-1)^{n+1} \frac{1}{n} \Delta^n + \cdots.$$

This formula suggests that  $hD$  and the first  $n$  terms on the right-hand side might be equivalent when applied to any polynomial of degree  $n$ . To verify this we use the Newton formula, (8), for any polynomial  $f(x)$  of degree  $n$ :

$$\begin{aligned} f(x + sh) &= \left[ 1 + s\Delta + \frac{s(s-1)}{2!} \Delta^2 + \cdots \right. \\ &\quad \left. + \frac{s(s-1)\cdots(s-n+1)}{n!} \Delta^n \right] f(x). \end{aligned}$$

Differentiate with respect to  $s$  and evaluate at  $s = 0$  to get

$$hf'(x) = \left[ \Delta - \frac{1}{2}\Delta^2 + \frac{1}{3}\Delta^3 - \cdots + (-1)^{n+1} \frac{1}{n} \Delta^n \right] f(x),$$

which was to be shown. The relation (11) may now be employed to obtain forward difference approximations to the derivative of a tabulated function. The general problem of approximating the derivatives of a function is considered in more detail in the next section.

Symbolic methods may also be employed to determine formulae for the approximate evaluation of integrals. Thus we define

$$(12) \quad Jf(x) \equiv \int_x^{x+h} f(\xi) d\xi,$$

and by using (5) we have, formally,

$$\begin{aligned}
(13) \quad Jf(x) &= h \int_0^1 E^s f(x) ds \\
&= h \int_0^1 E^s ds f(x) = h \int_0^1 e^{s \ln E} ds f(x) \\
&= h \frac{E - I}{\ln E} f(x) \\
&= \frac{h\Delta}{\ln(I + \Delta)} f(x).
\end{aligned}$$

By using (11) this result implies  $J = \Delta/D$  or  $JD = DJ = \Delta$  which may be verified directly. However, if we write

$$\ln(I + \Delta) = \Delta(I - R),$$

where by definition

$$R = \frac{1}{2}\Delta - \frac{1}{3}\Delta^2 + \cdots + (-1)^n \frac{1}{n} \Delta^{n-1} + \cdots$$

then (13) becomes, symbolically,

$$\begin{aligned}
(14) \quad J &= \frac{h}{I - R} \\
&= h(I + R + R^2 + \cdots).
\end{aligned}$$

Again this might be interpreted as meaning that when applied to any polynomial  $f(x)$  of degree  $n$ ,  $J$  is equivalent to the first  $n + 1$  terms on the right, or more simply just those terms involving  $\Delta^k$  for  $k \leq n$ . Usually, (14) is written in powers of  $\Delta$ , i.e.,

$$(15) \quad J = h(I + \frac{1}{2}\Delta - \frac{1}{12}\Delta^2 + \frac{1}{24}\Delta^3 - \frac{1}{720}\Delta^4 + \cdots).$$

To justify (14) we note that  $Jx^n = \Delta x^{n+1}/(n + 1)$ . Now (11) and the definition of  $R$  give

$$\begin{aligned}
hDx^{n+1} &= h(n + 1)x^n \\
&= \Delta(I - R)x^{n+1} \\
&= \Delta x^{n+1} - R\Delta x^{n+1}.
\end{aligned}$$

Thus,  $\Delta x^{n+1} = h(n + 1)x^n + R\Delta x^{n+1}$  and iterating this result yields, since  $R^{n+1}\Delta x^{n+1} = 0$ ,

$$\begin{aligned}
\Delta x^{n+1} &= h(n + 1)x^n + R[h(n + 1)x^n + R\Delta x^{n+1}] \\
&\quad \vdots \\
&= h(n + 1)(I + R + R^2 + \cdots + R^n)x^n.
\end{aligned}$$

Thus, we have shown that

$$(16) \quad Jx^n = h(I + R + \dots + R^n)x^n.$$

Since this holds for all integers  $n$ , the validity of (14) when applied to polynomials follows. Different expressions for  $J$  can be obtained by using other identities to eliminate  $E$  in the derivation of (13). However, Chapter 7 is devoted to the detailed study of approximation methods for the evaluation of integrals.

The symbols  $1/D$  and  $1/\Delta$  are not operators in the same sense as  $D$ ,  $\Delta$ ,  $E$ , etc., since

$$(17) \quad \frac{1}{D}f(x) = \{F(x)\} \equiv \text{the set of polynomials } F(x) \text{ such that } F'(x) = f(x).$$

$$(18) \quad \frac{1}{\Delta}f(x) = \{G(x)\} \equiv \text{the set of polynomials } G(x) \text{ such that } G(x + h) - G(x) = f(x).$$

Nevertheless, if  $f(x)$  is a polynomial of degree  $n$ ,  $\{F(x)\}$  and  $\{G(x)\}$  have the same structure, that is,

$$\begin{aligned} \{F(x)\} &\equiv \{P_{n+1}(x) + c\}, \quad c \text{ any constant,} \\ &P_{n+1}(x) \text{ a fixed polynomial of degree } n + 1. \\ \{G(x)\} &\equiv \{Q_{n+1}(x) + d\}, \quad d \text{ any constant,} \\ &Q_{n+1}(x) \text{ a fixed polynomial of degree } n + 1. \end{aligned}$$

Hence,

$$(E^p - E^q) \frac{1}{D} \quad \text{and} \quad (E^p - E^q) \frac{1}{\Delta}$$

are well defined operators. We leave as Problems 2 and 3 the proof that the corresponding formal power series in  $\Delta$  respectively satisfy (19) and (20):

$$\begin{aligned} (19) \quad (E^p - E^q) \frac{1}{D}f(x) &= [(I + \Delta)^p - (I + \Delta)^q] \frac{h}{\log(I + \Delta)} f(x) \\ &= \int_{x+qh}^{x+ph} f(\xi) d\xi, \end{aligned}$$

where  $f(x)$  is a polynomial;

$$\begin{aligned} (20) \quad (E^p - E^q) \frac{1}{\Delta}f(x) &= [(I + \Delta)^p - (I + \Delta)^q] \frac{1}{\Delta}f(x) \\ &= \sum_{j=q}^{p-1} f(x + jh), \end{aligned}$$

if  $p > q$  are integers and  $f(x)$  is a polynomial.

Equations (19) and (20) permit the development of formulae for integration and summation of polynomials [see Problems (4) and (5)]. Another kind of representation of (20), with  $q = 0$ , arises from replacing the term  $1/\Delta$  on the far left by the formal power series in  $D$  by setting, from (10),

$$(21) \quad \begin{aligned} \frac{1}{\Delta} &= \frac{1}{e^{hD} - I} = \frac{1}{hD + \frac{h^2 D^2}{2!} + \frac{h^3 D^3}{3!} + \dots} \\ &= \frac{1}{hD} - \frac{1}{2} + \frac{1}{12} hD - \frac{1}{720} h^3 D^3 + \frac{1}{30,240} h^5 D^5 - \dots \end{aligned}$$

We obtain the formula,

$$(22) \quad \begin{aligned} \sum_{j=0}^{p-1} f(x + jh) &= \frac{1}{h} \int_x^{x+ph} f(\xi) d\xi - \frac{1}{2}[f(x + ph) - f(x)] \\ &\quad + \frac{h}{12} [f'(x + ph) - f'(x)] \\ &\quad - \frac{h^3}{720} [f^{(3)}(x + ph) - f^{(3)}(x)] \\ &\quad + \frac{h^5}{30,240} [f^{(5)}(x + ph) - f^{(5)}(x)] - \dots, \end{aligned}$$

called the *Euler-Maclaurin summation formula*.

## PROBLEMS, SECTION 4

1. Verify that the factorial polynomials  $W_0(x) \equiv 1$ ,  $W_n(x) \equiv x(x - h)\dots[x - (n - 1)h]$ ,  $n = 1, 2, \dots$ , satisfy  $\Delta W_0(x) = 0$ ,  $\Delta W_{n+1}(x) = h(n + 1)W_n(x)$ , and may be used as a basis for polynomials on which the calculus of difference operators is applied, e.g., with

$$\begin{aligned} P_n(x) &\equiv \sum_{j=0}^n \alpha_j W_j(x), \\ \Delta P_n(x) &= h \sum_{j=1}^n j \alpha_j W_{j-1}(x). \end{aligned}$$

2. Verify (19). Hint:  $\{(I + \Delta)^{p-1} + (I + \Delta)^{p-2} + \dots + (I + \Delta)^q\}(I + \Delta) - I = (I + \Delta)^p - (I + \Delta)^q$ .]  
 3. Verify (20). [Hint: Same as in Problem 2.]  
 4. Use (19) with  $p = 2$ ,  $q = 0$  to find Simpson's rule valid for all polynomials of degree  $\leq 3$ ;

$$\int_x^{x+2h} f(\xi) d\xi = \frac{h}{3} [f(x) + 4f(x + h) + f(x + 2h)].$$

5. Use (20) with  $f(x) = x^3$ ,  $h = 1$ ,  $q = 0$ ,  $p = n + 1$ ,  $x = 1$  to get a simple explicit expression for  $\sum_{j=1}^n j^3$ . (Construct a table of differences for  $j = 1, 2, 3, 4$ .)

6. Use (22) to derive the formula for  $\sum_{j=1}^n j^3$ .

7.\* Prove the Euler-Maclaurin summation formula (22) is correct for polynomials. [Assume that you can use (21) to formally get an infinite series for  $1/(e^{hD} - I)$ .]

## 5. NUMERICAL DIFFERENTIATION

A problem of importance in many applications is to approximate the derivative of a function, being given only several values of the function. An obvious approach to this problem is to employ the derivative of an interpolation polynomial as the desired approximation to the derivative of the function. This can also be done for higher derivatives, but clearly the approximation must, in general, deteriorate as the order of the derivative increases. We have seen in Section 3 that the interpolation error factor is least near the center of the interval of interpolation (for equally spaced points), and indeed an analogous result is also true for numerical differentiation.

Denote by  $P_n(x)$  the  $n$ th degree interpolation polynomial for  $f(x)$  with respect to the  $n + 1$  distinct points  $x_0, x_1, \dots, x_n$ . Then as an approximation to

$$\frac{d^k f(x)}{dx^k} \equiv f^{(k)}(x),$$

with  $k < n$ , we use  $P_n^{(k)}(x)$ . However, to assess this approximation we require some convenient representation for the error:

$$(1) \quad R_n^{(k)}(x) \equiv f^{(k)}(x) - P_n^{(k)}(x).$$

If  $f^{(n+1)}(x)$  is continuous in the interval,  $I_x$ , which includes the  $x_j$  and  $x$ , it has been shown in Theorem 2.1 of Chapter 5 that

$$R_n(x) = \prod_{j=0}^n (x - x_j) \frac{f^{(n+1)}(\xi)}{(n+1)!},$$

where  $\xi = \xi(x)$  is an unknown point in  $I_x$  for each  $x$ . It is tempting to differentiate this expression for  $R_n(x)$  in order to obtain  $R_n^{(k)}(x)$  but this is not generally legitimate. First of all,  $\xi(x)$  may not be single valued, let alone differentiable  $k$  times and secondly,  $f(x)$  may not be  $n + 1 + k$

times differentiable. If  $f(x)$  does have these differentiability properties, then another alternative is presented by recalling (1.7) in the form

$$R_n(x) = \prod_{j=0}^n (x - x_j) f[x_0, \dots, x_n, x].$$

It now follows by an application of Theorem 1.2 that this representation is  $k$  times differentiable. However, the resulting expression is rather complicated and only useful in the case of first derivatives,  $k = 1$ , in which  $x = x_i$  is one of the points of interpolation. The error becomes in this special case

$$\begin{aligned} (2) \quad f'(x_i) - P_n'(x_i) &= R_n'(x_i) = \prod_{\substack{j=0 \\ (j \neq i)}}^n (x_i - x_j) f[x_0, x_1, \dots, x_n, x_i] \\ &= \prod_{\substack{j=0 \\ (j \neq i)}}^n (x_i - x_j) \frac{f^{(n+1)}(\eta)}{(n+1)!}. \end{aligned}$$

The last expression in (2) can be deduced from Theorems 1.2 and 1.1.

To obtain practical error estimates for numerical differentiation in the more general case, we return to Rolle's theorem which was the basis for the original interpolation error estimates of Theorem 2.1 in Chapter 5. The results may be stated as

**THEOREM 1.** *Let the interpolation points be ordered by  $x_0 < x_1 < \dots < x_n$ . Let  $f^{(n+1)}(x)$  be continuous. Then for each  $k \leq n$ ,*

$$(3) \quad R_n^{(k)}(x) = \prod_{j=0}^{n-k} (x - \xi_j) \frac{f^{(n+1)}(\eta)}{(n+1-k)!};$$

where the  $n+1-k$  distinct points,  $\xi_j$ , are independent of  $x$  and lie in the intervals

$$(4) \quad x_j < \xi_j < x_{j+k}, \quad j = 0, 1, \dots, n-k;$$

and  $\eta = \eta(x)$  is some point in the interval containing  $x$  and the  $\xi_j$ .

*Proof.* Since  $R_n(x) = f(x) - P_n(x)$  has  $n+1$  continuous derivatives and vanishes at  $x = x_j$ ,  $j = 0, 1, \dots, n$ , we may apply Rolle's theorem  $k \leq n$  times. In applying this theorem we can keep track of the location of the implied zeros of the higher derivatives of  $R_n(x)$  by means of Table 1. The  $k$ th column lists the open intervals,  $(x_j, x_{j+k})$ , in each of which (by means of Rolle's theorem) at least one distinct root,  $\xi_j$ , of  $R_n^{(k)}(x)$  must lie. Thus the points  $\xi_j$  of (4) are defined and we note that they depend only upon the function  $f(x)$  and the interpolation points  $x_j$  but not upon  $x$ .

**Table 1** *Zeros of Higher Derivatives of  $R_n(x)$* 

$R_n(x)$	$R_n^{(1)}(x)$	$R_n^{(2)}(x)$	...	$R_n^{(k)}(x)$
$x_0$				
$x_1$	$(x_0, x_1)$			
$x_2$	$(x_1, x_2)$	$(x_0, x_2)$		
$x_3$	$(x_2, x_3)$	$(x_1, x_3)$		
$\vdots$	$\vdots$	$\vdots$		
$x_k$	$(x_{k-1}, x_k)$	$(x_{k-2}, x_k)$	...	$(x_0, x_k)$
$\vdots$	$\vdots$	$\vdots$		$\vdots$
$x_n$	$(x_{n-1}, x_n)$	$(x_{n-2}, x_n)$		$(x_{n-k}, x_n)$

We now define the function

$$F(z) = R_n^{(k)}(z) - \alpha \prod_{j=0}^{n-k} (z - \xi_j),$$

and note that  $F(\xi_j) = 0$  for  $j = 0, 1, \dots, n - k$ . For any fixed  $x$  distinct from the  $\xi_j$ , we pick  $\alpha = \alpha(x)$  such that  $F(x) = 0$ . Then  $F(z)$  has  $n - k + 2$  distinct zeros and we may apply Rolle's theorem again [noting that  $F(z)$  has  $n - k + 1$  continuous derivatives]. We deduce that  $F^{(n-k+1)}(z)$  has a zero, say at  $\eta$ , in the interval containing  $x$  and the  $\xi_j$ . From this result follows:

$$\begin{aligned} 0 &= F^{(n-k+1)}(\eta) = R_n^{(n+1)}(\eta) - \alpha(n - k + 1)! \\ &= f^{(n+1)}(\eta) - \alpha(n - k + 1)!, \end{aligned}$$

or

$$\alpha = \frac{f^{(n+1)}(\eta)}{(n - k + 1)!}.$$

By using this value of  $\alpha$  in  $F(x) = 0$  the result (3) follows for all  $x$ . That is, (3) holds also for  $x = \xi_j$  with arbitrary  $\eta$  since  $F(\xi_j) = 0$  for arbitrary  $\alpha$ . ■

The expression (3) for the error in numerical differentiation is valid for all  $x$  and so is of much more general applicability than expressions of the form (2). Using the known intervals (4) it is possible to obtain bounds on the error. For instance, if  $x$  and the  $x_j$  all lie in  $[a, b]$  and in this interval  $|f^{(n+1)}(x)| \leq M$  then clearly,

$$|R_n^{(k)}(x)| \leq \frac{M|b - a|^{n-k+1}}{(n - k + 1)!}.$$

Sharper estimates may be obtained by a more careful use of the inequalities (4) in bounding the error factor,  $\prod_{j=0}^{n-k} (x - \xi_j)$ .

There are other ways to determine numerical differentiation formulae and their errors. Suppose the value  $f^{(k)}(a)$  is to be approximated by using the values  $f(x_i)$ ,  $i = 1, 2, \dots, m$ . With an  $f(x)$  that has  $n + 1$  continuous derivatives where  $n + 1 \geq m$  we define  $h_i \equiv x_i - a$  and use Taylor's theorem to write

$$\begin{aligned} f(x_i) &= f(a + h_i) \\ &= f(a) + h_i f'(a) + \frac{h_i^2}{2!} f''(a) + \cdots + \frac{h_i^n}{n!} f^{(n)}(a) \\ &\quad + \frac{h_i^{n+1}}{(n+1)!} f^{(n+1)}(a + \theta_i h_i), \quad i = 1, 2, \dots, m. \end{aligned}$$

Here, of course,  $0 < \theta_i < 1$ . We now form a linear combination of these equations with weights,  $\alpha_i$ , to be determined.

$$\begin{aligned} \sum_{i=1}^m \alpha_i f(x_i) &= \left( \sum_{i=1}^m \alpha_i \right) f(a) + \left( \sum_{i=1}^m \alpha_i h_i \right) f'(a) + \cdots \\ (5) \quad &\quad + \left( \sum_{i=1}^m \alpha_i h_i^n \right) \frac{f^{(n)}(a)}{n!} + \frac{1}{(n+1)!} \sum_{i=1}^m \alpha_i h_i^{n+1} f^{(n+1)}(a + \theta_i h_i). \end{aligned}$$

We choose the  $\alpha_i$  in order that the linear combination of the values  $f(x_i)$  be the most accurate approximation to  $f^{(k)}(a)$ . Thus we impose the  $m$  conditions on the  $m$  unknowns  $\alpha_i$ :

$$(6) \quad \sum_{i=1}^m \alpha_i h_i^\nu = \nu! \delta_{\nu k}, \quad \nu = 0, 1, \dots, m-1.$$

It is clear that the system (6) has a unique solution since the coefficient determinant is a *Vandermonde* determinant. Thus a necessary and sufficient condition for (6) to have a non-trivial solution is that it be non-homogeneous, i.e.,  $m > k$ . Hence, *in order to approximate a  $k$ th derivative we need more than  $k$  points*. With the solution of the system (6) in (5) we obtain, recalling that  $n + 1 \geq m$

$$\begin{aligned} f^{(k)}(a) &= \sum_{i=1}^m \alpha_i f(x_i) - \frac{1}{m!} \left( \sum_{i=1}^m \alpha_i h_i^m \right) f^{(m)}(a) - \cdots \\ (7) \quad &\quad - \frac{1}{n!} \left( \sum_{i=1}^m \alpha_i h_i^n \right) f^{(n)}(a) - \frac{1}{(n+1)!} \sum_{i=1}^m \alpha_i h_i^{n+1} f^{(n+1)}(a + \theta_i h_i), \\ &\quad m > k. \end{aligned}$$

This procedure is equivalent to what may be called the *method of undetermined coefficients*: if we are given  $m$  function values,  $f(x_i)$ , we seek that linear combination of the values at these points which would give the exact value of the derivative  $f^{(k)}(a)$  for all polynomials of as high a degree as possible, at the fixed point  $a$ . Specifically for the first derivative, since

$$\frac{dx^\nu}{dx} \Big|_{x=a} = \nu a^{\nu-1},$$

we seek  $\alpha_i$  such that

$$(8) \quad \sum_{i=1}^m \alpha_i x_i^\nu = \nu a^{\nu-1}, \quad \nu = 0, 1, \dots, m-1.$$

This system also has a unique solution and it is, in fact, the same as the solution of the system (6) with  $k = 1$  (assuming that the quantities  $x_j$ ,  $h_j$ , and  $a$  are related by  $h_j = x_j - a$ ). This verification is posed as Problem 1. In the present derivation of the approximation formula no estimate of the error term is obtained but this could be remedied. It should also be observed that the method of undetermined coefficients can be used to determine approximations to higher derivatives.

### 5.1. Differentiation Using Equidistant Points

Naturally the numerical differentiation formulae are somewhat simplified when equally spaced data points are used. For instance, the operator identity (4.11) yields approximations of the form

$$(9a) \quad f'(x) = \frac{1}{h} [\Delta f(x) - \frac{1}{2}\Delta^2 f(x) + \dots + (-1)^{n+1} \frac{1}{n} \Delta^n f(x)] + R_n'(x).$$

Here the data required are  $f(x)$ ,  $f(x + h)$ ,  $\dots$ ,  $f(x + nh)$ , so that this formula only approximates the derivative at a tabular point and uses only data on one side of this point. This formula is obtained by differentiating Newton's forward difference formula (3.12) and evaluating the result at  $t = 0$ . Thus the error determined in (2) is applicable and becomes in this case

$$(9b) \quad R_n'(x) = \frac{(-1)^n h^n}{n+1} f^{(n+1)}(\eta), \quad x < \eta < x + nh.$$

Another example is furnished by differentiating the Gaussian form of the interpolating polynomial, say (3.15a) with  $n = 2m$ , and again setting  $t = 0$ ,

$$(10a) \quad f'(x_0) = \frac{1}{h} \left[ \Delta f(x_0) - \frac{1}{2!} \Delta^2 f(x_{-1}) - \frac{1}{3!} \Delta^3 f(x_{-1}) + \dots \right. \\ \left. + (-1)^m \frac{m!(m-1)!}{(2m)!} \Delta^{2m} f(x_{-m}) \right] + R_{2m}'(x_0).$$

Here the tabular points involved are symmetrically placed about  $x_0$  and again the error formula (2) is valid:

$$(10b) \quad R_{2m}'(x_0) = (-1)^m \frac{(m!)^2}{(n+1)!} h^n f^{(n+1)}(\eta),$$

$$x_0 - mh < \eta < x_0 + mh.$$

For  $n$  and  $m \gg 1$  Stirling's approximation for  $n!$  implies, since  $n = 2m$ ,

$$\frac{(m!)^2}{(2m+1)!} \approx \frac{n^{-\frac{1}{2}} \sqrt{2\pi}}{2^{n+1}}.$$

Thus a comparison of (9b) and (10b) indicates, for differentiation, the superiority of centering the data points about the point of approximation. An important special case of (10) occurs for  $n = 2$ ; this can be written as

$$(11) \quad f'(x) = \frac{f(x+h) - f(x-h)}{2h} - \frac{h^2}{6} f^{(3)}(\eta), \quad x-h < \eta < x+h.$$

The approximation formula in (11) is called the centered difference approximation to the first derivative.

The second derivative, or in fact, any even order derivative, can be approximated by a centered formula obtained by differentiating the other Gaussian interpolating polynomial, (15b) with  $n = 2m + 1$ . For example, with  $n = 3$  the approximation of  $f''(x_0)$  becomes on setting  $x_0 = x$ :

$$(12a) \quad f''(x) = \frac{\Delta^2 f(x-h)}{h^2} + R_3^{(2)}(x)$$

$$= \frac{f(x+h) - 2f(x) + f(x-h)}{h^2} + R_3^{(2)}(x).$$

The error term can now be estimated by Theorem 1:

$$(12b) \quad |R_3^{(2)}(x)| \leq h^2 |f^{(4)}(\eta)|, \quad x-h < \eta < x+2h.$$

But a better bound for the error can be found by the Taylor expansion procedure indicated by equations (6) and (7). That is, if we set  $k = 2$ ,  $m = 4$ ,  $a = x$ , and  $h_i = (i-2)h$ ,  $i = 1, 2, 3, 4$ , we find that

$$\alpha_1 = \alpha_3 = \frac{1}{h^2}, \quad \alpha_2 = \frac{-2}{h^2}, \quad \alpha_4 = 0.$$

The error expression, given by the last terms in (7), becomes, upon setting  $n = 3$ ,

$$R_3^{(2)}(x) = -\frac{h^2}{4!} [f^{(4)}(\xi_1) + f^{(4)}(\xi_2)], \quad x-h < \xi_1 < x < \xi_2 < x+h.$$

But  $f^{(4)}(x)$  was assumed continuous in this derivation so for some  $\xi$  in  $\xi_1 < \xi < \xi_2$  we must have  $\frac{1}{2}[f^{(4)}(\xi_1) + f^{(4)}(\xi_2)] = f^{(4)}(\xi)$ . Thus the error is

$$(12c) \quad R_3^{(2)}(x) = -\frac{h^2}{12}f^{(4)}(\xi), \quad x - h < \xi < x + h.$$

Note the improvement over the bound (12b) both in the factor  $\frac{1}{12}$  and the decreased range of the argument of the fourth derivative. It is also of interest to observe that the same approximation (12a) and error (12c) are obtained for  $m = 3$  and  $n = 3$  in (7) with the above choice of  $h_i$ ; that is, improving upon the accuracy by using data at one additional point is not always possible. (See Problem 2.)

## PROBLEMS, SECTION 5

1. Verify that the set of coefficients  $\{\alpha_i\}$  produced by solving the system (8) is the same as the solution  $\{\alpha_i\}$  of system (6) for  $k = 1$ ,  $h_i \equiv x_i - a$ .

2. Verify that if  $\{\alpha_i\}$  produces the differentiation formula of maximum accuracy

$$f^{(k)}(x) = \sum_{i=-r}^r \alpha_i f(x + ih),$$

then

$$f^{(k)}(x) = \sum_{i=-r}^{r+1} \beta_i f(x + ih)$$

can be no more accurate, and is of the same accuracy only if  $\beta_{r+1} = 0$ ,  $\beta_i = \alpha_i$  for  $i = -r, \dots, +r$ . Show also that the coefficients  $\alpha_i$  satisfy  $\alpha_p = \alpha_{-p}$  if  $k$  is even;  $\alpha_p = -\alpha_{-p}$  if  $k$  is odd.

## 6. MULTIVARIATE INTERPOLATION

The problems of polynomial interpolation and approximate differentiation for functions of several independent variables are important but the methods are less well developed than in the case of functions of a single variable. An immediate indication of the difficulties inherent in the higher dimensional case can be seen in the lack of uniqueness in the general interpolation problem. That is, we ask if  $p_1, p_2, \dots, p_n$  are  $n$  distinct points, say in the  $x, y$ -plane, then is there a unique polynomial of specified degree which attains specified values, say  $f(p_j)$ , at these points? Clearly the answer, in general, must be *no* since if all of the points  $[p_j, f(p_j)]$  lie on a straight line in  $x, y, z$ -space then there are infinitely many planes (i.e., linear polynomials) and perhaps higher degree polynomials of the form  $z = P(x, y)$  containing this line. We shall not dwell on these aspects of

interpolation in higher dimensions but shall show how to construct appropriate polynomials when the points of interpolation are specially chosen. It will also be found that in these special cases the interpolation polynomials are unique. For simplicity, we concentrate on functions of two variables but extension to more dimensions offers no difficulty.

Let us be given the  $(m + 1)(n + 1)$  distinct points  $p_{ij} \equiv (x_i, y_j)$ ;  $i = 0, 1, \dots, m, j = 0, 1, \dots, n$  and corresponding function values  $f(x_i, y_j)$ . These points form a rectangular array which is the set of intersections of the vertical lines  $x = x_i$  with the horizontal lines  $y = y_j$ , in the  $x, y$ -plane. We seek a polynomial,  $P(x, y)$ , of degree at most  $m$  in  $x$  and at most  $n$  in  $y$  such that

$$P(x_i, y_j) = f(x_i, y_j), \quad i = 0, 1, \dots, m, \quad j = 0, 1, \dots, n.$$

Such a polynomial must have the form

$$(1) \quad P(x, y) = \sum_{i=0}^m \sum_{j=0}^n a_{ij} x^i y^j,$$

with  $(m + 1)(n + 1)$  coefficients,  $a_{ij}$ , to be determined. This problem is easily solved, due to the special form of the points  $p_{ij}$ , with the use of the Lagrange interpolation coefficients. Let us write the Lagrange coefficients for the points  $\{x_i\}$  and  $\{y_j\}$  as

$$(2) \quad \begin{aligned} X_{m,i}(x) &= \prod_{\substack{k=0 \\ (k \neq i)}}^m \frac{x - x_k}{x_i - x_k}, \quad i = 0, 1, \dots, m; \\ Y_{n,j}(y) &= \prod_{\substack{k=0 \\ (k \neq j)}}^n \frac{y - y_k}{y_j - y_k}, \quad j = 0, 1, \dots, n. \end{aligned}$$

Then clearly, the polynomial  $X_{m,i}(x)Y_{n,j}(y)$  is of degree  $m$  in  $x$ , of degree  $n$  in  $y$  and vanishes when  $(x, y) = p_{\nu\mu}$  unless  $\nu = i$  and  $\mu = j$  in which case it is unity. Thus the required polynomial satisfying  $P(x_i, y_j) = f(x_i, y_j)$  can be written as

$$(3) \quad P(x, y) = \sum_{i=0}^m \sum_{j=0}^n X_{m,i}(x) Y_{n,j}(y) f(x_i, y_j).$$

Since the number of coefficients in the general polynomial (1) of degree  $m$  in  $x$  and  $n$  in  $y$  is equal to the number of conditions imposed we may expect that the interpolation polynomial (3) is unique. A formal proof of this fact is indicated in Problem 1. The extension to more independent variables is obvious.

Another representation of the interpolation polynomial (3) can be

obtained by using Newton's divided difference formulae. With the  $m + 1$  distinct points  $x_j$ , we have, recalling (1.7) and (1.8),

$$(4) \quad f(x, y) = \sum_{k=0}^m \omega_{k-1}(x) f[x_0, x_1, \dots, x_k; y] \\ + \omega_m(x) f[x_0, \dots, x_m, x; y].$$

Here we have introduced

$$\omega_{-1}(x) \equiv 1; \quad \omega_k(x) = \omega_{k-1}(x)(x - x_k), \quad k = 0, 1, \dots;$$

and the divided differences of a function of several variables are formed by keeping all but one variable fixed and taking the indicated differences with respect to the free variable. Hence  $f[x_0, x_1, \dots, x_k; y]$  as a function of the independent variable  $y$  has the Newton representation, using the  $n + 1$  points  $y_j$ :

$$(5) \quad f[x_0, x_1, \dots, x_k; y] = \sum_{j=0}^n \omega_{j-1}(y) f[x_0, x_1, \dots, x_k; y_0, y_1, \dots, y_j] \\ + \omega_n(y) f[x_0, \dots, x_k; y_0, \dots, y_n, y]$$

We use (5) for  $k = 0, 1, \dots, m$  in (4) to obtain

$$f(x, y) = Q(x, y) + R(x, y)$$

where

$$(6) \quad Q(x, y) \equiv \sum_{k=0}^m \sum_{j=0}^n \omega_{k-1}(x) \omega_{j-1}(y) f[x_0, \dots, x_k; y_0, \dots, y_j]$$

and

$$(7a) \quad R(x, y) = \omega_n(y) \sum_{k=0}^m \omega_{k-1}(x) f[x_0, \dots, x_k; y_0, \dots, y_n, y] \\ + \omega_m(x) f[x_0, \dots, x_m, x; y].$$

It is clear that  $R(x_i, y_j) = 0$  at the  $(m + 1)(n + 1)$  points  $(x_i, y_j)$  and hence by the uniqueness proof mentioned we can conclude that  $P(x, y) \equiv Q(x, y)$ . The derivation of the interpolation polynomial also yields an expression for the interpolation error,  $R(x, y)$ . To simplify this expression we again use Newton's formula and the  $m + 1$  points  $x_i$  to write

$$f[x; y_0, \dots, y_n, y] = \sum_{i=0}^m \omega_{i-1}(x) f[x_0, \dots, x_i; y_0, \dots, y_n, y] \\ + \omega_m(x) f[x_0, \dots, x_m, x; y_0, \dots, y_n, y].$$

If we multiply this identity by  $\omega_n(y)$  and subtract the result from (7a) we obtain finally,

$$(7b) \quad R(x, y) = \omega_m(x) f[x_0, \dots, x_m, x; y] + \omega_n(y) f[x; y_0, \dots, y_n, y] \\ - \omega_m(x) \omega_n(y) f[x_0, \dots, x_m, x; y_0, \dots, y_n, y].$$

If  $f(x, y)$  has continuous partial derivatives of orders  $m + 1$  and  $n + 1$ , respectively, in  $x$  and  $y$  and the appropriate mixed derivative of order  $m + n + 2$  then by applying the obvious extension of Theorem 1.1 the error becomes

$$(7c) \quad R(x, y) = \frac{\omega_m(x)}{(m+1)!} \frac{\partial^{m+1} f(\xi, y)}{\partial x^{m+1}} + \frac{\omega_n(y)}{(n+1)!} \frac{\partial^{n+1} f(x, \eta)}{\partial y^{n+1}} \\ - \frac{\omega_m(x)\omega_n(y)}{(m+1)!(n+1)!} \frac{\partial^{m+n+2} f(\xi', \eta')}{\partial x^{m+1} \partial y^{n+1}}.$$

This error formula is not of the form of the two-dimensional Taylor series error term, as was the case in one dimension, since different orders of differentiation occur here.

By specializing the interpolation points to be equally spaced we can obtain special forms of (3) and (6). These forms may be written in terms of the difference operators of Section 4, generalized so that they operate with respect to a particular independent variable. An example of such a representation is to be found in Problem 2.

The interpolation problem solved above, by (3) or (6), does not specify the degree of the polynomial in question but rather the maximum degree in  $x$  and  $y$ , separately. If a polynomial in two variables is to have total degree at most  $n$ , say, then it must have the form

$$(8) \quad P_n(x, y) = \sum_{k=0}^n \sum_{j=0}^{n-k} b_{kj} x^k y^j.$$

We note that the coefficients  $b_{kj}$  can be naturally arranged in a triangular array of  $\frac{1}{2}(n+1)(n+2)$  numbers. [In contrast, the  $a_{ij}$  in (1) formed a rectangular array of  $(m+1)(n+1)$  quantities.] We shall show that with an appropriate “triangular” array of points,  $(x_k, y_j)$ , the interpolation problem for polynomials of the form (8) can be uniquely solved.

Let  $\{x_j\}$  and  $\{y_j\}$  be two sets of  $n + 1$  distinct points, where  $j = 0, 1, \dots, n$ . Then we consider the array of points:

$$(9) \quad p_{kj} \equiv (x_k, y_j); \quad j + k = 0, 1, \dots, n.$$

[This array is actually triangular only if the values  $x_j$  and  $y_j$  are ordered by  $j$  and uniformly spaced, which we do not assume.] There are  $\frac{1}{2}(n+1)(n+2)$  such points and with them we pose the interpolation problem: *find a polynomial in  $x$  and  $y$  of degree at most  $n$  such that  $P_n(x_k, y_j) = f(x_k, y_j)$  for  $0 \leq j + k \leq n$ .* Newton’s divided difference formulae easily yield the solution of this problem. We obtain, upon replacing  $m$  by  $n$  in (4) and  $n$  by  $n - k$  in (5):

$$f(x, y) = P_n(x, y) + R_n(x, y)$$

where

$$(10) \quad P_n(x, y) = \sum_{k=0}^n \sum_{j=0}^{n-k} \omega_{k-1}(x) \omega_{j-1}(y) f[x_0, \dots, x_k; y_0, \dots, y_j];$$

and

$$(11) \quad R_n(x, y) = \sum_{k=0}^n \omega_{k-1}(x) \omega_{n-k}(y) f[x_0, \dots, x_k; y_0, \dots, y_{n-k}, y] \\ + \omega_n(x) f[x_0, \dots, x_n, x; y], \\ = \sum_{k=0}^{n+1} \frac{\omega_{k-1}(x) \omega_{n-k}(y)}{k! (n - k + 1)!} \left(\frac{\partial}{\partial x}\right)^k \left(\frac{\partial}{\partial y}\right)^{n-k+1} f(\xi_k, \eta_k).$$

The polynomial (10) has degree at most  $n$ . If we assume the indicated partial derivatives of  $f(x, y)$  to be continuous, then (11) yields the error which vanishes at all points in (9). The uniqueness of this polynomial follows from Problem 3. Thus the interpolation problem is solved by a polynomial of the form (8) on any set of points of the form (9). If these points are equally spaced and in monotone order, then the polynomial (10) can be simplified by introducing difference operators (see Problem 4). The remainder term in (11) is now analogous to that in Taylor's formula. In fact, if we let  $x_v \rightarrow x_0$  and  $y_v \rightarrow y_0$  for  $v = 1, 2, 3, \dots, n$  then (10) formally goes over into the Taylor expansion.

To approximate the partial derivatives of functions of several independent variables we could proceed as in Section 5 for functions of one variable. By using the error expressions of the form (7b) we could also obtain representations for the error in these numerical differentiation methods (if the function is sufficiently differentiable). However, in practice it turns out that relatively low order approximations to partial derivatives are usually all that are required. In these circumstances, it is easy to use the method of undetermined coefficients or the Taylor expansion method developed in Section 5. If no mixed derivatives occur and the points to be used are on a coordinate line in the direction of differentiation then the one-dimensional analysis is valid. For mixed derivatives the points employed in the expansion procedure must not be collinear. In Chapter 9, where partial differential equations are treated, specific applications are made in several examples.

## PROBLEMS, SECTION 6

1. Show that every polynomial  $Q(x, y)$  of degree  $m$  in  $x$  and  $n$  in  $y$  which vanishes at the  $(m + 1)(n + 1)$  distinct points  $(x_i, y_j)$ ;  $i = 0, 1, \dots, m$ ;  $j = 0, 1, \dots, n$ ; vanishes identically.

[Hint: Any such polynomial has the form

$$Q(x, y) = \sum_{v=0}^m \sum_{\mu=0}^n a_{v\mu} x^v y^\mu = \sum_{v=0}^m b_v(y) x^v.$$

Set  $y = y_j$  and then note that the polynomial  $Q(x, y_j)$  of degree  $m$  in  $x$  vanishes at  $m + 1$  distinct points. Thus,  $b_v(y_j) = 0$  for  $v = 0, 1, \dots, m$ . Next show that all  $a_{v\mu} = 0$ .]

2. We define the difference operators  $\Delta_x$  and  $\Delta_y$  by:

$$\begin{aligned}\Delta_x f(x, y) &\equiv f(x + h, y) - f(x, y), \\ \Delta_y f(x, y) &\equiv f(x, y + k) - f(x, y).\end{aligned}$$

If  $x_i = x_0 + ih$  and  $y_j = y_0 + jk$  then show that the interpolation polynomial of degree  $m$  in  $x$  and  $n$  in  $y$  for  $f(x, y)$  using the points  $(x_i, y_j); i = 0, 1, \dots, m; j = 0, 1, \dots, n$  is:

$$P(x_0 + sh, y_0 + tk) = \sum_{v=0}^m \sum_{\mu=0}^n \frac{\pi_v(s)\pi_\mu(t)}{v! \mu!} \Delta_x^v \Delta_y^\mu f(x_0, y_0).$$

3. State and prove the analog of Problem 1 for polynomials in  $x$  and  $y$  of degree at most  $n$  using points of the form (9).

4. Use the difference operators of Problem 2 and special equally spaced points of the form (9) to derive from (10):

$$P_n(x_0 + sh, y_0 + tk) = \sum_{v=0}^n \sum_{\mu=0}^{n-v} \frac{\pi_v(s)\pi_\mu(t)}{v! \mu!} \Delta_x^v \Delta_y^\mu f(x_0, y_0).$$

Define the corresponding backward difference operators  $\nabla_x$  and  $\nabla_y$ ; use them to write an interpolation polynomial of degree  $n$  in the plane; describe the set of interpolation points employed.

# 7

## Numerical Integration

### 0. INTRODUCTION

Simple explicit formulae cannot be given for the indefinite integrals of most functions. Furthermore, in many problems the integrand,  $f(x)$ , is not known precisely but perhaps is given by tabular data or defined as the solution of some differential equation (which cannot be solved explicitly). Thus, we seek appropriate numerical procedures to approximate the value of the definite integral, say

$$(1) \quad I\{f\} \equiv \int_a^b f(x) dx.$$

Unless otherwise specified  $[a, b]$  is a finite closed interval.

The types of approximation to (1) that we shall consider are all essentially of the form

$$(2) \quad I_n\{f\} \equiv \sum_{j=1}^n \alpha_j f(x_j).$$

When employed as an approximation to an integral, a sum of this form is called a numerical *quadrature* or numerical *integration formula*. For brevity, “numerical” is usually dropped. The  $n$  distinct points,  $x_j$ , are called the *quadrature points* or *nodes* and the quantities  $\alpha_j$  are called the *quadrature coefficients*. The basic problems in numerical integration are concerned with choosing the nodes and coefficients so that  $I_n\{f\}$  will be a “close” approximation to  $I\{f\}$  for a large class of functions,  $f(x)$ . As with polynomial approximation we note that different criteria may be used to measure the *quadrature error*,

$$E_n\{f\} \equiv I\{f\} - I_n\{f\}$$

even though it is a scalar; these criteria suggest different types of quadrature formulae.

One particularly useful notion which measures the error of a quadrature formula is its so-called *degree of precision*; this is by definition the maximum integer  $m$  such that  $E_n\{x^k\} = 0$  for  $k = 0, 1, \dots, m$  but  $E_n\{x^{m+1}\} \neq 0$ . Thus, if a formula has degree of precision  $m$  all polynomials of degree at most  $m$  are integrated exactly by that formula.

In fact, an expression for the error  $E_n\{f\}$  of such a scheme is given in

**THEOREM 1.** *If (2) has degree of precision  $m$  and  $f(x)$  has a continuous derivative of order  $m + 1$ , then*

$$(3) \quad E_n\{f\} \equiv I\{f\} - I_n\{f\} = \frac{1}{(m+1)!} \int_c^d f^{(m+1)}(\zeta) G_{n,m}(\zeta) d\zeta$$

where

$$G_{n,m}(\zeta) \equiv (m+1)[I\{(x-\zeta)_+^m\} - I_n\{(x-\zeta)_+^m\}],$$

with

$$(x-\zeta)_+^m \equiv \begin{cases} 0 & x \leq \zeta \\ (x-\zeta)^m & x \geq \zeta \end{cases}$$

and  $[c, d]$  is the smallest interval containing  $[a, b]$  and all  $x_j$ .

*Proof.* By Taylor's theorem (with remainder)

$$f(x) = T_m(x) + R_m(x),$$

where

$$T_m(x) \equiv \sum_{k=0}^m \frac{1}{k!} f^{(k)}(c)(x-c)^k,$$

$$R_m(x) \equiv \frac{1}{m!} \int_c^x f^{(m+1)}(\zeta) (x-\zeta)^m d\zeta.$$

Clearly,

$$(4) \quad R_m(x) = \frac{1}{m!} \int_c^d f^{(m+1)}(\zeta) (x-\zeta)_+^m d\zeta, \quad c \leq x \leq d.$$

$I\{\cdot\}$  and  $I_n\{\cdot\}$  are *linear operators*.† Hence, since  $I_n\{\cdot\}$  has degree of precision  $m$ ,  $I\{T_m\} = I_n\{T_m\}$  and

$$E_n\{f\} = I\{R_m\} - I_n\{R_m\}.$$

But for  $R_m(x)$  as given in (4), we find the expression (3), by interchanging the order of the operations on the variables  $x$  and  $\zeta$ . ■

†  $J\{\cdot\}$  is called a linear operator iff for all scalars  $a, b$  and functions  $f(x), g(x)$

$$J\{af(x) + bg(x)\} \equiv aJ\{f\} + bJ\{g\}.$$

It is left to Problem 1, to show that  $G_{n,m}(c) = G_{n,m}(d) = 0$ . In the following sections, simpler expressions than (3) for the error are found in special cases.

If a “close” approximation to  $f(x)$  in  $a \leq x \leq b$  is known, then the integral of the approximating function will be “close” to the integral of  $f(x)$ . That is, if

$$|f(x) - g(x)| \leq \epsilon,$$

then

$$\left| \int_a^b f(x) dx - \int_a^b g(x) dx \right| \leq |b - a|\epsilon.$$

This simple result is the motivation for developing most numerical integration methods. Of course, it is desirable that the approximating function should have a simple explicit indefinite integral. Hence polynomial approximations are naturally suggested and of these the interpolation polynomials are most frequently employed. Although there are quadrature formulae of great utility which are not necessarily motivated by the use of simple interpolation polynomials, we shall see nevertheless that *all such methods of general value are* what we will call *interpolatory*. There is considerable freedom to choose the position of the interpolation points relative to the interval of integration, so as may be expected there are a large number of numerical integration formulae. The choice of which formula to employ in a given case should depend upon its accuracy and relative ease of application.

In Sections 1 through 4, we consider *simple* quadrature formulae; in Section 5 we treat *composite* quadrature formulae. A composite formula is obtained by applying a simple formula to successive subintervals of  $[a, b]$ . In this fashion the problem of uniformly approximating the integrand  $f(x)$  over  $[a, b]$  is treated by using polynomials of a fixed “low” degree over each of the “small” subintervals into which the interval  $[a, b]$  is divided.

In many integration problems the integrand cannot be accurately approximated by a polynomial. Such cases may arise, for example, if  $f(x)$  is discontinuous at some points of the interval. Special considerations are required in these problems and we study some of them in Section 6.

In Section 7, we briefly treat the subject of approximating multiple integrals where the current state of the theory is not fully developed.

## PROBLEMS, SECTION 0

- Under the conditions of Theorem 1, show that

$$G_{n,m}(c) = G_{n,m}(d) = 0.$$

## 1. INTERPOLATORY QUADRATURE

Let  $n + 1$  distinct points  $x_j$  be ordered by

$$x_0 < x_1 < \cdots < x_n.$$

With these points as interpolation points, we form the interpolation polynomial  $P_n(x)$  of degree at most  $n$  [for the continuous function  $f(x)$ ] such that  $f(x_j) = P_n(x_j)$ ,  $j = 0, 1, \dots, n$ . Then as an approximation to the integral (0.1), we set

$$(1) \quad I_{n+1}\{f\} \equiv \int_a^b P_n(x) dx.$$

This integral is easily evaluated. In fact, by using the Lagrange form for the interpolation polynomial

$$(2a) \quad P_n(x) = \sum_{j=0}^n \phi_{n,j}(x) f(x_j)$$

$$(2b) \quad \phi_{n,j}(x) = \frac{\omega_n(x)}{(x - x_j)\omega_n'(x_j)} \quad j = 0, 1, \dots, n,$$

where  $\omega_n(x) \equiv (x - x_0)(x - x_1) \cdots (x - x_n)$ , we obtain from (1) the quadrature formula

$$(3a) \quad I_{n+1}\{f\} = \sum_{j=0}^n w_{n,j} f(x_j),$$

with the coefficients given by

$$(3b) \quad w_{n,j} = \int_a^b \phi_{n,j}(x) dx.$$

It is clear that the coefficients  $w_{n,j}$  are determined completely by the endpoints of the interval of integration and by the interpolation points  $x_j$ , which also are the nodes of the formula (3a); the coefficients are independent of the integrand. Any quadrature formula of the form (3a and b) is called an *interpolatory quadrature formula*.

The error in approximating the continuous function  $f(x)$  by  $P_n(x)$  is, by (1.8) of Chapter 6 and the definition of  $\omega_n(x)$  above,

$$f(x) - P_n(x) = \omega_n(x)f[x_0, \dots, x_n, x].$$

Integrate this equation over  $[a, b]$  and use (0.1) and (1) to obtain the *interpolatory quadrature error*

$$(4a) \quad \begin{aligned} E_{n+1}\{f\} &\equiv I\{f\} - I_{n+1}\{f\} \\ &= \int_a^b \omega_n(x)f[x_0, \dots, x_n, x] dx. \end{aligned}$$

If  $f(x)$  is a polynomial of degree  $n$  or less then  $E_{n+1}\{f\} = 0$  follows from the corollary to Theorem 1.1 of Chapter 6. Thus we have shown that *any interpolatory quadrature formula using  $n + 1$  nodes has degree of precision at least  $n$* . We shall, in fact, see later that even higher degrees of precision are possible if the nodes are specially placed with respect to the interval of integration. From (4a) a simple error bound is found,

$$(4b) \quad |E_{n+1}\{f\}| \leq \max_{x \in [a,b]} |f[x_0, x_1, \dots, x_n, x]| \int_a^b |\omega_n(x)| dx,$$

where without loss of generality we assume  $a \leq b$ .

To examine the error in more detail let us consider the *special case in which  $\omega_n(x)$  does not vanish in the open interval  $(a, b)$*  [i.e.,  $\omega_n(x)$  does not change sign there]:

If  $f'(x)$  is continuous on  $[a, b]$ , it follows that  $f[x_0, \dots, x_n, x]$  is continuous on  $[a, b]$  by Theorem 1.3 of Chapter 6. Thus the mean value theorem for integrals can be employed in (4a) to yield

$$(5a) \quad E_{n+1}\{f\} = f[x_0, \dots, x_n, \eta] \int_a^b \omega_n(x) dx, \quad a < \eta < b.$$

If, in addition,  $f^{(n+1)}(x)$  is continuous on the smallest closed interval,  $[c, d]$ , containing  $[a, b]$  and the nodes  $\{x_j\}$ , then by (1.9) of Chapter 6, (4a) becomes

$$(5b) \quad E_{n+1}\{f\} = \frac{1}{(n+1)!} \int_a^b \omega_n(x) f^{(n+1)}(\xi) dx, \quad \xi \equiv \xi(x) \in (c, d).$$

Now apply the same mean value theorem to (5b), where  $\omega_n(x)$  is of one sign and  $f^{(n+1)}(\xi(x))$  is continuous in  $x$  because

$$f^{(n+1)}(\xi(x)) \equiv (n+1)! f[x_0, x_1, \dots, x_n, x],$$

to find

$$(5c) \quad E_{n+1}\{f\} = \frac{f^{(n+1)}(\zeta)}{(n+1)!} \int_a^b \omega_n(x) dx, \quad \zeta \in (c, d).$$

In the general case, some nodes will lie in the interval of integration and the simple error formula (5c) is not generally valid. Our aim is to find a suitable replacement for (5c).

Specifically let there be  $r - 1 > 0$  interpolation points or nodes in the open interval  $(a, b)$ , as in Figure 1,

$$x_0 < \dots < x_i \leq a < x_{i+1} < \dots < x_{i+r-1} < b \leq x_{i+r} < \dots < x_n.$$

For convenience of notation we introduce the  $r + 1$  quantities,  $\xi_j$ ,

$$\xi_0 = a; \quad \xi_k = x_{i+k}, \quad k = 1, 2, \dots, r - 1; \quad \xi_r = b.$$

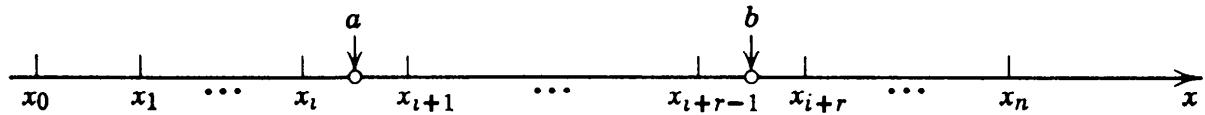


Figure 1

Now the error expression in (4a) can be written as

$$\begin{aligned} E_{n+1}\{f\} &= \int_{\xi_0}^{\xi_r} \omega_n(x) f[x_0, \dots, x_n, x] dx \\ &= \sum_{k=1}^r \int_{\xi_{k-1}}^{\xi_k} \omega_n(x) f[x_0, \dots, x_n, x] dx. \end{aligned}$$

In each of the intervals  $[\xi_{k-1}, \xi_k]$  the quantity  $\omega_n(x)$  is of one sign and it changes sign at the points  $\xi_1, \xi_2, \dots, \xi_{r-1}$ . Thus as in the derivation of (5a) we now conclude that

$$(6a) \quad E_{n+1}\{f\} = \sum_{k=1}^r C_k f[x_0, \dots, x_n, \eta_k], \quad \xi_{k-1} < \eta_k < \xi_k,$$

where

$$(6b) \quad C_k = \int_{\xi_{k-1}}^{\xi_k} \omega_n(x) dx, \quad k = 1, 2, \dots, r.$$

If  $f^{(n+1)}(x)$  is continuous in  $[x_0, x_n]$  then, as in (5c), we obtain

$$E_{n+1}\{f\} = \sum_{k=1}^r \frac{C_k}{(n+1)!} f^{(n+1)}(\zeta_k), \quad x_0 < \zeta_k < x_n.$$

However, it is clear from (6b) that the constants  $C_k$  alternate in sign; that is,

$$\text{sign } C_{k+1} = \text{sign } [-C_k], \quad k = 1, 2, \dots, r-1.$$

So the last form for the error can be written as two sums, each with coefficients of the same sign:

$$(6c) \quad E_{n+1}\{f\} = \left\{ \begin{array}{l} \frac{1}{(n+1)!} \{C_1 f^{(n+1)}(\zeta_1) + C_3 f^{(n+1)}(\zeta_3) + \dots\} \\ + \frac{1}{(n+1)!} \{C_2 f^{(n+1)}(\zeta_2) + C_4 f^{(n+1)}(\zeta_4) + \dots\}. \end{array} \right.$$

To simplify further, we require the following:

**LEMMA 1.** *Let  $g(x)$  be a continuous function in  $[a, b]$  and let  $\alpha_1, \alpha_2, \dots, \alpha_n$  be any set of non-negative numbers such that*

$$\sum_{k=1}^n \alpha_k = A.$$

Then for each set of  $n$  points  $x_k \in [a, b]$  there exists a  $\xi \in [a, b]$  such that

$$\sum_{k=1}^n \alpha_k g(x_k) = Ag(\xi).$$

*Proof.* Since  $g(x)$  is continuous on a closed interval, it has there a finite maximum,  $M$ , a finite minimum,  $m$ , and actually takes on these values and all intermediate values as  $x$  ranges over  $[a, b]$ . Thus for each of the  $x_k$

$$m \leq g(x_k) \leq M, \quad k = 1, 2, \dots, n.$$

Since the numbers  $\alpha_k$  are non-negative, this implies

$$\alpha_k m \leq \alpha_k g(x_k) \leq \alpha_k M.$$

Sum these inequalities for  $k = 1, 2, \dots, n$ , to find

$$Am \leq \sum_{k=1}^n \alpha_k g(x_k) \leq AM.$$

Hence the value of the sum must be equal to  $Ag(\xi)$  for some  $\xi \in [a, b]$ . ■

It should be observed that this lemma is analogous to the mean value theorem for integrals and that this proof copies the usual proof of that theorem.

Returning to (6c) we now have, by an obvious application of Lemma 1 to each of the sums in brackets

$$(7a) \quad E_{n+1}\{f\} = \frac{1}{(n+1)!} [K_o f^{(n+1)}(\zeta_o) - K_e f^{(n+1)}(\zeta_e)],$$

where

$$(7b) \quad K_o \equiv C_1 + C_3 + \dots, \quad K_e \equiv -C_2 - C_4 - \dots,$$

$$x_0 < \zeta_e, \zeta_o < x_n.$$

The constants  $K_e$  and  $K_o$  have the same sign and so some cancellation is suggested by (7a). In fact, since

$$|K_o - K_e| = \left| \sum_{j=1}^r C_j \right| = \left| \int_a^b \omega_n(x) dx \right|,$$

the error expression (7a) formally reduces to (5c) if  $\zeta_o = \zeta_e = \xi$ . In general,  $\zeta_o$  and  $\zeta_e$  are unknown points and  $f^{(n+1)}(x)$  may change sign in  $(x_0, x_n)$  so that the above reduction in the error may not occur. However, there are important special cases where, in fact, this maximum suggested cancellation does occur. We shall consider them later (see Theorem 2).

If the interpolation points or nodes are equally spaced, the above results

can be modified to exhibit the dependence of the error on the spacing of the points. That is, let the interpolation points be of the form

$$x_j = x_0 + jh, \quad j = 0, 1, \dots, n,$$

and introduce the change of variable from  $x$  to  $t$ ,

$$x = x_0 + th.$$

Now, from (3.10) of Chapter 6, we have

$$\omega_n(x) = h^{n+1} \pi_n(t),$$

and the integrals  $C_k$  in (6b) can be written as

$$C_k = h^{n+2} B_k$$

where

$$B_k = \int_{i+k-1}^{i+k} \pi_n(t) dt, \quad k = 2, 3, \dots, r-1; \quad B_1 = \int_{t_a}^{i+1} \pi_n(t) dt;$$

$$B_r = \int_{i+r-1}^{t_b} \pi_n(t) dt.$$

The limits of integration  $t_a$  and  $t_b$  are given by

$$t_a = \frac{a - x_0}{h}, \quad t_b = \frac{b - x_0}{h},$$

and lie in the interval

$$i \leq t_a < i+1, \quad i+r-1 < t_b \leq i+r.$$

Since  $\pi_n(t)$  is of one sign in these intervals,  $B_1$  and  $B_r$  can be bounded by

$$|B_1| \leq \left| \int_i^{i+1} \pi_n(t) dt \right|, \quad |B_r| \leq \left| \int_{i+r-1}^{i+r} \pi_n(t) dt \right|,$$

and these bounds are independent of  $h$ . By using these results in (7), we obtain the error representation

$$(8a) \quad E_{n+1}\{f\} = \frac{h^{n+2}}{(n+1)!} [L_o f^{(n+1)}(\zeta_o) - L_e f^{(n+1)}(\zeta_e)],$$

where  $x_0 < \zeta_e$ ,  $\zeta_o < x_n$  and

$$(8b) \quad L_o \equiv B_1 + B_3 + \dots, \quad L_e \equiv -B_2 - B_4 - \dots.$$

The constants  $L_o$  and  $L_e$  have the same sign and, in fact,

$$|L_o - L_e| = \left| \sum_{j=1}^r B_j \right| = \left| \int_{t_a}^{t_b} \pi_n(t) dt \right| \leq \int_{t_a}^{t_b} |\pi_n(t)| dt.$$

We have thus shown, in general, that if  $f^{n+1}(x)$  is continuous in the smallest interval,  $[c, d]$ , containing all the  $x_j$ ,  $a$  and  $b$ , then an interpolatory quadrature formula which uses  $n + 1$  equally spaced nodes, of spacing  $h$ , has an error of the form (8). Furthermore, by (8) or (4b):

$$|E_{n+1}\{f\}| \leq \frac{h^{n+2}}{(n+1)!} \int_{t_a}^{t_b} |\pi_n(t)| dt \max_{\xi \in [c, d]} |f^{(n+1)}(\xi)|.$$

This estimate is valid independent of the location of the nodes relative to the interval of integration. We give the analogue of formula (5c) which is valid in the special case that  $(a, b)$  contains none of the uniformly spaced points  $\{x_i\}$ :

$$(8c) \quad E_{n+1}\{f\} = \frac{h^{n+2}}{(n+1)!} f^{(n+1)}(\zeta) \int_{t_a}^{t_b} \pi_n(t) dt, \quad \zeta \in (c, d).$$

In the next subsection, we treat the Newton-Cotes formulae, and we find representations of  $E_{n+1}\{f\}$  that are of the form (5c) or (8c), even though  $\omega_n(x)$  changes sign in  $(a, b)$ .

### 1.1. Newton-Cotes Formulae

Let the interpolation points,  $x_j$ , be equally spaced, say as before,

$$(9a) \quad x_j = x_0 + jh, \quad j = 0, 1, \dots, n;$$

but now let the endpoints of the interval of integration be placed such that

$$(9b) \quad x_0 = a, \quad x_n = b, \quad h = \frac{b - a}{n}.$$

With this choice of nodes the quadrature formula (3) as an approximation to the integral (0.1) is called a *closed Newton-Cotes formula*. Note that all of the nodes are in the integration interval  $[a, b]$  and the word “closed” means that the endpoints  $a$  and  $b$  are the extreme nodes of the formula (3).

To examine the error, we again introduce the change of variable  $x = x_0 + th$  and obtain  $\omega_n(x) = h^{n+1}\pi_n(t)$ . Now, however,  $t$  ranges over the interval  $[0, n]$  and so we deduce properties of  $\omega_n(x)$  over  $[a, b]$  analogous to those of  $\pi_n(t)$  developed in Lemmas 3.1 and 3.2 of Chapter 6. With the notation

$$x_{n/2} = \frac{a + b}{2} = x_0 + \frac{n}{2} h,$$

these properties are restated in

**LEMMA 2.**

$$\omega_n(x_{n/2} + \xi) = (-1)^{n+1} \omega_n(x_{n/2} - \xi). \quad \blacksquare$$

**LEMMA 3.** (a) For  $a < \xi + h \leq x_{n/2}$  and  $\xi \neq x_j, j = 0, 1, \dots, n$ :

$$|\omega_n(\xi + h)| < |\omega_n(\xi)|;$$

(b) For  $x_{n/2} \leq \xi < b$  and  $\xi \neq x_j, j = 0, 1, \dots, n$ :

$$|\omega_n(\xi)| < |\omega_n(\xi + h)|. \quad \blacksquare$$

Let us introduce the functions

$$(10) \quad \Omega_n(x) = \int_a^x \omega_n(\xi) d\xi, \quad n = 1, 2, \dots,$$

which will be used to estimate the error in the closed Newton-Cotes formulae. For these functions, we have

**LEMMA 4.** For  $n$  even

- (a)  $\Omega_n(a) = \Omega_n(b) = 0$ ;
- (b)  $\Omega_n(x) > 0, \quad a < x < b$ .

*Proof.* From the definition (10) it follows that  $\Omega_n(a) = 0$ . Since  $n$  is even, by Lemma 2, the integrand in  $\Omega_n(b)$  is antisymmetric about the midpoint of the interval of integration and hence  $\Omega_n(b) = 0$ .

For part (b) we observe that  $a, x_1, x_2, \dots, x_{n-1}, b$  are the only zeros of  $\omega_n(x)$ , and hence  $\omega_n(x) < 0$  for  $x < a$  (since  $\omega_n(x)$  is of odd degree). Then  $\omega_n(x) > 0$  for  $a < x < x_1$  and thus

$$\Omega_n(x) > 0 \quad \text{for } a < x \leq x_1.$$

But by Lemma 3, we see that the negative contribution of  $\omega_n(x)$  over  $[x_1, x_2]$  to  $\Omega_n(x)$  is in magnitude less than the positive contribution over  $[a, x_1]$ . Therefore,

$$\Omega_n(x) > 0 \quad \text{for } a < x < x_2,$$

This argument can be repeated to cover the interval  $a < x < x_{n/2}$ . For  $x > x_{n/2}$ , Lemma 2 is employed.  $\blacksquare$

Notice that these arguments can be used to yield

**LEMMA 5.** For  $n$  odd:

- (a)  $\Omega_n(a) = 0, \quad \Omega_n(b) = 2\Omega_n(x_{n/2})$ ;
- (b)  $\Omega_n(x) < 0, \quad a < x \leq b$ .  $\blacksquare$

However, we shall not require this lemma in our analysis of the error in quadrature formulae.

We are now prepared to estimate the error,  $E_{n+1}$ , given by (4a), for the closed Newton-Cotes formulae. We first treat the case of  $n$  even and assume that the integrand,  $f(x)$ , has  $n + 2$  continuous derivatives. By using (10), integration by parts [note that the continuity of  $(d/dx)f[x_0, x_1, \dots, x_n, x]$  is assured by Problem 1.6 of Chapter 6], and Lemma 4a, the error is

$$\begin{aligned} E_{n+1}\{f\} &= \int_a^b \frac{d\Omega_n(x)}{dx} f[x_0, \dots, x_n, x] dx \\ &= \Omega_n(x) f[x_0, \dots, x_n, x] \Big|_a^b - \int_a^b \Omega_n(x) \frac{d}{dx} f[x_0, \dots, x_n, x] dx \\ &= - \int_a^b \Omega_n(x) \frac{d}{dx} f[x_0, \dots, x_n, x] dx. \end{aligned}$$

Hence,

$$E_{n+1}\{f\} = - \int_a^b \Omega_n(x) \frac{f^{(n+2)}(\xi(x))}{(n+2)!} dx$$

(from Problem 1.7 and Corollary 2 to Theorem 1.2, all of Chapter 6).

Now

$$f^{(n+2)}(\xi(x)) \equiv \frac{d}{dx} f[x_0, \dots, x_n, x](n+2)!$$

is continuous by Problem (1.6) of Chapter 6. By Lemma 4b,  $\Omega_n(x) \geq 0$ . Hence, we may apply the mean value theorem for integrals in the above to get

$$E_{n+1}\{f\} = - \frac{f^{(n+2)}(\eta)}{(n+2)!} \int_a^b \Omega_n(x) dx, \quad a < \eta < b.$$

In addition, integration by parts and Lemma 4 yield

$$\begin{aligned} \int_a^b \Omega_n(x) dx &= x\Omega_n(x) \Big|_a^b - \int_a^b x \frac{d}{dx} \Omega_n(x) dx \\ &= - \int_a^b x \omega_n(x) dx > 0. \end{aligned}$$

These results have established

**THEOREM 1.** *Let the points of (9) divide  $[a, b]$  into an even number of equal intervals. Let  $f(x)$  have a continuous derivative of order  $n + 2$  on  $[a, b]$ . Then the error, (4a), in the closed Newton-Cotes quadrature formula, (3), for  $n$  even is*

$$E_{n+1}\{f\} = \frac{K_n}{(n+2)!} f^{(n+2)}(\eta), \quad a < \eta < b;$$

where

$$K_n \equiv \int_a^b x \omega_n(x) dx < 0. \quad \blacksquare$$

We deduce from this theorem the interesting result that the closed Newton-Cotes formula with an even number,  $n$ , of intervals has degree of precision  $n + 1$  (even though the interpolation polynomial employed is of degree  $n$ ).

To treat the case of odd  $n$ , we could employ Lemma 5. This would lead to an error expression containing two terms involving different order derivatives of  $f(x)$ . However, to obtain the simpler form of Theorem 1 we first recall that  $\omega_n(x)$  does not change sign in  $[b - h, b]$ . Then (4a) yields by the mean value theorem for integrals and (1.9) of Chapter 6,

$$\begin{aligned} E_{n+1}\{f\} &= \int_a^{b-h} \omega_n(x)f[x_0, \dots, x_n, x] dx + \int_{b-h}^b \omega_n(x)f[x_0, \dots, x_n, x] dx \\ &= \int_a^{b-h} \omega_n(x)f[x_0, \dots, x_n, x] dx + \frac{f^{(n+1)}(\xi')}{(n+1)!} \int_{b-h}^b \omega_n(x) dx, \\ &\quad a < \xi' < b. \end{aligned}$$

To treat the first integral, we write

$$\omega_n(x) = \omega_{n-1}(x)(x - x_n) \quad \text{and} \quad \Omega_{n-1}(x) = \int_a^x \omega_{n-1}(\xi) d\xi.$$

Then the properties of divided differences given in (1.5) and (1.6) of Chapter 6 permit

$$\begin{aligned} \int_a^{b-h} \omega_n(x)f[x_0, \dots, x_n, x] dx &= \int_a^{b-h} \frac{d\Omega_{n-1}(x)}{dx} (f[x_0, \dots, x_{n-1}, x] \\ &\quad - f[x_0, \dots, x_n]) dx \end{aligned}$$

Now  $n - 1$  is even, and so  $\Omega_{n-1}(a) = \Omega_{n-1}(b - h) = 0$ , or

$$\int_a^{b-h} \frac{d\Omega_{n-1}(x)}{dx} dx = 0.$$

Hence we may neglect the integral involving the constant  $f[x_0, \dots, x_n]$ . For the remaining integral, an integration by parts and application of the mean value theorem for integrals as before yield

$$\begin{aligned} \int_a^{b-h} \omega_n(x)f[x_0, \dots, x_n, x] dx &= -\frac{f^{(n+1)}(\xi'')}{(n+1)!} \int_a^{b-h} \Omega_{n-1}(x) dx, \\ &\quad a < \xi'' < b. \end{aligned}$$

Thus we have deduced that

$$E_{n+1}\{f\} = -[Af^{(n+1)}(\xi') + Bf^{(n+1)}(\xi'')],$$

where

$$A = -\frac{1}{(n+1)!} \int_{b-h}^b \omega_n(x) dx, \quad B = \frac{1}{(n+1)!} \int_a^{b-h} \Omega_{n-1}(x) dx.$$

However, since  $x = b$  is the largest zero of  $\omega_n(x)$  and  $\omega_n(x) > 0$  for  $x > b$ , it follows that  $\omega_n(x) \leq 0$  in  $[b-h, b]$ , and so  $A > 0$ . That  $B > 0$  follows from Lemma 4, since  $n-1$  is even. Thus, if  $f^{(n+1)}(x)$  is continuous on  $[a, b]$  an application of Lemma 1 implies that there exists a point  $\xi$  in  $[\xi', \xi'']$  such that

$$E_{n+1}\{f\} = -(A + B)f^{(n+1)}(\xi).$$

Since

$$\omega_n(x) = \frac{d\Omega_{n-1}(x)}{dx} (x - b),$$

we have through integration by parts and Lemma 4

$$\begin{aligned} \int_a^{b-h} \omega_n(x) dx &= \Omega_{n-1}(x)(x - b) \Big|_a^{b-h} - \int_a^{b-h} \Omega_{n-1}(x) dx \\ &= - \int_a^{b-h} \Omega_{n-1}(x) dx. \end{aligned}$$

Thus

$$A + B = -\frac{1}{(n+1)!} \int_a^b \omega_n(x) dx.$$

In summary, we have

**THEOREM 2.** *If the points of (9) divide  $[a, b]$  into an odd number of equal intervals and  $f(x)$  has a continuous derivative of order  $(n+1)$  on  $[a, b]$  then the error, (4a), in the closed Newton-Cotes quadrature formula (3), for  $n$  odd is*

$$E_{n+1}\{f\} = \frac{K_n}{(n+1)!} f^{(n+1)}(\xi), \quad a < \xi < b;$$

where

$$K_n = \int_a^b \omega_n(x) dx < 0. \quad \blacksquare$$

The formula covered by this theorem has degree of precision  $n$ . Note that the result in Theorem 2 is formally similar to that in (5c).

To express the dependence of the errors given in Theorems 1 and 2 on the interval size,  $h$ , we use the change of variable,  $x = x_0 + th$ , and find

COROLLARY. *Under the hypotheses of Theorems 1 and 2, respectively,*

$$(11) \quad E_{n+1}\{f\} = \begin{cases} \frac{M_n}{(n+2)!} h^{n+3} f^{(n+2)}(\xi), & M_n \equiv \int_0^n t \pi_n(t) dt < 0, \\ & n \text{ even;} \\ \frac{M_n}{(n+1)!} h^{n+2} f^{(n+1)}(\xi), & M_n \equiv \int_0^n \pi_n(t) dt < 0, \\ & n \text{ odd. } \blacksquare \end{cases}$$

Since the closed formulae are exact for polynomials of degree at most  $n+1$  when  $n+1$  is odd, and are exact for polynomials of degree at most  $n$  when  $n+1$  is even, it is generally preferable to employ the odd formulae, i.e., those with an odd number of interpolation points. Also, it clearly does not pay, in general, to add one point to a scheme with even  $n$ ; rather, points should be added in pairs.

Another useful integration formula with equal intervals is found by using the points

$$(12a) \quad x_j = x_0 + jh, \quad j = 0, 1, \dots, n;$$

where

$$(12b) \quad h = \frac{b-a}{n+2}, \quad x_0 = a + h, \quad x_n = b - h.$$

The endpoints are then labeled  $x_{-1} = a$  and  $x_{n+1} = b$ . Since we do not employ the endpoints in formula (3), it is now called an *open Newton-Cotes formula*. In this procedure, all  $n+1$  points of interpolation are interior to the interval of integration.

To examine the error we introduce, in place of  $\Omega_n(x)$ , the functions  $J_n(x)$  defined by

$$(13) \quad J_n(x) = \int_a^x \omega_n(\xi) d\xi, \quad n = 1, 2, \dots$$

These differ from the functions in (10) since now  $a < x_0$  and  $x_n < b$ . However, as in the proof of Lemma 4, it follows that for  $n$  even

$$J_n(a) = J_n(b) = 0; \quad J_n(x) < 0, \quad a < x < b.$$

Then, exactly as in the derivation of Theorem 1, we have

**THEOREM 1'.** *Replace “(9)” by “(12)” and “closed” by “open” in the statement of Theorem 1. Then the formula for  $E_{n+1}\{f\}$  becomes*

$$E_{n+1}\{f\} = \frac{K_n'}{(n+2)!} f^{(n+2)}(\xi), \quad a < \xi < b; \quad n \text{ even,}$$

where

$$K_n' = \int_a^b x \omega_n(x) dx > 0. \quad \blacksquare$$

Similarly, for  $n$  odd, by procedures analogous to those for closed-type formulae, we find

**THEOREM 2'.** *Use the hypothesis of Theorem 1', but with  $n$  odd. Then*

$$E_{n+1}\{f\} = \frac{K_n'}{(n+1)!} f^{(n+1)}(\xi), \quad a < \xi < b; \quad n \text{ odd},$$

where

$$K_n' = \int_a^b \omega_n(x) dx > 0. \quad \blacksquare$$

These errors for the open formula may be expressed in terms of the spacing  $h$  as

**COROLLARY.** *Under the hypothesis of Theorems 1' and 2', respectively,*

$$E_{n+1}\{f\} = \frac{M_n'}{(n+2)!} h^{n+3} f^{(n+2)}(\xi),$$

(14a)

$$M_n' = \int_{-1}^{n+1} t \pi_n(t) dt > 0, \quad n \text{ even};$$

$$E_{n+1}\{f\} = \frac{M_n'}{(n+1)!} h^{n+2} f^{(n+1)}(\xi),$$

(14b)

$$M_n' = \int_{-1}^{n+1} \pi_n(t) dt > 0, \quad n \text{ odd}. \quad \blacksquare$$

Again we find that for  $n$  even, the degree of precision is  $n + 1$ , while for  $n$  odd it is only  $n$ . A comparison of (11) and (14) indicates that only the values of the coefficients  $M_n$  and  $M_n'$  differ in the form of the error estimates for open and closed Newton-Cotes formulae based on the same number,  $n + 1$ , of nodes. [However, for any fixed number of intervals in  $[a, b]$  say  $m$ , the closed formulae use  $m + 1$  nodes and the open formulae use  $m - 1$  nodes. Hence, the closed method has a degree of precision two more than the open method on this basis, but requires two more evaluations of the function  $f(x)$ .]

There are useful quadrature formulae that are neither open nor closed but which have uniformly spaced nodes [e.g., see Problems 2, 3, and 5]. The formulae of Problems 2 and 3 are the basis for Adam's method for the numerical solution of ordinary differential equations (see Table 2.1 of Chapter 8).

## 1.2. Determination of the Coefficients

Let the interpolation points be equally spaced, say of the form  $x_j = x_0 + jh$ ,  $j = 0, 1, \dots, n$ , and let the endpoints  $a, b$  of the integration interval also be of this form, say  $a = x_p = x_0 + ph$  and  $b = x_q = x_0 + qh$ , (but not necessarily interpolation points). Then the coefficients,  $w_{n,j}$ , for quadrature formulae of the form (3) can be written as, using  $x = x_0 + th$ ,

$$\begin{aligned} w_{n,j} &= \int_a^b \phi_{n,j}(x) dx \\ &= h \int_p^q \prod_{\substack{k=0 \\ (k \neq j)}}^n \frac{t - k}{j - k} dt \\ &= (-1)^{n-j} \frac{h}{n!} \binom{n}{j} \int_p^q \frac{\pi_n(t)}{t - j} dt, \quad j = 0, 1, \dots, n. \end{aligned}$$

Thus if we define the quantities

$$(15a) \quad A_{n,j}(p, q) = \frac{(-1)^{n-j}}{n!} \binom{n}{j} \int_p^q \frac{\pi_n(t)}{t - j} dt, \quad j = 0, 1, \dots, n;$$

the coefficients are simply

$$(15b) \quad w_{n,j} = h A_{n,j}(p, q).$$

In the special case of the closed Newton-Cotes schemes we have  $p = 0$  and  $q = n$ ; for the open schemes of Section 1.1,  $p = -1$  and  $q = n + 1$ . It should be noted that the  $A_{n,j}(p, q)$  are independent of the spacing,  $h$ , and thus may be tabulated as functions of the parameters  $n, j, p$ , and  $q$ . Appropriate coefficients for a particular spacing,  $h$ , are then determined by using (15b). Since the integrand in (15a) is a polynomial of degree  $n$  the  $A_{n,j}(p, q)$  will all be rational numbers for rational  $p$  and  $q$ . Further, we note that  $A_{n,j}(0, n) = A_{n,n-j}(0, n)$  and more generally,  $A_{n,j}(-r, n+r) = A_{n,n-j}(-r, n+r)$  for any  $r$ . For  $p$  and  $q$  of these important special forms, we need only tabulate the values for  $j \leq n/2$ . Tables 1 and 2 list the simplest closed and open Newton-Cotes formulae. The coefficients  $A_{n,j}(n, n+1)$  are to be found in Problem 6, for  $n = 1, 2, 3$ , and 4.

An alternative indirect procedure, called the *method of undetermined coefficients*, can also be employed to determine the coefficients for the quadrature formula (3). This method is quite practical for unequally spaced nodes as well as for fairly large values of  $n$ . In addition, it can be used to prove

**Table 1** *Closed Newton-Cotes Formulae*

---


$$\int_{x_0}^{x_1} f(x) dx = \frac{h}{2} (f_0 + f_1) - \frac{h^3}{12} f^{(2)}(\xi), \quad x_0 < \xi < x_1, \quad (\text{trapezoidal rule})$$

$$\int_{x_0}^{x_2} f(x) dx = \frac{h}{3} (f_0 + 4f_1 + f_2) - \frac{h^5}{90} f^{(4)}(\xi), \quad x_0 < \xi < x_2,$$

$$(\text{Simpson's rule})$$

$$\int_{x_0}^{x_3} f(x) dx = \frac{3h}{8} (f_0 + 3f_1 + 3f_2 + f_3) - \frac{3h^5}{80} f^{(4)}(\xi), \quad x_0 < \xi < x_3.$$

$$\int_{x_0}^{x_4} f(x) dx = \frac{2h}{45} (7f_0 + 32f_1 + 12f_2 + 32f_3 + 7f_4) - \frac{8h^7}{945} f^{(6)}(\xi),$$

$$x_0 < \xi < x_4.$$

$$\int_{x_0}^{x_5} f(x) dx = \frac{5h}{288} (19f_0 + 75f_1 + 50f_2 + 50f_3 + 75f_4 + 19f_5)$$

$$- \frac{275h^7}{12096} f^{(6)}(\xi), \quad x_0 < \xi < x_5.$$


---

**Table 2** *Open Newton-Cotes Formulae*

---


$$\int_{x_0}^{x_2} f(x) dx = 2hf_1 + \frac{h^3}{3} f^{(2)}(\xi), \quad x_0 < \xi < x_2, \quad (\text{midpoint rule})$$

$$\int_{x_0}^{x_3} f(x) dx = \frac{3h}{2} (f_1 + f_2) + \frac{3h^3}{4} f^{(2)}(\xi), \quad x_0 < \xi < x_3.$$

$$\int_{x_0}^{x_4} f(x) dx = \frac{4h}{3} (2f_1 - f_2 + 2f_3) + \frac{28h^5}{90} f^{(4)}(\xi), \quad x_0 < \xi < x_4.$$

$$\int_{x_0}^{x_5} f(x) dx = \frac{5h}{24} (11f_1 + f_2 + f_3 + 11f_4) + \frac{95h^5}{144} f^{(4)}(\xi), \quad x_0 < \xi < x_5.$$

$$\int_{x_0}^{x_6} f(x) dx = \frac{6h}{20} (11f_1 - 14f_2 + 26f_3 - 14f_4 + 11f_5) + \frac{41h^7}{140} f^{(6)}(\xi),$$

$$x_0 < \xi < x_6.$$

$$\int_{x_0}^{x_7} f(x) dx = \frac{7h}{1440} (611f_1 - 453f_2 + 562f_3 + 562f_4 - 453f_5 + 611f_6)$$

$$+ \frac{5257h^7}{8640} f^{(6)}(\xi), \quad x_0 < \xi < x_7.$$


---

**THEOREM 3.** *A quadrature formula which uses  $n + 1$  distinct nodes is an interpolatory formula iff it has degree of precision at least  $n$ .*

*Proof.* The necessity has been demonstrated by equation (4b). To prove sufficiency, we let the  $n + 1$  distinct points  $x_j$ ,  $j = 0, 1, \dots, n$ , be the given nodes. If the quadrature formula, with these points and coefficients  $\alpha_j$ , has degree of precision at least  $n$  in approximating  $\int_a^b f(x) dx$ , then we must have

$$(16) \quad \begin{aligned} \sum_{j=0}^n \alpha_j &= (b - a) \\ \sum_{j=0}^n \alpha_j x_j &= \frac{1}{2}(b^2 - a^2) \\ &\vdots \\ \sum_{j=0}^n \alpha_j x_j^n &= \frac{1}{n+1} (b^{n+1} - a^{n+1}). \end{aligned}$$

This may be considered as a system of  $n + 1$  equations for the determination of the  $n + 1$  coefficients,  $\alpha_j$ . Since the coefficient matrix of the system (16) is of the Vandermonde form, and the  $x_j$  are distinct, there exists a *unique solution*. On the other hand, note that the interpolatory formula (3), with the same points  $x_j$ , is exact when applied to the powers  $1, x, \dots, x^n$ . Hence, the system (16) is satisfied with the  $\alpha_j$  replaced by the  $w_{n,j}$ . The fact that (16) has a unique solution shows  $\alpha_j = w_{n,j}$ . ■

We shall see that most of the popular quadrature formulae are interpolatory. This follows by means of Theorem 3 and its extension, in Section 4, to weighted quadrature formulae.

The *method of undetermined coefficients* consists in solving the system (16) for the  $\alpha_j$ . To determine the degree of precision of an interpolatory quadrature formula, we simply form the quantities

$$E_{n+1}\{x^k\} \equiv \frac{1}{k+1} (b^{k+1} - a^{k+1}) - \sum_{j=0}^n \alpha_j x_j^k, \quad k = n+1, n+2, \dots;$$

and determine the least integer  $k$  such that  $E_{n+1}\{x^k\} \neq 0$ . The degree of precision is then  $k - 1$ .

If it is known that the error in the integration formula has the form  $E_{n+1}\{f\} = A_m f^{(m)}(\xi)$  for some integer  $m$ , then we can determine the coefficient  $A_m$  by this method. That is, we must have  $m = k$  where  $k - 1$  is the determined degree of precision. Hence, use  $f(x) \equiv x^k$  to get

$$(17) \quad E_{n+1}\{x^k\} = \frac{1}{k+1} (b^{k+1} - a^{k+1}) - \sum_{j=0}^n \alpha_j x_j^k = k! A_k,$$

which can be solved for  $A_k$ . For example, in (11) for the closed Newton-Cotes formulae, we may use the quantities  $A_{n+2}$  or  $A_{n+1}$  to evaluate the appropriate coefficient  $M_n$ .

As an illustration of the application of the method of undetermined coefficients, we consider the closed formula with one segment,  $n = 1$ , and the two nodes  $x_0 = a$  and  $x_1 = b$ . The system (16) now becomes

$$\begin{aligned} \alpha_0 + \alpha_1 &= (b - a); \\ a\alpha_0 + b\alpha_1 &= \frac{1}{2}(b^2 - a^2); \end{aligned}$$

and the solution of this system is

$$\alpha_0 = \alpha_1 = \frac{1}{2}(b - a).$$

To determine the degree of precision we apply the formula to  $x^2, x^3, \dots$ , and get first

$$\begin{aligned} E_2\{x^2\} &= \frac{1}{3}(b^3 - a^3) - \frac{1}{2}(b - a)(a^2 + b^2) \\ &= -\frac{1}{6}(b - a)^3 \neq 0. \end{aligned}$$

Thus the degree of precision is 1, as we knew since it is a closed Newton-Cotes formula with  $n = 1$ . The error can be written as

$$E_2\{f\} = A_2 f^{(2)}(\xi)$$

where  $A_2 = (M_1/2!)h^3$  and  $h = b - a$ . By using  $f(x) = x^2$  we find that

$$E_2\{x^2\} = -\frac{1}{6}(b - a)^3 = 2A_2.$$

Thus  $A_2 = -\frac{1}{12}(b - a)^3$  and  $M_1 = -\frac{1}{6}$ . The formula determined above is the familiar *trapezoidal rule* which can now be written as

$$(18) \quad \int_a^b f(x) dx = \frac{b - a}{2} [f(a) + f(b)] - \frac{(b - a)^3}{12} f^{(2)}(\xi), \quad a < \xi < b.$$

## PROBLEMS, SECTION 1

1. Add to the hypothesis of Lemma 1:  $r$  of the coefficients  $(\alpha_k)$  are non-zero. What is the smallest integer  $r$  for which the stronger conclusion  $\xi \in (a, b)$  is valid?
2. From equation (4.19) of Chapter 6, derive the formula

$$\int_x^{x+h} f(\xi) d\xi = h[d_0 f(x) + d_1 \Delta f + d_2 \Delta^2 f + \cdots + d_m \Delta^m f] + h^{m+2} d_{m+1} f^{(m+1)}(\xi)$$

where  $d_0 = 1$ , and recursively

$$d_k - \frac{d_{k-1}}{2} + \frac{d_{k-2}}{3} + \cdots + (-1)^k \frac{d_0}{k+1} = 0, \quad k = 1, 2, \dots, m+1;$$

$$x < \xi < x + mh; \quad \Delta^k f \equiv \sum_{j=0}^k (-1)^{k-j} \binom{k}{j} f(x + jh).$$

[Hint: The coefficients  $\{d_i\}$  satisfy the identity in  $\Delta$ ,

$$\Delta \equiv \log(I + \Delta)(d_0 I + d_1 \Delta + \cdots + d_m \Delta^m + \cdots).$$

Check the following listing:

$k$	0	1	2	3	4	5
$d_k$	1	$\frac{1}{2}$	$-\frac{1}{12}$	$\frac{1}{24}$	$-\frac{19}{720}$	$\frac{3}{160}$

3. From equation (4.19) of Chapter 6, derive the formula

$$\int_{x-h}^x f(\xi) d\xi = h[e_0 f(x) + e_1 \Delta f + e_2 \Delta^2 f + \cdots + e_m \Delta^m f] + h^{m+2} e_{m+1} f^{(m+1)}(\eta), \quad \text{where } e_0 = 1,$$

and recursively,

$$e_k - \frac{e_{k-1}}{2} + \frac{e_{k-2}}{3} + \cdots + (-1)^k \frac{e_0}{k+1} = (-1)^k, \quad k = 1, 2, \dots, m+1;$$

$$x - h < \eta < x + mh; \quad \Delta \text{ as defined in Problem 2.}$$

[Hint: The coefficients  $\{e_i\}$  satisfy the identity

$$\begin{aligned} \Delta - \Delta^2 + \Delta^3 + \cdots + (-1)^m \Delta^{m+1} + \cdots \\ \equiv \left( \Delta - \frac{\Delta^2}{2} + \frac{\Delta^3}{3} + \cdots \right) (e_0 I + e_1 \Delta + e_2 \Delta^2 + \cdots). \end{aligned}$$

Check the following listing:

$k$	0	1	2	3	4	5
$e_k$	1	$-\frac{1}{2}$	$-\frac{5}{12}$	$-\frac{3}{8}$	$\frac{251}{720}$	$-\frac{95}{288}$

4. With  $\{d_i\}$  and  $\{e_i\}$  as defined in Problems 2 and 3, and  $e_{-1} = 0$ , show that

$$d_m = e_m + e_{m-1}.$$

5. From Everett's form of the interpolation polynomial, (3.17) of Chapter 6, define the coefficients  $\{r_i\}$  and  $C_m$  of the formula

$$\begin{aligned} \frac{1}{h} \int_{x_0}^{x_1} f(x) dx = r_0(f_0 + f_1) + r_1(\delta^2 f_0 + \delta^2 f_1) + \cdots \\ + r_m(\delta^{2m} f_0 + \delta^{2m} f_1) + C_m h^{2m+2} f^{(2m+2)}(\xi), \end{aligned}$$

where

$$x_{-m} < \xi < x_{m+1}, \quad x_j = x_0 + jh.$$

Explicitly find  $r_0$  and  $r_1$ .

6. Make a table listing the coefficients  $A_{n,j}(n, n+1)$ , for  $n = 1, 2, 3$ , and 4. [Hint: Use the result of Problem 3.]

7. Prove Lemma 5. [Hint: Let  $n = 2m+1$  and  $0 < \epsilon \leq \frac{1}{2}$ . Show that  $|\pi_n(t+1)/\pi_n(t)| < 1$  for  $t = m - \epsilon$ .]

## 2. ROUNDOFF ERRORS AND UNIFORM COEFFICIENT FORMULAE

In almost all practical applications of integration formulae of the form (0.2) the exact function values,  $f(x)$ , will not be available. This fact is usually due to limitations in the calculation of these function values (or in their measurement). Thus the quantities actually employed may be

written as  $\tilde{f}(x_j) = f(x_j) + \rho_j$ , where  $\rho_j$  is the local roundoff error made in computing (or error in measuring)  $f(x_j)$ . The error in approximating (0.1) by (0.2) with these values is then

$$\int_a^b f(x) dx - \sum_{j=1}^n \alpha_j \tilde{f}(x_j) = E_n\{f\} - R_n\{f\},$$

where  $E_n\{f\} = I\{f\} - I_n\{f\}$  is the quadrature or *truncation error* and the *accumulated roundoff error* is:

$$(1) \quad R_n\{f\} = \sum_{j=1}^n \alpha_j \rho_j.$$

If, as is frequently the case, we know that the local errors are bounded, say  $|\rho_j| \leq \rho$  for  $j = 1, 2, \dots, n$ , then

$$(2) \quad |R_n\{f\}| \leq \rho \sum_{j=1}^n |\alpha_j|.$$

Let us also assume that the formula (0.2) has degree of precision  $\geq 0$ , i.e., that at least a constant is integrated correctly. Then with the integrand  $f(x) \equiv 1$  we find

$$(3) \quad \sum_{j=1}^n \alpha_j = (b - a).$$

Thus if all the coefficients,  $\alpha_j$ , are of one sign the bound (2) becomes

$$(4) \quad |R_n\{f\}| \leq \rho |b - a|.$$

If the coefficients are not all of one sign, then clearly,

$$\sum_{j=1}^n |\alpha_j| > \left| \sum_{j=1}^n \alpha_j \right| = |b - a|$$

and a larger maximum accumulated roundoff is possible. To attain the maximum value requires that

$$\rho_j = \rho \operatorname{sign} \alpha_j$$

which is, of course, very special. By comparing (4) and (2), we find a practical advantage in having the coefficients of a quadrature formula all of one sign, especially if the truncation error is smaller than the rounding error.

By introducing *statistical notions of roundoff* (or measurement) errors, we can, in fact, show that it is of even greater advantage to have all of the coefficients of the same value. There are several ways in which the statistical notion of “randomness” of the local errors can be introduced. For

instance, suppose we ask for those coefficients,  $\alpha_j$ , for which some measure of  $R_n\{f\}$  is a minimum for all functions  $f$ , of some particular class,  $F$ . Since the errors  $\rho_j = \rho_j\{f\}$  usually depend in an extremely complicated way on the functions  $f$ , direct attempts at such a minimization do not seem possible. However, as  $f$  varies over the class  $F$  the errors  $\rho_j\{f\}$ , can be expected to vary in an erratic manner. By making specific assumptions about the nature of this variation and introducing a measure of "volume" in  $F$  we can calculate various "averages" of  $R_n\{f\}$  over  $F$ .

Specifically, let us consider for  $F$  a one parameter family of functions of  $x$ , say  $f(x; \tau)$ , where the parameter  $\tau$  ranges over  $0 \leq \tau \leq T$ . The roundoff error in evaluating  $f(x_k; \tau)$  for each  $\tau$  and  $k = 1, 2, \dots, n$  will be denoted by  $\rho_k(\tau)$ . We assume that  $|\rho_k(\tau)| \leq \rho$  and that all values in this range are equally likely for each value of  $\tau$ ; or in particular, we assume that the "average" roundoff over the family  $F$  vanishes; i.e., that

$$(5a) \quad \bar{\rho}_k \equiv \frac{1}{T} \int_0^T \rho_k(\tau) d\tau = 0, \quad k = 1, 2, \dots, n.$$

Further, we assume that the roundoff errors at distinct points  $x_j$  and  $x_k$  are uncorrelated; i.e.,

$$(5b) \quad \int_0^T \rho_j(\tau) \rho_k(\tau) d\tau = 0, \quad \text{if } j \neq k.$$

In effect, this means that the error committed at  $x_j$  is independent of the error at  $x_k$  for all the functions,  $f(x; \tau)$  in  $F$ . Finally, let us make the assumption that all the local errors have the same mean-square value, say,

$$(5c) \quad \sigma^2 \equiv \frac{1}{T} \int_0^T \rho_j^2(\tau) d\tau, \quad j = 1, 2, \dots, n.$$

Note that  $\sigma \leq \rho$ .

We now consider some measures of the accumulated roundoff error for the family  $F$ . We define for any  $\tau$  in  $0 \leq \tau \leq T$ :

$$\begin{aligned} R_n(\tau) &= R_n\{f(x; \tau)\} \\ &= \sum_{j=1}^n \alpha_j \rho_j(\tau). \end{aligned}$$

The *mean-accumulated roundoff* for the family  $F$  is, by (5a),

$$\bar{R}_n \equiv \frac{1}{T} \int_0^T R_n(\tau) d\tau = \sum_{j=1}^n \alpha_j \bar{\rho}_j = 0.$$

Thus the coefficients,  $\alpha_j$ , of the quadrature formula have no effect on the

average accumulated roundoff. Next we compute the *mean-square roundoff error*, using (5b and c), to get

$$(6) \quad \begin{aligned} r_n^2 &\equiv \frac{1}{T} \int_0^T R_n^2(\tau) d\tau \\ &= \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j \frac{1}{T} \int_0^T \rho_i(\tau) \rho_j(\tau) d\tau \\ &= \sigma^2 \sum_{j=1}^n \alpha_j^2. \end{aligned}$$

The coefficients clearly have an effect on this measure of the roundoff error. Let us seek  $\{\alpha_j\}$  which minimize the sum in the last line of (6) but which also satisfy (3). This problem is easily solved by using the method of Lagrange multipliers and we find for the minimizing coefficients:

$$(7) \quad \alpha_1 = \alpha_2 = \cdots = \alpha_n = \frac{b - a}{n}.$$

Thus to reduce the *root mean-square roundoff error*,  $r_n$ , as much as possible, while retaining at least zero order precision, the coefficients should be equal. Using (7) in (6) yields for the minimum of  $r_n$  the value

$$(8) \quad r_n = \frac{\sigma|b - a|}{\sqrt{n}}.$$

This result is somewhat surprising as it indicates that the root mean-square roundoff error actually decreases as the number of quadrature points (and hence of sources of error) increases! It should be noted that when the weights are equal the bound (4) applies. Compare the maximum bound,  $\rho|b - a|$ , with the statistical result in (8) to find a reduction in the dependence on  $n$  by the factor  $1/\sqrt{n}$  for the statistical estimate. This is a common feature of statistical estimates of roundoff. It is frequently found in practice that the statistical estimate is a more realistic approximation of the error than is the maximum-type estimate.

These results can be interpreted in a slightly different, perhaps more familiar and intuitive way. We think of a family of calculations of the quadrature formula applied to the same function,  $f(x)$ . Each calculation of the family is determined by the particular rounding employed in the set of values  $f(x_j)$ ,  $j = 1, 2, \dots, n$ . That is,  $\rho_j(\tau)$  is the roundoff error in the computation characterized by the parameter value  $\tau$  when computing  $f(x_j)$ . Of course, any fixed program for an electronic computer is represented by a single value of the parameter. Thus, the intended interpretation is not repeating the same calculation, but rather, altering the computational procedure slightly each time. The above averages are then averages over an

appropriate family of calculations. With this interpretation numerical experiments using “randomly” generated and independent rounding errors are easily devised.

We shall find that equal coefficient formulae of the form (0.2) cannot approximate integrals of the form (0.1), with degree of precision  $n$  for all values of  $n$  (see, however, Subsection 4.1).

## 2.1. Determination of Uniform Coefficient Formulae

An integration formula with uniform coefficients and  $n$  nodes has the simple form

$$(9a) \quad I_n\{f\} \equiv \alpha_n \sum_{j=1}^n f(x_j).$$

In order that this yield the exact result for

$$I\{f\} = \int_a^b f(x) dx$$

when  $f(x) \equiv 1$ , it is clear that

$$(9b) \quad \alpha_n = \frac{b - a}{n}.$$

We now try to determine the  $n$  nodes,  $x_j$ , such that (9) has as high a degree of precision as possible. Thus we impose the  $n$  conditions,  $I_n\{x^\nu\} = I\{x^\nu\}$  for  $\nu = 1, 2, \dots, n$  to get

$$(10) \quad \alpha_n \sum_{j=1}^n x_j^\nu = \frac{1}{\nu + 1} (b^{\nu+1} - a^{\nu+1}), \quad \nu = 1, 2, \dots, n.$$

If these equations have as solution  $n$  distinct real values,  $x_j$ , then the corresponding quadrature formula (9) has degree of precision at least  $n$ , while only  $n$  nodes are employed (as in the Newton-Cotes formulae for an even number of uniform intervals or odd number of nodes). Thus, by Theorem 1.3, we can conclude that such equal coefficient formulae are interpolatory. The error estimates of Section 1 are then applicable [say, of the form (1.4)].

Let us set  $n = 1$  in (9) and (10). We find  $\alpha_1 = b - a$ ,  $x_1 = \frac{1}{2}(b + a)$ , and the quadrature formula is simply the midpoint rule,

$$I_1\{f\} = (b - a)f\left(\frac{b + a}{2}\right),$$

which is clearly exact for linear functions.

For  $n = 2$ , the system (10) is easily reduced to a quadratic equation which yields the nodes

$$x_1 = \frac{b+a}{2} - \frac{b-a}{2\sqrt{3}}, \quad x_2 = \frac{b+a}{2} + \frac{b-a}{2\sqrt{3}}.$$

Note that in each of these cases the nodes are symmetrically located about the center of the interval of integration.

In general, the solution of the system (10) determines the  $n$ th degree polynomial

$$P_n(x) \equiv (x - x_1)(x - x_2) \cdots (x - x_n)$$

whose roots are the required nodes. This polynomial can be written in the form

$$(11) \quad P_n(x) = x^n + \sigma_1 x^{n-1} + \sigma_2 x^{n-2} + \cdots + \sigma_n,$$

where the coefficients are the classical elementary symmetric functions of the roots, i.e.,

$$\begin{aligned} \sigma_1 &= -(x_1 + x_2 + \cdots + x_n) \\ \sigma_2 &= (x_1 x_2 + x_1 x_3 + \cdots + x_{n-1} x_n) \\ &\vdots \\ \sigma_n &= (-1)^n x_1 x_2 \cdots x_n. \end{aligned}$$

However, the values of these symmetric functions can be obtained from the sums of the powers of the roots

$$S_\nu \equiv x_1^\nu + x_2^\nu + \cdots + x_n^\nu, \quad \nu = 1, 2, \dots, n,$$

which are directly determined in (10). The relations between the  $\sigma_j$  and the  $S_\nu$  are known as *Newton's identities* (see Problem 2),

$$\begin{aligned} S_1 + \sigma_1 &= 0 \\ S_2 + S_1 \sigma_1 + 2\sigma_2 &= 0 \\ &\vdots \\ S_n + S_{n-1} \sigma_1 + S_{n-2} \sigma_2 + \cdots & \\ &+ S_1 \sigma_{n-1} + n\sigma_n = 0. \end{aligned} \tag{12}$$

Thus the determination of the nodes has been reduced to finding the roots of the polynomial (11). The coefficients of this polynomial are recursively computed from (12) by using the known values of the  $S_\nu$ .

The nodes for any interval  $[a, b]$  are easily obtained from the nodes for

the special interval  $[-1, 1]$ . For this purpose we introduce the usual linear change of variable

$$(13) \quad x = \frac{b+a}{2} + y \frac{b-a}{2},$$

and then note that

$$I\{f\} \equiv \int_a^b f(x) dx = \frac{b-a}{2} \int_{-1}^1 g(y) dy \equiv \frac{b-a}{2} J\{g(y)\},$$

where

$$g(y) \equiv f\left(\frac{b+a}{2} + y \frac{b-a}{2}\right).$$

The  $n$ -point uniform coefficient quadrature formula which approximates  $J\{g\}$  is written as

$$J_n\{g\} = \beta_n \sum_{j=1}^n g(y_j).$$

In order that this formula have degree of precision at least  $n$ , we must have  $J\{y^\nu\} = J_n\{y^\nu\}$  for  $\nu = 0, 1, \dots, n$ . Thus  $\beta_n = 2/n$  and

$$(14) \quad S_\nu \equiv \sum_{j=1}^n y_j^\nu = \frac{n}{2(\nu+1)} [1 + (-1)^\nu], \quad \nu = 1, 2, \dots, n.$$

We note that the odd order power sums now vanish, i.e.,  $S_1 = S_3 = \dots = 0$ . Newton's identities (12) become in this case

$$\begin{aligned} \sigma_1 &= 0 \\ \frac{n}{3} + 2\sigma_2 &= 0 \\ (15) \quad \sigma_3 &= 0 \\ \frac{n}{5} + \frac{n}{3}\sigma_2 + 4\sigma_4 &= 0 \\ \sigma_5 &= 0 \\ &\vdots \end{aligned}$$

Thus we find that the odd order elementary symmetric functions vanish and the polynomial for the determination of the nodes,  $y_j$ , becomes

$$(16) \quad P_n(y) = y^n + \sigma_2 y^{n-2} + \sigma_4 y^{n-4} + \dots$$

The roots,  $y_j$ , of  $P_n(y) = 0$  are thus symmetric with respect to the origin and if  $n$  is odd, then  $y = 0$  is a root. Using the transformation (13) we obtain for the nodes of the general quadrature formula (9a) the values

$$x_j = \frac{b+a}{2} + y_j \frac{b-a}{2}.$$

From the properties of the  $y_j$ , it follows that the nodes  $x_j$  are symmetrically located with respect to the midpoint of the interval of integration and that for  $n$  odd the midpoint is a node. If  $n$  is an even integer then by the symmetry of the  $y_j$ , we have

$$\sum_{j=1}^n y_j^{n+1} = 0.$$

That is,  $J_n\{y^{n+1}\} = J\{y^{n+1}\}$  and the degree of precision is  $n + 1$ , when an even number,  $n$ , of nodes is employed. The same must, of course, be true in the general case (9) for  $n$  even.

In order to determine an  $n$ -point quadrature formula of the form (9) which has degree of precision at least  $n$ , the polynomial (16) must have  $n$  real distinct roots. However, for  $n = 8$  and for all  $n \geq 10$  it can be shown that a pair of complex roots occurs. For  $n \leq 7$  and  $n = 9$  the roots have the required properties and are known to many decimals. We list in Table 1 these roots in  $0 \leq y \leq 1$ ; the others are obtained by symmetry.

**Table 1.**

	$n = 1$	$n = 2$	$n = 3$	$n = 4$
$y_j$	0.0	0.57735 02692	0.0 0.70710 67812	0.18759 24741 0.79465 44723
	$n = 5$	$n = 6$	$n = 7$	$n = 9$
$y_j$	0.0 0.37454 14096 0.83249 74870	0.26663 54015 0.42251 86538 0.86624 68181	0.0 0.32391 18105 0.52965 67753 0.88386 17008	0.0 0.16790 61842 0.52876 17831 0.60101 86554 0.91158 93077

Quadrature formulae with uniform coefficients and any number of nodes arise in Subsection 4.1, in another setting.

## PROBLEMS, SECTION 2

1. Verify that (7) characterizes the solution to the problem of minimizing

$$\sum_{j=1}^n \alpha_j^2 \text{ subject to } \sum_{j=1}^n \alpha_j = b - a.$$

**2. Verify Newton's identities (12).**

[Hint: Let  $P_n(x) \equiv \prod_{j=1}^n (x - x_j) \equiv \sum_{k=0}^n \sigma_k x^{n-k}$ ,  $\sigma_0 \equiv 1$ .

$$P_n'(x) = \frac{d}{dx} \prod_{j=1}^n (x - x_j) = \sum_{j=1}^n \frac{P_n(x)}{x - x_j} = \sum_{j=1}^n \frac{P_n(x) - P_n(x_j)}{x - x_j},$$

But,

$$\begin{aligned} \frac{P_n(x) - P_n(x_j)}{x - x_j} &= \sum_{k=0}^n \sigma_k \left( \frac{x^{n-k} - x_j^{n-k}}{x - x_j} \right) \\ &= \sum_{k=0}^{n-1} \sigma_k (x^{n-k-1} + x_j x^{n-k-2} + \cdots + x_j^{n-k-2} x + x_j^{n-k-1}) \\ &= x^{n-1} + \sum_{p=1}^{n-1} (\sigma_p + \sigma_{p-1} x_j + \cdots + \sigma_1 x_j^{p-1} + x_j^p) x^{n-p-1}. \end{aligned}$$

Hence

$$P_n'(x) = nx^{n-1} + \sum_{p=1}^{n-1} (n\sigma_p + S_1\sigma_{p-1} + \cdots + S_{p-1}\sigma_1 + S_p) x^{n-p-1}.$$

**3. GAUSSIAN QUADRATURE; MAXIMUM DEGREE OF PRECISION**

In Section 1 it is shown that, given  $n + 1$  fixed nodes, we can determine the coefficients of a quadrature formula which has degree of precision at least  $n$  (by forming the appropriate interpolatory formula). In Subsection 2.1 we investigated the problem of determining  $n$  nodes such that all the coefficients are equal and the degree of precision is at least  $n$ . Now we allow a choice of both  $n$  nodes and  $n$  coefficients in order to determine formulae with maximum degree of precision. Of course, the degree of precision for such a formula will not be less than the corresponding degree for the interpolatory formula using the same nodes. Hence by Theorem 1.3 we conclude that *the quadrature formula with maximum degree of precision is interpolatory*.

If the formula is to have the  $n$  nodes  $x_1, x_2, \dots, x_n$ , it can be written as

$$(1) \quad \int_a^b f(x) dx = \sum_{j=1}^n \alpha_j f(x_j) + E_n\{f\}.$$

However, since it must be interpolatory, the error can be written as, recalling (1.4a),

$$(2) \quad E_n\{f\} = \int_a^b p_n(x) f[x_1, \dots, x_n, x] dx,$$

where we have introduced

$$(3) \quad p_n(x) \equiv (x - x_1)(x - x_2) \cdots (x - x_n).$$

Clearly,  $E_n\{f\} = 0$  if  $f(x)$  is a polynomial of degree  $n - 1$  or less. We seek points  $x_j$  such that the error also vanishes when  $f(x)$  is any polynomial of degree  $n + r$  where  $r = 1, 2, \dots, m$  and  $m$  is to be as large as possible.

To determine such nodes we first recall that the  $n$ th divided difference of any polynomial of degree  $n + r$  is a polynomial of degree at most  $r$ . (This follows by repeated application,  $n$  times, of the result in Problem 1.2 of Chapter 6.) Thus from (2) we conclude that the necessary and sufficient conditions for  $E_n\{f\}$  to vanish for all polynomials,  $f$ , of degree  $n + m$  are that

$$(4) \quad \int_a^b p_n(x)x^r dx = 0, \quad r = 0, 1, \dots, m.$$

However, these are just the conditions that the polynomial  $p_n(x)$  be orthogonal, over  $[a, b]$ , to all polynomials of degree at most  $m$ . In fact, if we take for  $p_n(x)$  the  $n$ th orthogonal polynomial, then (4) is satisfied for  $m = n - 1$ . Further, (4) cannot be satisfied for  $m = n$  or else  $p_n(x)$  would have to vanish identically which is impossible. These results may be summarized as

**THEOREM 1.** *The quadrature formula in (1) can have the maximum degree of precision  $2n - 1$ . This is attained iff the  $n$  nodes,  $x_j$ , are the zeros of  $p_n(x)$ , the  $n$ th orthogonal polynomial over  $[a, b]$ , and the formula is interpolatory.* ■

The formulae determined by Theorem 1 are called Gaussian quadrature formulae. From Theorem 3.4 of Chapter 5 it follows that the nodes are all interior to the interval of integration,  $(a, b)$ . The coefficients in these formulae are easily obtained (once the nodes are determined) since they are interpolatory; we get as in (1.3b)

$$(5) \quad \alpha_j = \frac{1}{p_n'(x_j)} \int_a^b \frac{p_n(x)}{x - x_j} dx, \quad j = 1, 2, \dots, n.$$

Although it is not apparent from this expression, we have

**THEOREM 2.** *The coefficients,  $\alpha_j$ , in the Gaussian quadrature formulae are positive for all  $j = 1, 2, \dots, n$  and all  $n$ .*

*Proof.* Since the Gaussian quadrature formula with  $n$  nodes has degree of precision  $2n - 1$ , it yields the exact value for  $\int_a^b f(x) dx$  when  $f(x)$  is any polynomial of degree  $2n - 1$  or less. In particular, then, it is exact for

$$q_j(x) \equiv \frac{p_n(x)}{(x - x_j)^2}, \quad j = 1, 2, \dots, n,$$

which are polynomials of degree  $2n - 2$ ; i.e.,

$$\int_a^b q_j(x) dx = \sum_{k=1}^n \alpha_k q_j(x_k).$$

However, it is clear that

$$q_j(x_k) = 0, \quad \text{for } k \neq j;$$

$$q_j(x_j) = \prod_{\substack{i=1 \\ (i \neq j)}}^n (x_j - x_i)^2 = [p_n'(x_j)]^2 > 0.$$

Thus we find

$$\alpha_j = \frac{1}{q_j(x_j)} \int_a^b q_j(x) dx = \frac{1}{[p_n'(x_j)]^2} \int_a^b \frac{p_n^2(x)}{(x - x_j)^2} dx > 0. \quad \blacksquare$$

Note that the only property of Gaussian quadrature used in this proof is the fact that the formula with  $n$  nodes has degree of precision at least  $2n - 2$ . Thus we may also conclude that any quadrature formula of the form in (1) using  $n$  nodes and having degree of precision  $2n - 2$  has positive coefficients. Another formula for computing the coefficients  $\alpha_j$  is derived in the next section [see equations (4.5)].

We can obtain expressions for the error in Gaussian quadrature which are more useful than that given in (2). The first such result can be stated as

**THEOREM 3.** *Let  $f'(x)$  be continuous in the closed interval  $[a, b]$ . Let  $\xi_1, \xi_2, \dots, \xi_n$  be any  $n$  distinct points in  $[a, b]$  which do not coincide with the zeros,  $x_1, x_2, \dots, x_n$ , of the  $n$ th orthogonal polynomial,  $p_n(x)$ , over  $[a, b]$ . Then the error in  $n$  point Gaussian quadrature applied to  $\int_a^b f(x) dx$  is*

$$(6) \quad E_n\{f\} = \int_a^b p_n(x)(x - \xi_1) \cdots (x - \xi_n) f[x_1, \dots, x_n, \xi_1, \dots, \xi_n, x] dx.$$

*Proof.* Using the  $2n$  distinct points  $x_j$  and  $\xi_j$ , the function  $f(x)$  can be written as

$$f(x) = P_{2n-1}(x) + R_{2n-1}(x),$$

where  $P_{2n-1}(x)$  is the interpolation polynomial of degree at most  $2n - 1$  agreeing with  $f(x)$  at the  $2n$  points  $x_j$  and  $\xi_j$ , and  $R_{2n-1}(x)$  is the interpolation error. With Newton's form for the remainder we write this error as

$$R_{2n-1}(x) = \prod_{j=1}^n [(x - x_j)(x - \xi_j)] f[x_1, \dots, x_n, \xi_1, \dots, \xi_n, x].$$

These expressions in (1) yield

$$(7) \quad \int_a^b P_{2n-1}(x) dx + \int_a^b R_{2n-1}(x) dx \\ = \sum_{j=1}^n \alpha_j P_{2n-1}(x_j) + \sum_{j=1}^n \alpha_j R_{2n-1}(x_j) + E_n\{f\}.$$

However, since the degree of precision of the quadrature formula is  $2n - 1$  we must have

$$\int_a^b P_{2n-1}(x) dx = \sum_{j=1}^n \alpha_j P_{2n-1}(x_j).$$

Also, since  $f'(x)$  is continuous, it follows that  $f[x_1, \dots, x_n, \xi_1, \dots, \xi_n, x_j]$  has a finite value for  $j = 1, 2, \dots, n$  and so  $R_{2n-1}(x_j) = 0$ . By using these results in (7), we obtain (6). ■

We note that there is great freedom in the choice of the  $n$  points  $\xi_j$  in Theorem 3. Further, the conditions on  $f(x)$  could be relaxed somewhat to require only continuity of  $f(x)$  on  $[a, b]$  and differentiability at the points  $x_j$  and  $\xi_j$ , and the error representation (6) remains valid (see Problems 1.4 through 1.6 of Chapter 6). By requiring more differentiability of  $f(x)$ , the result in Theorem 3 can be simplified. The most common such simplification is stated as the

**COROLLARY.** *Let  $f(x)$  have a continuous derivative of order  $2n$  in  $[a, b]$ . Then the error in  $n$ -point Gaussian quadrature applied to  $\int_a^b f(x) dx$  is*

$$(8) \quad E_n\{f\} = \frac{f^{(2n)}(\xi)}{(2n)!} \int_a^b p_n^2(x) dx,$$

where  $\xi$  is some point in  $(a, b)$ .

*Proof.* Under the assumed continuity conditions on  $f(x)$  the integrand in (6) is a continuous function of the  $n$  points  $\xi_1, \xi_2, \dots, \xi_n$ . Thus it is legitimate in this integral to let  $\xi_j \rightarrow x_j$  for  $j = 1, 2, \dots, n$ , and obtain, by applying the mean value theorem for integrals,

$$E_n\{f\} = \int_a^b p_n^2(x) f[x_1, \dots, x_n, x_1, \dots, x_n, x] dx \\ = f[x_1, \dots, x_n, x_1, \dots, x_n, \eta] \int_a^b p_n^2(x) dx, \quad a < \eta < b.$$

The result (8) now follows from the extension of Corollary 2 of Theorem 1.2 in Chapter 6. ■

It should be recalled in all of these results that  $p_n(x)$  is *not the normalized orthogonal polynomial of degree  $n$*  over  $[a, b]$ , but, by (3), is the one with leading coefficient unity. So if the  $n$ th degree *orthonormal* polynomial is

$$Q_n(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0,$$

then, since  $p_n(x)$  and  $Q_n(x)$  have the same zeros,

$$p_n(x) = \frac{1}{a_n} Q_n(x).$$

Thus we deduce that

$$(9) \quad \int_a^b p_n^2(x) dx = \frac{1}{a_n^2} \int_a^b Q_n^2(x) dx = \frac{1}{a_n^2}.$$

For example if  $a = -1$  and  $b = 1$  then the Legendre polynomials,  $P_n(x)$ , are the relevant orthogonal polynomials. It can be shown that

$$P_n(x) = \frac{1}{2^n n!} \frac{d^n}{dx^n} (x^2 - 1)^n$$

and they are normalized by forming  $\sqrt{\frac{2n+1}{2}} P_n(x)$ . Thus we find in this case that

$$a_n = \frac{(2n)!}{2^n (n!)^2} \sqrt{\frac{2n+1}{2}},$$

and the error expression (8) becomes for Gaussian quadrature over  $[-1, 1]$ :

$$(10) \quad E_n\{f\} = \frac{2}{(2n+1)!} \left[ \frac{2^n (n!)^2}{(2n)!} \right]^2 f^{(2n)}(\xi), \quad -1 < \xi < 1.$$

#### 4. WEIGHTED QUADRATURE FORMULAE

It is of practical and theoretical interest to consider the approximate evaluation, for a fixed weight function  $w(x)$ , of integrals of the form

$$(1) \quad W\{g\} = \int_a^b g(x) w(x) dx,$$

by quadrature formulae of the form

$$(2) \quad W_n\{g\} = \sum_{j=1}^n \beta_j g(x_j).$$

We again call the points  $x_j$  the nodes and the  $\beta_j$  the coefficients of the formula. However, only the factor  $g(x)$  in the integrand enters directly into the

evaluation of the integration formula (2). The weight factor,  $w(x)$ , enters into the determination of the coefficients and nodes. Once these are determined the formula may be applied to integrals of the form (1) with different functions  $g(x)$  but the same weight function. Formulae of the form (2), when applied to approximate integrals of the form (1), are called *weighted quadrature formulae*. To evaluate integrals of the form (0.1) by such formulae, write

$$\int_a^b f(x) dx = \int_a^b \frac{f(x)}{w(x)} w(x) dx,$$

and use (2) with  $g(x) \equiv f(x)/w(x)$ . As we shall see, there are frequently advantages to such procedures.

Many of the previous results are valid, with but slight changes, for the weighted quadrature formulae. Their degree of precision is defined as before; i.e., (2) has degree of precision  $m$  if

$$W_n\{x^k\} = W\{x^k\}, \quad \text{for } k = 0, 1, \dots, m,$$

but not for  $k = m + 1$ .

Given  $n + 1$  distinct points  $x_0, x_1, \dots, x_n$  the *weighted interpolatory formula* with these points as nodes and weight function  $w(x)$  over  $[a, b]$  is, say,

$$(3a) \quad W_{n+1}\{g\} = \sum_{j=0}^n w_{n,j} g(x_j);$$

where

$$(3b) \quad w_{n,j} = \int_a^b \phi_{n,j}(x) w(x) dx, \quad j = 0, 1, \dots, n.$$

Here the  $\phi_{n,j}(x)$  are the Lagrange interpolation coefficients for the points  $x_0, x_1, \dots, x_n$ . Since  $g(x) = P_n(x) + \omega_n(x)g[x_0, \dots, x_n, x]$  where  $P_n(x)$  is the Lagrange interpolation polynomial of degree  $n$ , the error in (3a and b) is

$$(3c) \quad \begin{aligned} E_{n+1}\{g\} &= W\{g\} - W_{n+1}\{g\} \\ &= \int_a^b [g(x) - P_n(x)] w(x) dx \\ &= \int_a^b \omega_n(x) g[x_0, \dots, x_n, x] w(x) dx. \end{aligned}$$

By assuming sufficient differentiability of  $g(x)$  we can simplify this error

expression. Also if  $\omega_n(x)w(x)$  does not change sign in  $[a, b]$ , even further simplification is possible. The case of equally spaced nodes does not yield particularly simple error expressions for arbitrary weight functions,  $w(x)$ , and so we do not study these formulae further. It should be observed, however, that a modification of the method of undetermined coefficients still applies (i.e., the right-hand sides in (1.16) are changed from  $\int_a^b x^v dx$  to  $\int_a^b x^v w(x) dx$ ). Hence, *the result of Theorem 1.3 is valid* when modified to refer to weighted formulae.

If the weight function,  $w(x)$ , is positive in  $[a, b]$  and say for simplicity continuous, then the Gaussian quadrature formulae also generalize in an obvious manner. These generalizations are best derived by seeking the  $n$  nodes,  $x_j$ , and coefficients,  $\alpha_j$ , such that the weighted formula (2) will have the maximum degree of precision. We find now with  $q_n(x) \equiv (x - x_1) \cdots (x - x_n)$  that

**THEOREM 1.** *The weighted quadrature formula (2) has degree of precision at most  $2n - 1$ . This maximum degree of precision is attained iff the  $n$  nodes,  $x_j$ , are the zeros of  $q_n(x)$ , the  $n$ th orthogonal polynomial with respect to the weight  $w(x)$  over  $[a, b]$ . The formula is a weighted interpolatory one.*

*Proof.* The details of the proof are left as an exercise to the reader. They follow closely the proof of Theorem 3.1. ■

The coefficients of the weighted formula of maximum precision are given by

$$(4) \quad \beta_j = \frac{1}{q_n'(x_j)} \int_a^b \frac{q_n(x)}{x - x_j} w(x) dx, \quad j = 1, 2, \dots, n.$$

Exactly as in Theorem 3.2 it follows that  $\beta_j > 0$ . The coefficients,  $\beta_j$ , are called the Christoffel numbers. The formulae (2) are of the type frequently called *weighted Gaussian quadrature* with special names applied for special weight functions (see Subsection 4.1).

The coefficients  $\beta_j$  of the weighted Gaussian formulae can be expressed in a simpler form than that given by (4). For this purpose let  $P_n(x)$  denote the  $n$ th orthonormal polynomial over  $[a, b]$  with respect to the given weight function,  $w(x)$ . If the leading coefficient of  $P_n(x)$  is  $a_n$ , we have  $P_n(x) = a_n q_n(x)$  and hence from (4)

$$\beta_j = \frac{1}{P_n'(x_j)} \int_a^b \frac{P_n(x)}{x - x_j} w(x) dx.$$

Now set  $\xi = x_k$  in the Christoffel-Darboux relation (3.25) of Chapter 5,

multiply the result by  $w(x)/(x - x_k)$  and integrate over  $a \leq x \leq b$ . This yields, since  $P_n(x_k) = 0$  and

$$P_0(x) = \left[ \int_a^b w(x) dx \right]^{-\frac{1}{2}};$$

$$1 = \frac{-a_n}{a_{n+1}} \int_a^b \frac{P_n(x)}{(x - x_k)} w(x) dx P_{n+1}(x_k), \quad k = 1, 2, \dots, n.$$

Use this in the previous expression for  $\beta_j$  to obtain

$$(5a) \quad \beta_j = \frac{-a_{n+1}}{a_n P_n'(x_j) P_{n+1}(x_j)}.$$

From the three term recursion of Theorem 3.5 in Chapter 5 we find that, since  $x_j$  is a zero of  $P_n(x)$ ,

$$P_{n+1}(x_j) = -\frac{a_{n+1} a_{n-1}}{a_n^2} P_{n-1}(x_j),$$

and (5a) becomes

$$(5b) \quad \beta_j = \frac{a_n}{a_{n-1} P_n'(x_j) P_{n-1}(x_j)}.$$

It should be observed that the coefficients,  $\alpha_j$ , of the ordinary Gaussian quadrature formulae are also given by the above formulae, (5), in which the  $P_n(x)$  are orthonormal with respect to the uniform weight,  $w(x) \equiv 1$ .

The errors of the weighted Gaussian formulae are derived exactly as in Theorem 3.3 and its corollary. Thus under appropriate continuity conditions on  $g(x)$  we have

$$(6) \quad E_n\{g\} = W\{g\} - W_n\{g\}$$

$$= \int_a^b q_n(x)(x - \xi_1) \cdots (x - \xi_n)$$

$$\times g[x_1, \dots, x_n, \xi_1, \dots, \xi_n, x] w(x) dx$$

$$= \int_a^b q_n^2(x) g[x_1, \dots, x_n, x_1, \dots, x_n, x] w(x) dx$$

$$= g[x_1, \dots, x_n, x_1, \dots, x_n, \eta] \int_a^b q_n^2(x) w(x) dx$$

$$= \frac{g^{(2n)}(\xi)}{(2n)!} \int_a^b q_n^2(x) w(x) dx, \quad a < \xi < b.$$

#### 4.1. Gauss-Chebyshev Quadrature

The polynomials orthogonal over  $[-1, 1]$  with respect to the weight  $w(x) = (1 - x)^{-p}(1 + x)^{-q}$ , provided  $p < 1$  and  $q < 1$ , are known as the

Jacobi polynomials. The special case,  $p = q = \frac{1}{2}$ , arises in the treatment of integrals of the form

$$(7) \quad W\{g\} \equiv \int_{-1}^1 \frac{g(x)}{\sqrt{1-x^2}} dx.$$

That is, consider the orthonormal polynomials over  $[-1, 1]$  with respect to the weight function  $(1-x^2)^{-\frac{1}{2}}$ , say  $Q_n(x)$ , such that

$$\int_{-1}^1 Q_n(x) Q_m(x) \frac{dx}{\sqrt{1-x^2}} = \delta_{n,m}.$$

Introduce the change of variable

$$x = \cos \theta, \quad 0 \leq \theta \leq \pi,$$

and these integrals reduce to the form

$$\int_0^\pi Q_n(\cos \theta) Q_m(\cos \theta) d\theta = \delta_{n,m}.$$

In Problem 3.9 and equation (4.13) of Chapter 5 we verify that the polynomials are

$$(8) \quad \begin{aligned} Q_n(x) &\equiv \sqrt{\frac{2}{\pi}} \cos(n \cos^{-1} x), \quad n = 1, 2, \dots, \\ Q_0(x) &\equiv \frac{1}{\sqrt{\pi}}. \end{aligned}$$

The nodes for the  $n$ -point quadrature formula of maximum degree of precision are, by Theorem 1, the points  $x_j$ , such that

$$Q_n(x_j) = 0, \quad -1 \leq x_j \leq 1.$$

But from (8) the zeros are

$$(9) \quad x_j \equiv \cos \theta_j = \cos\left(\frac{2j-1}{2n}\pi\right), \quad j = 1, 2, \dots, n.$$

The Christoffel numbers,  $\beta_j$ , for this best formula are most easily evaluated by using (5). That is, from (4.13) of Chapter 5 and (8)

$$Q_n(x) = 2^{n-1} \sqrt{\frac{2}{\pi}} x^n + \dots, \quad \text{for } n = 1, 2, \dots$$

Hence

$$a_n = \frac{2^n}{\sqrt{2\pi}}, \quad \text{for } n = 1, 2, \dots,$$

$$a_0 = \frac{1}{\sqrt{\pi}}.$$

But by using (5b)

$$\beta_j = \frac{2}{Q_n'(x_j)Q_{n-1}(x_j)}, \quad \text{for } n = 2, 3, \dots;$$

therefore, from (9) with  $\frac{d}{dx} = \frac{d\theta}{dx} \frac{d}{d\theta}$ ,

$$\beta_j = \frac{\pi}{n} \frac{\sin \theta_j}{\sin n\theta_j \cos (n-1)\theta_j}.$$

Now  $\cos(n-1)\theta_j = \sin \theta_j \sin n\theta_j$ , whence

$$(10) \quad \beta_j = \frac{\pi}{n}, \quad j = 1, 2, \dots, n.$$

The quadrature formulae thus derived are

$$(11) \quad W_n\{g\} = \frac{\pi}{n} \sum_{j=1}^n g(x_j), \quad n = 1, 2, \dots$$

It is of great interest to note that *each such formula has uniform coefficients* and that the  $n$ -point Gauss-Chebyshev formula (11) has degree of precision  $2n - 1$ . Thus the problem posed in Section 2, of choosing coefficients  $\beta_j$ , to minimize the mean-square roundoff error in evaluating the sum (2), is solved by the same coefficients that yield maximum precision in approximating integrals of the form (7).

## PROBLEM, SECTION 4

1. Carry out the proof of Theorem 1.

## 5. COMPOSITE QUADRATURE FORMULAE

By Theorem 1.3 (and its generalization for weighted quadrature) we see that all of the formulae considered thus far have been interpolatory. Thus, in effect, the integrand has been approximated by a single polynomial over the entire interval of integration and the integral of this polynomial is the approximation to the integral. (This is the justification of the name *simple quadrature formula*.) In order to get reasonable accuracy over a large integration interval, low degree polynomial approximations would in general not suffice. We learned in Subsection 3.4 of Chapter 6 that a high order interpolation polynomial may be a poor approximation to a smooth function in the case that the nodes are uniformly spaced. Hence we avoid

using integration formulae based on interpolation polynomials of high degree and uniformly spaced nodes. On the other hand, the coefficients and nodes for formulae with maximum degree of precision are not available for large orders and are difficult to compute with great accuracy. Nevertheless, it is possible to devise quadrature formulae which have simple coefficients and nodes, and yet yield accurate approximations. These are the so-called *composite rules* which, in brief, are devised by dividing the integration interval into subintervals (usually of equal length) and then applying some formula of relatively low degree of precision over each of the subintervals. There are many composite quadrature formulae, and we only examine those most commonly used.

Let the integral to be approximated be

$$(1) \quad \int_a^b f(x) dx.$$

Given integers  $m$  and  $n$ , define

$$(2a) \quad H \equiv \frac{b - a}{m}, \quad h \equiv \frac{H}{n}.$$

and divide  $[a, b]$  into  $m$  subintervals, each of length  $H$ , by the points

$$(2b) \quad y_j = a + jH, \quad j = 0, 1, \dots, m.$$

Each of these subintervals is divided into finer subintervals of length  $h$  by the points

$$(2c) \quad x_k = a + kh, \quad k = 0, 1, \dots, mn.$$

Now (1) may be written as

$$(3) \quad \int_a^b f(x) dx = \sum_{j=1}^m \int_{y_{j-1}}^{y_j} f(x) dx.$$

By using the appropriate points  $x_k$  of (2c) each of the  $m$  integrals on the right-hand side of (3) can be approximated by a closed Newton-Cotes formula with  $n + 1$  nodes. That is, by adapting the notation of Section 1,

$$(4) \quad \int_{y_{j-1}}^{y_j} f(x) dx = \sum_{k=0}^n w_{n,k}^{(j)} f(y_{j-1} + kh) + E_{n+1}^{(j)}\{f\},$$

$$j = 1, 2, \dots, m;$$

where  $E_{n+1}^{(j)}\{f\}$  is the error in the  $(n + 1)$ -point formula applied to the  $j$ th integral and  $w_{n,k}^{(j)}$  is the  $k$ th coefficient for the  $j$ th integral.

The coefficients,  $w_{n,k}^{(j)}$ , are independent of  $j$ . In fact, from (1.15), we may write

$$(5) \quad w_{n,k}^{(j)} = h A_{n,k}(0, n) \equiv h A_{n,k},$$

where  $A_{n,k}$  depends only upon the integers  $n$  and  $k$ . That is, corresponding coefficients in each subinterval are equal. With the use of (4) and (5), equation (3) becomes

$$(6) \quad \int_a^b f(x) dx = h \sum_{j=1}^m \sum_{k=0}^n A_{n,k} f(y_{j-1} + kh) + \sum_{j=1}^m E_{n+1}^{(j)}\{f\}$$

$$= h \sum_{k=0}^n A_{n,k} \left[ \sum_{j=1}^m f(y_{j-1} + kh) \right] + \sum_{j=1}^m E_{n+1}^{(j)}\{f\}.$$

However, since  $y_j = y_{j-1} + nh$  it is seen that the values of the integrand at the points  $y_j$  with  $j = 1, 2, \dots, m - 1$  appear twice in (6). We account for this repetition and rewrite the sum in the form

$$(7) \quad \int_a^b f(x) dx = h \left\{ A_{n,0} f(y_0) + A_{n,n} f(y_m) + (A_{n,0} + A_{n,n}) \sum_{j=1}^{m-1} f(y_j) \right.$$

$$\left. + \sum_{k=1}^{n-1} A_{n,k} \left[ \sum_{j=1}^m f(y_{j-1} + kh) \right] \right\} + E_{m,n+1}\{f\}.$$

Here we have introduced the *composite error*

$$(8) \quad E_{m,n+1}\{f\} \equiv \sum_{j=1}^m E_{n+1}^{(j)}\{f\}.$$

Since the same closed Newton-Cotes formula has been employed over each interval  $[y_{j-1}, y_j]$ , we deduce from (1.11) and Lemma 1.1 applied to (8), that

$$(9a) \quad \begin{aligned} E_{m,n+1}\{f\} &= \frac{mM_n}{(n+2)!} h^{n+3} f^{(n+2)}(\xi), \\ M_n &\equiv \int_0^n t \pi_n(t) dt < 0, \quad n \text{ even}; \\ E_{m,n+1}\{f\} &= \frac{mM_n}{(n+1)!} h^{n+2} f^{(n+1)}(\xi), \\ M_n &\equiv \int_0^n \pi_n(t) dt < 0, \quad n \text{ odd}. \end{aligned}$$

Here  $a < \xi < b$  and we have assumed that the indicated derivative of  $f(x)$  is continuous on  $[a, b]$ . By (2a) this error can be written as

$$(9b) \quad E_{m,n+1}\{f\} = \begin{cases} \frac{b-a}{(n+2)!} \frac{M_n}{n^{n+3}} H^{n+2} f^{(n+2)}(\xi), & n \text{ even}; \\ \frac{b-a}{(n+1)!} \frac{M_n}{n^{n+2}} H^{n+1} f^{(n+1)}(\xi), & n \text{ odd}. \end{cases}$$

Thus for fixed  $n$ , the error can be made arbitrarily small by letting  $H \rightarrow 0$ . In this manner, we find that composite quadrature formulae may be very accurate when applied to integrals whose integrands do not possess high order derivatives.

The most common composite formulae are those with  $n = 1$  (trapezoidal rule) and  $n = 2$  (Simpson's rule). For the trapezoidal rule we have

$$h = H = \frac{b - a}{m}, \quad A_{1,0} = A_{1,1} = \frac{1}{2}, \quad M_1 = -\frac{1}{6};$$

and (7) and (9) yield

$$(10) \quad \int_a^b f(x) dx = \frac{h}{2} \left\{ f(a) + f(b) + 2 \sum_{j=1}^{m-1} f(a + jh) \right\} - \frac{(b - a)}{12} h^2 f^{(2)}(\xi).$$

For Simpson's rule, with  $n = 2$  in (5) and (9),

$$h = \frac{H}{2} = \frac{b - a}{2m}, \quad A_{2,0} = A_{2,2} = \frac{1}{3}, \quad A_{2,1} = \frac{4}{3}, \quad M_2 = -\frac{4}{15};$$

so that (7) becomes

$$(11) \quad \int_a^b f(x) dx = \frac{h}{3} \left\{ f(a) + f(b) + 2 \sum_{j=1}^{m-1} f(a + 2jh) + 4 \sum_{j=1}^m f(a + [2j - 1]h) \right\} - \frac{(b - a)}{180} h^4 f^{(4)}(\xi).$$

We note that in formulae (10) and (11) the coefficients are all positive and so the roundoff errors are not generally magnified. In fact, the Newton-Cotes closed formulae have positive coefficients for  $n \leq 8$ .

In practice, the nodal tabulation of  $f(x)$  in  $[a, b]$  may not permit the use of the composite Simpson's rule because the number of net points,  $N + 1$ , is even. That is, the uniformly spaced points are  $x_j = x_0 + jh$ , such that  $x_0 = a$ ,  $x_N = b$ .

In this case we could use the closed formula

$$\int_a^{a+3h} f(x) dx = \frac{3h}{8} [f(a) + 3f(a + h) + 3f(a + 2h) + f(a + 3h)] + E_4,$$

with

$$E_4 = -\frac{3h^5}{80} f^{(4)}(\xi), \quad a < \xi < a + 3h.$$

The remaining integral

$$\int_{a+3h}^b f(x) dx$$

can then be evaluated by the composite Simpson's rule.

The error term  $E_4$  is comparable to the error term

$$E_3 = -(h^5/90)f^{(4)}(\eta).$$

This illustrates the general principle of forming composite rules with simple formulae of comparable accuracy.

### 5.1. Periodic Functions and the Trapezoidal Rule

Experience has shown that if  $f(x)$  is periodic, i.e.,  $f(x + L) = f(x)$ , the formula for the integral over a period,

$$(12) \quad \int_0^L f(x) dx \cong \frac{L}{N} \sum_{j=0}^{N-1} f(x_j), \quad x_j = jh, \quad h = \frac{L}{N},$$

is remarkably accurate. One possible explanation is that (12) arises from the composition of formulae having an arbitrarily high degree of precision. That is, from the Euler-Maclaurin summation formula (4.22) of Chapter 6 (also see Problem 4.7 of Chapter 6) for  $p = 1$ ,

$$\begin{aligned} \int_{x_0}^{x_0+h} f(x) dx &= hf(x_0) + \frac{h}{2} [f(x_1) - f(x_0)] \\ &\quad - \frac{h^2}{12} [f'(x_1) - f'(x_0)] \\ &\quad + \frac{h^4}{720} [f^{(3)}(x_1) - f^{(3)}(x_0)] \\ &\quad - \dots. \end{aligned}$$

If the above is composed for all of the intervals  $(x_j, x_{j+1})$  where  $0 \leq j \leq N - 1$ , we find that the terms in brackets cancel in the interior for any function  $f(x)$ , but also cancel at  $x_0$  and  $x_N$  since  $f(x)$  is periodic.

We have, in fact,

**THEOREM 1.** *If  $f(x) \in C^{2m+2}[0, L]$  and  $f(x)$  is  $L$ -periodic, then the composite trapezoidal rule, (12), has the error*

$$e_N \equiv \int_0^L f(x) dx - \frac{L}{N} [\frac{1}{2}f(x_0) + f(x_1) + \dots + f(x_{N-1}) + \frac{1}{2}f(x_N)]$$

where

$$e_N = h^{2m+2} LC_m f^{(2m+2)}(\xi),$$

with some  $\xi$  such that  $0 \leq \xi \leq L$ , and with a constant  $C_m$  that is independent of  $N$  and  $f(x)$ .

*Proof.* Note the central difference integration formula that is derived in Problem 1.5,

$$\begin{aligned} \int_{x_j}^{x_{j+1}} f(x) dx &= \frac{h}{2}(f_j + f_{j+1}) + hr_1(\delta^2 f_j + \delta^2 f_{j+1}) + \dots \\ &\quad + hr_m(\delta^{2m} f_j + \delta^{2m} f_{j+1}) + C_m h^{2m+3} f^{(2m+2)}(\xi_j). \end{aligned}$$

Add the formulae for each interval  $(x_j, x_{j+1})$ ,  $0 \leq j \leq N - 1$ . The difference terms contribute nothing to this sum. That is, with the notation

$$\psi_j \equiv \frac{1}{2}f_j + f_{j+1} + \dots + f_{j+N-1} + \frac{1}{2}f_{j+N},$$

the integral becomes

$$\begin{aligned} \int_{x_0}^{x_N} f(x) dx &= h[\psi_0 + 2r_1\delta^2\psi_0 + 2r_2\delta^4\psi_0 + \dots + 2r_m\delta^{2m}\psi_0] \\ &\quad + h^{2m+3}C_m \sum_{j=0}^{N-1} f^{(2m+2)}(\xi_j). \end{aligned}$$

But it is easy to see from the periodicity of  $f(x)$  that

$$\psi_j = \sum_{s=0}^{N-1} f_{j+s} = \psi_p$$

for all integers  $p$ . Hence, in particular,

$$\delta^{2k}\psi_0 = 0 \quad \text{for all integers } k \geq 1.$$

Therefore, by Lemma 1.1 and the definition of  $h$  in (12),

$$\begin{aligned} (13) \quad \int_{x_0}^{x_N} f(x) dx &= h[\frac{1}{2}f_0 + f_1 + f_2 + \dots + f_{N-1} + \frac{1}{2}f_N] \\ &\quad + h^{2m+3}C_m \sum_{j=0}^{N-1} f^{(2m+2)}(\xi_j) \\ &= h[\frac{1}{2}f_0 + f_1 + f_2 + \dots + f_{N-1} + \frac{1}{2}f_N] \\ &\quad + h^{2m+2}LC_m f^{(2m+2)}(\xi), \quad 0 \leq \xi \leq L. \quad \blacksquare \end{aligned}$$

## 5.2. Convergence for Continuous Functions

In the event that the function  $f(x)$  is merely continuous (or piecewise continuous, with jump discontinuities), we can still prove convergence of composite quadrature formulae that have non-negative coefficients and degree of precision  $s > 0$ , as in

**THEOREM 2.** *With the notation of 2 (a, b, and c), let*

$$(14) \quad S_{m,n}\{f\} \equiv \sum_{j=1}^m S_{m,n}^{(j)}\{f\},$$

where

$$(15) \quad S_{m,n}^{(j)}\{f\} \equiv \sum_{k=0}^n \alpha_{n,k}^{(j)} f(y_{j-1} + kh).$$

If  $f(x)$  is continuous in  $[a, b]$ ,  $\alpha_{n,k}^{(j)} \geq 0$  and if  $S_{m,n}^{(j)}$  has degree of precision  $s$  (i.e.,

$$S_{m,n}^{(j)}\{g\} = \int_{y_{j-1}}^{y_j} g(\xi) d\xi$$

for  $g(\xi) \equiv \xi^p$ ,  $0 \leq p \leq s$ ), then

$$\lim_{m \rightarrow \infty} S_{m,n}\{f\} = \int_a^b f(x) dx.$$

*Proof.* As  $m$  tends to infinity the closed intervals  $[y_{j-1}, y_j]$  become arbitrarily small. Hence given any  $\epsilon > 0$  there is an  $M$  such that if  $m \geq M$  there exist polynomials  $\{P_s^{(j)}(x)\}$  for  $j = 1, 2, \dots, m$ , of degree at most  $s$ , such that

$$\max_{y_{j-1} \leq x \leq y_j} |f(x) - P_s^{(j)}(x)| \leq \epsilon \quad \text{for } j = 1, 2, \dots, m.$$

Now

$$(16) \quad \begin{aligned} \left| \int_{y_{j-1}}^{y_j} f(x) dx - S_{m,n}^{(j)}\{f\} \right| &\leq \left| \int_{y_{j-1}}^{y_j} f(x) dx - \int_{y_{j-1}}^{y_j} P_s^{(j)}(x) dx \right| \\ &\quad + \left| \int_{y_{j-1}}^{y_j} P_s^{(j)}(x) dx - S_{m,n}^{(j)}\{f\} \right| \\ &\leq \epsilon |y_j - y_{j-1}| + |S_{m,n}^{(j)}\{P_s^{(j)} - f\}| \end{aligned}$$

The fact that  $S_{m,n}^{(j)}$  has degree of precision  $s$  was used to obtain the last term on the right-hand side.

The fact that  $S_{m,n}^{(j)}\{g\}$  is exact for  $g(\xi)$  identically constant and that  $\alpha_{n,k}^{(j)} \geq 0$ , implies that

$$\sum_{k=0}^n |\alpha_{n,k}^{(j)}| = |y_j - y_{j-1}|.$$

Hence (16) yields

$$\left| \int_{y_{j-1}}^{y_j} f(x) dx - S_{m,n}^{(j)}\{f\} \right| \leq 2\epsilon |y_j - y_{j-1}|.$$

Therefore,

$$\left| \int_a^b f(x) dx - S_{m,n}\{f\} \right| \leq 2\epsilon |b - a|. \quad \blacksquare$$

By picking  $P_s^{(j)}(x) \equiv f(\xi_j)$ , where  $\xi_j \equiv (y_{j-1} + y_j)/2$ , we find the

COROLLARY. Under the hypothesis of Theorem 2,

$$\left| \int_a^b f(x) dx - S_{m,n}\{f\} \right| \leq 2|b-a|\omega(f; \delta),$$

where  $\omega$  is the modulus of continuity and  $\delta \equiv \max_j |y_{j-1} - y_j|/2$ . ■

We leave to Problems 8 and 9 the formulation of generalizations of Theorem 2.

## PROBLEMS, SECTION 5

1. At what interval in  $x$  and to how many decimal places must  $f(x)$  be tabulated in order to evaluate  $\int_0^5 f(x) dx$  correctly to six decimal places by using:

- (a) composite trapezoidal rule,
- (b) composite Simpson's rule, for  $f(x) \equiv \cos x$ ?

2. The composite midpoint rule is based on the single node, open Newton-Cotes formula with error,  $E_1$ :

$$\int_a^{a+2h} f(x) dx \equiv 2hf(a+h) + E_1.$$

Find the expression for the composite formula and error term when the integral to be evaluated is

$$\int_a^{a+2mh} f(x) dx, \quad m = \frac{b-a}{2h}.$$

3. Answer Problem 1 for the composite midpoint rule (see Problem 2).

4\*. Use the notation of Problem 1.5 to derive the *composite trapezoidal rule with end corrections*:

$$\begin{aligned} \int_{x_0}^{x_N} f(x) dx &= h \left[ (\frac{1}{2}f_0 + f_1 + \cdots + f_{N-1} + \frac{1}{2}f_N) \right. \\ &\quad \left. + \sum_{k=1}^m r_k(d_{N,k} - d_{0,k}) \right] + C_m(x_N - x_0)h^{2m+2}f^{(2m+2)}(\xi), \end{aligned}$$

where

$$d_{j,k} \equiv \Delta^{2k-1}f_{j-k} + \Delta^{2k-1}f_{j-k+1},$$

and

$$x_{-k} \leq \xi \leq x_{N+k}.$$

[Hint: Use equation (3.16a) of Chapter 6 to get

$$\delta^{2k}f_s = \Delta^{2k-1}f_{s-k+1} - \Delta^{2k-1}f_{s-k},$$

whence

$$\sum_{s=1}^{N-1} \delta^{2k}f_s = \Delta^{2k-1}f_{N-k} - \Delta^{2k-1}f_{1-k}.$$

Since the coefficients of  $\{f_j\}$  which occur in  $\Delta^n f_0$  are the alternating binomial coefficients  $(-1)^{n-j} \binom{n}{j}$ , we see that

$$d_{j,k} = \sum_{p=-k}^k c_p^{(k)} f_{j+p},$$

where

$$c_0^{(k)} = 0, \quad c_p^{(k)} = -c_{-p}^{(k)}, \quad p = 1, 2, \dots, k.]$$

**5.\*** Given the integer  $N > 0$ , let  $h = 1/N$ ,  $x_0 = 0$ ,  $x_j = x_0 + jh$ ;

$$t_N \equiv h(\frac{1}{2}f_0 + f_1 + \dots + f_{N-1} + \frac{1}{2}f_N).$$

(a) From Problem 4, verify that if  $f(x) \in C^{2m+2}$ ,

$$e_N \equiv \int_0^1 f(x) dx - t_N = \sum_{j=1}^m a_j h^{2j} + \mathcal{O}(h^{2m+2}).$$

[Hint: Expand the values  $f_p$  and  $f_{N+p}$ , which appear in the end corrections, by Taylor's series about  $x_0$  and  $x_N$  respectively. Collect terms with like powers of  $h$ .]

(b) Let

$$s_N \equiv h[f_{\frac{1}{2}} + f_{\frac{3}{2}} + \dots + f_{(2N-1)/2}].$$

$s_N$  is the *composite midpoint rule* for evaluating

$$\int_{x_0}^{x_N} f(x) dx$$

(with intervals  $h/2$ —see Problem 2). Verify that

$$t_{2N} = \frac{1}{2}(t_N + s_N),$$

and

$$e_{2N} \equiv \int_0^1 f(x) dx - t_{2N} = \sum_{j=1}^m a_j \left(\frac{h}{2}\right)^{2j} + \mathcal{O}(h^{2m+2}).$$

(c) Show that

$$\begin{aligned} \frac{4e_{2N} - e_N}{3} &= \int_0^1 f(x) dx - \left(\frac{4t_{2N} - t_N}{3}\right) \\ &= \sum_{j=2}^m b_j h^{2j} + \mathcal{O}(h^{2m+2}), \end{aligned}$$

and  $b_r = 0$  if  $a_r = 0$ .

Call  $t_{2N}^{(1)} \equiv (4t_{2N} - t_N)/3$  the *first extrapolation* of the trapezoidal rule.

**6.\* Romberg's method:** With the notation of Problem 5, consider the sequence of subdivisions obtained by halving, i.e.,  $N = 1, 2, 4, \dots, 2^k, \dots$ . Define

$$t_{2^k}^{(0)} \equiv t_{2^k} \quad k = 0, 1, 2, \dots;$$

$$t_{2^k}^{(1)} \equiv \frac{4t_{2^k}^{(0)} - t_{2^{k-1}}^{(0)}}{3} \quad k = 1, 2, 3, \dots;$$

$$t_{2^k}^{(2)} \equiv \frac{4^2 t_{2^k}^{(1)} - t_{2^{k-1}}^{(1)}}{4^2 - 1} \quad k = 2, 3, 4, \dots;$$

⋮

$$t_{2^k}^{(p)} \equiv \frac{4^p t_{2^k}^{(p-1)} - t_{2^{k-1}}^{(p-1)}}{4^p - 1} \quad k = p, p+1, \dots.$$

$t_{2^k}^{(p)}$  is called the *p*th extrapolation of the trapezoidal rule. Show, by induction on  $p$ , that for fixed  $p \geq 0$ ,

$$\int_0^1 f(x) dx - t_{2^k}^{(p)} = \mathcal{O}(2^{-k(2p+2)}) \quad \text{if } p \leq m.$$

Romberg's method consists in successively constructing the rows of the triangular matrix  $R$ :  $(r_{k,p})$ , with  $r_{k,p} \equiv t_{2^k}^{(p)}$ . If  $a_p \neq 0$ ,  $f(x) \in C^{2m+2}[-\delta, 1+\delta]$  for some  $\delta > 0$ , and  $0 < p \leq m$ , then, by (c), the entries in the *p*th column converge more rapidly than those of the  $(p-1)$ st column to  $\int_0^1 f(x) dx$ .

In Romberg's method we may achieve the degree of precision of the end correction formula and avoid the evaluation of  $f(x)$  outside of the interval  $[x_0, x_N]$ . In practice, the rows of  $R$  may be successively computed until the elements in some column have "converged" enough.

7.\* Prove that the *p*th extrapolation of the trapezoidal rule is a quadrature formula with non-negative weights and degree of precision at least  $2p+1$ .

That is, to approximate  $\int_0^1 f(x) dx$ ,

$$t_{2^k}^{(p)} \equiv 2^{-k} \sum_{j=0}^{2^k} c_{k,j}^{(p)} f(j2^{-k}),$$

where

$$\sum_{j=0}^{2^k} c_{k,j}^{(p)} = 2^k, \quad c_{k,j}^{(p)} \geq 0.$$

8.\* State and prove a generalization of Theorem 2, to cover the case where  $f(x)$  is piecewise continuous (with only a finite number of jump discontinuities) and where the quadrature formula is not based on uniformly spaced nodes.

9.\* Under the conditions of Theorem 2, if  $f(x)$  has a continuous derivative of order  $r$ , show that

(a) If  $r \leq s$ ,

$$\left| \int_a^b f(x) dx - S_{m,n}\{f\} \right| \leq 2|b-a| \frac{\delta^r}{r!} \omega(f^{(r)}; \delta),$$

where  $\delta = \max_j |y_{j-1} - y_j|/2$ .

[Hint: Pick

$$P_s^{(j)}(x) = f(\xi_j) + f'(x)(x - \xi_j) + \dots + \frac{f^{(r)}(\xi_j)}{r!} (x - \xi_j)^r,$$

with  $\xi_j = (y_{j-1} + y_j)/2$ . Verify that

$$|f(x) - P_s^{(j)}(x)| \leq \frac{\delta^r}{r!} \omega(f^{(r)}; \delta) \quad \text{for } y_{j-1} \leq x \leq y_j.]$$

(b) If  $r > s$ ,

$$\left| \int_a^b f(x) dx - S_{m,n}\{f\} \right| \leq 2|b-a| \frac{\delta^{s+1}}{(s+1)!} K,$$

where  $K \equiv \max_x |f^{(s+1)}(x)|$ .

## 6. SINGULAR INTEGRALS; DISCONTINUOUS INTEGRANDS

In deriving most of the quadrature formulae of this chapter and the appropriate error formulae, it was either stated or implied that the integrand and various of its higher order derivatives were continuous. (An exception is found in the weighted quadrature formulae where the weight need not be continuous.) If these conditions are violated, a quadrature method may still yield a good approximation but the error will, in general, be much larger than predicted. In less favorable circumstances, of course, the approximation will be meaningless. There are a number of cases of rather frequent occurrence in which such difficulties can be anticipated and satisfactorily resolved.

### 6.1. Finite Jump Discontinuities

If the integrand has a finite jump discontinuity at a known point (or any finite number of them), say at  $x = c$  in the interval of integration, then we write

$$(1) \quad \int_a^b f(x) dx = \int_a^c f(x) dx + \int_c^b f(x) dx.$$

Now if  $f(z)$  has sufficiently many continuous derivatives in  $[a, c]$  and  $[c, b]$  the two integrals on the right-hand side may be accurately approximated by any of a variety of quadrature formulae. This simple procedure can be considered as an application of a special composite rule, not necessarily one with equal spacing.

If the integrand is continuous but has a discontinuity in some low order derivative, a similar procedure can be employed. For example, if  $f(x) = |x|$  then  $f'(x)$  has a finite jump at  $x = 0$ . In this case, a composite rule with  $x = 0$  as an endpoint of a subinterval could be used.

### 6.2. Infinite Integrand

We consider the case in which  $f(x)$  becomes infinite as  $x \rightarrow a$ , the lower limit of integration. The upper limit can be similarly treated and an interior discontinuity, say at  $x = c$ , is reduced to the endpoint cases by using (1). We assume for the present that the integral is of the form

$$(2) \quad I \equiv \int_a^b \frac{g(x)}{(x - a)^\theta} dx, \quad 0 < \theta < 1,$$

where  $g(x)$  has continuous derivatives in  $[a, b]$  of "sufficiently high" order. The restriction on  $\theta$  insures that the integral in (2) exists for rather general functions  $g(x)$  [i.e., it is not required that  $g(a) = 0$ ].

METHOD I. For any positive number  $\epsilon$  in  $0 < \epsilon < b - a$  we write (2) as  $I = I_1 + I_2$  where

$$(3) \quad I_1 \equiv \int_a^{a+\epsilon} \frac{g(x)}{(x-a)^\theta} dx, \quad I_2 \equiv \int_{a+\epsilon}^b \frac{g(x)}{(x-a)^\theta}.$$

The range of integration in  $I_2$  is now such that the integrand has derivatives of high order there. Thus,  $I_2$  can be approximated by many of the standard procedures previously described, and in principle the error in this approximation can be estimated. It remains to approximate  $I_1$  to within a known error.

By Taylor's theorem we have for  $x \geq a$ :

$$\begin{aligned} g(x) &= g(a) + (x-a)g'(a) + \frac{(x-a)^2}{2!} g''(a) + \dots \\ &\quad + \frac{(x-a)^s}{s!} g^{(s)}(a) + \frac{(x-a)^{s+1}}{(s+1)!} g^{(s+1)}(\xi(x)). \end{aligned}$$

Use this expansion in  $I_1$  and perform the indicated integrations to find

$$(4) \quad \begin{aligned} I_1 &= \epsilon^{1-\theta} \left[ \frac{g(a)}{1-\theta} + \frac{\epsilon}{1!} \frac{g'(a)}{2-\theta} + \frac{\epsilon^2}{2!} \frac{g''(a)}{3-\theta} + \dots + \frac{\epsilon^s}{s!} \frac{g^{(s)}(a)}{s+1-\theta} \right] \\ &\quad + \frac{1}{(s+1)!} \int_a^{a+\epsilon} (x-a)^{s+1-\theta} g^{(s+1)}(\xi(x)) dx. \end{aligned}$$

If the first (bracketed) term on the right in (4) is used as the approximation to  $I_1$  we obtain the error bound

$$(5) \quad |E^{(1)}| \leq \frac{\epsilon^{s+2-\theta}}{(s+1)! (s+2-\theta)} \max_{a \leq \xi \leq a+\epsilon} |g^{(s+1)}(\xi)|.$$

For fixed  $s$  this bound is clearly an increasing function of  $\epsilon$ . Or for fixed  $\epsilon < 1$ , if the derivatives of  $g^{(s)}(x)$  do not grow too fast with  $s$  it will also be a decreasing function of  $s$ .

If the error in evaluating  $I_2$  is  $E^{(2)}$  we must now determine conditions on  $\epsilon, s$  and the quadrature formula such that  $|E^{(1)}| + |E^{(2)}| < \delta$ , where  $\delta$  is the maximum permissible error in approximating  $I$ . Of course, the parameters should be chosen such that  $|E^{(1)}| \approx |E^{(2)}|$  since then some cancellation of error may take place. For definiteness, let us assume that  $I_2$  is approximated by a composite rule using  $m$  subintervals and a closed  $(n+1)$ -point Newton-Cotes quadrature formula with equal spacing over each subinterval. Furthermore, we will assume  $n$  to be even. Then from (5.9) we must have

$$E^{(2)} = \frac{mM_n}{(n+2)!} h^{n+3} \frac{d^{n+2}}{dx^{n+2}} \frac{g(x)}{(x-a)^\theta} \Big|_{x=\xi}, \quad a+\epsilon < \xi < b;$$

where  $h = (b - a - \epsilon)/(mn)$ . From this expression we obtain the bound

$$(6) \quad |E^{(2)}| \leq \frac{mM_n}{(n+2)!} \left( \frac{b-a}{mn} \right)^{n+3} \max_{a+\epsilon \leq x \leq b} \left| \frac{d^{n+2}}{dx^{n+2}} \left[ \frac{g(x)}{(x-a)^\theta} \right] \right|,$$

in which the coefficient of the derivative term is independent of  $\epsilon$ . If the derivatives entering into (5) and (6) can be estimated, say

$$|g^{(s+1)}(\xi)| \leq M^{(s+1)},$$

and

$$\max_{a+\epsilon \leq x \leq b} \left| \frac{d^{n+2}}{dx^{n+2}} \frac{g(x)}{(x-a)^\theta} \right| = N^{(n+2)}(\epsilon),$$

then for fixed  $s$ ,  $n$ , and  $\delta$  we can find  $\epsilon$  and  $m$  such that

$$\frac{\epsilon^{s+2-\theta}}{(s+1)!(s+2-\theta)} M^{(s+1)} = mK_n \left( \frac{b-a}{mn} \right)^{n+3} N^{(n+2)}(\epsilon) = \frac{\delta}{2}.$$

The bound  $N^{(n+2)}(\epsilon)$  will, in general, become large for small  $\epsilon$  and so may imply an unusually large  $m$ . Thus, we consider an alternative procedure obtained by modifying these considerations.

**METHOD II.** Let us rewrite the Taylor expansion of  $g(x)$  as

$$(7a) \quad g(x) = G_s(x) + \frac{(x-a)^{s+1}}{(s+1)!} g^{(s+1)}(\xi(x)),$$

where

$$(7b) \quad G_s(x) \equiv g(a) + (x-a)g'(a) + \cdots + \frac{(x-a)^s}{s!} g^{(s)}(a).$$

Now the integral (2) is identically represented by

$$(8) \quad I = \int_a^b \frac{g(x) - G_s(x)}{(x-a)^\theta} dx + \int_a^b \frac{G_s(x)}{(x-a)^\theta} dx \equiv I_N + I_E.$$

The second integral,  $I_E$ , can be evaluated explicitly, just as was the first part of  $I_1$  in (4), and we have

$$(9) \quad I_E = (b-a)^{1-\theta} \left[ \frac{g(a)}{1-\theta} + \frac{(b-a)}{1!} \frac{g'(a)}{2-\theta} + \cdots + \frac{(b-a)^s}{s!} \frac{g^{(s)}(a)}{s+1-\theta} \right].$$

However, the first integral in (8),  $I_N$ , no longer has a singular integrand at  $x = a$  so it can be approximated by many of the standard quadrature formulae. In fact, the first  $s$  derivatives of

$$\frac{g(x) - G_s(x)}{(x-a)^\theta}$$

are finite at  $x = a$  and so the error in any closed formula with  $n + 2 \leq s$  is bounded. For example, in a composite formula employing Simpson's rule to approximate  $I_N$  the error becomes, from (5.9) with  $n = 2$ :

$$E^{(N)} = -\frac{m}{90} h^5 \frac{d^4}{dx^4} \left[ \frac{g(x) - G_s(x)}{(x - a)^\theta} \right]_{x=\xi}$$

where

$$h \equiv \frac{b - a}{2m}, \quad \text{and} \quad a < \xi < b.$$

In order that the indicated derivative remain bounded as  $\xi \rightarrow a$ , it is sufficient that  $s = 4$ . Of course, if  $0 \leq s < 4$  the quadrature formula will still yield an accurate approximation (see Theorem 5.1) but the above form for the error cannot be used.

**METHOD III.** As a third alternative, which is restricted to singular integrands of the form (2), we consider the change of variable

$$(x - a) = t^\phi \quad dx = \phi t^{\phi-1} dt.$$

Then (2) becomes, if  $\phi = k/(1 - \theta)$  for  $k$  any positive integer,

$$(10) \quad I = \phi \int_0^{(b-a)^{1/\phi}} g(a + t^\phi) t^{k-1} dt.$$

Now, since  $k \geq 1$ , the integrand of the above integral is continuous at  $t = 0$ . Thus numerical quadrature formulae may be directly applied to (10). In fact, if  $\theta = q/p$ , i.e., rational, and  $k = p - q$ , the integrand is smooth as long as  $g$  is smooth.

Methods I and II are applicable to other types of singularities. In fact, if the integral is of the form

$$(11) \quad \int_a^b g(x) S(x) dx$$

where  $S(a)$  is infinite, Methods I and II may be applied if integrals of the form

$$(12) \quad \int_a^{a+\epsilon} (x - a)^k S(x) dx, \quad k = 0, 1, \dots, s,$$

can be explicitly evaluated. For example, if  $S(x) \equiv \ln(x - a)$ , we can employ these methods.

**METHOD IV.** In many cases of interest the singular part of the integrand, i.e.,  $S(x)$  in (11), is of one sign throughout  $[a, b]$ . Then, in principle, the weighted quadrature methods outlined in Section 4 can be employed. In

particular, the weighted Gaussian quadrature formulae are frequently very effective for evaluating such singular integrals. Of course, such applications require the determination of the polynomials orthogonal over  $[a, b]$  with respect to the weight  $S(x)$ . For many special forms of  $S(x)$  these polynomials are known; for instance, in the case

$$(13) \quad S(x) \equiv (x - a)^{-\frac{1}{2}}(b - x)^{-\frac{1}{2}},$$

the Gauss-Chebyshev formula derived in Section 4.1 is the relevant scheme. However, even if the required polynomials are not well known, it may be of value to construct them and to devise the appropriate quadrature formula. This is especially true if many integrals containing the same singular part are to be evaluated.

### 6.3. Infinite Integration Limits

It is clear that an integral of the form

$$(14) \quad I = \int_a^\infty g(x) dx$$

cannot, in general, be accurately approximated by the standard quadrature methods (which employ a finite number of finite subintervals). The usual approach to such problems is to write again  $I = I_1 + I_2$  where now

$$(15) \quad I_1 \equiv \int_a^b g(x) dx, \quad I_2 \equiv \int_b^\infty g(x) dx.$$

Then if  $b$  is “sufficiently large,” it may be possible by analytical means to show that  $I_2$  is negligible. Or alternatively,  $g(x)$  may be approximated for  $x > b$  by some function from which  $I_2$  is then approximated; in this case good error estimates are usually difficult to obtain. Another procedure, too frequently disregarded, is to reduce  $I_2$  to an integral over a finite interval.

That is, introduce the change of variable  $x = 1/\xi$  and obtain

$$(16) \quad \begin{aligned} I_2 &= - \int_{1/b}^0 g\left(\frac{1}{\xi}\right) \xi^{-2} d\xi \\ &= \int_0^{1/b} f(\xi) d\xi. \end{aligned}$$

Here we have introduced the function

$$(17) \quad f(\xi) \equiv \frac{g(1/\xi)}{\xi^2}.$$

Now if  $f(\xi)$  is not singular at  $\xi = 0$ , then  $I_2$  may be evaluated, in the form (16), by standard quadrature methods. If  $f(\xi)$  is singular at  $\xi = 0$ , then

the evaluation of  $I_2$  might be reduced to the previous case of Subsection 6.2.

In fact, a sufficient condition for  $I_2$  defined in (15) to converge (absolutely) is that  $g(x)$  be continuous and that

$$(18) \quad \lim_{x \rightarrow \infty} x^{1+\epsilon} g(x) = 0 \quad \text{for some } \epsilon > 0.$$

This, by (17), is equivalent to

$$\lim_{\xi \rightarrow 0} \xi^{1-\epsilon} f(\xi) = 0.$$

Now this condition will be satisfied if  $f(\xi)$  behaves at  $\xi = 0$  like  $\xi^{-\theta}$  where  $\theta < 1 - \epsilon$ . If  $\epsilon < 1$ , the integral in (16) may have a singularity of the form indicated in (2).

Finally, we point out that special *weighted Gaussian quadrature formulae* may be effective for various integrals over infinite intervals. In particular, the orthogonal polynomials over  $[-\infty, \infty]$  with respect to the weight function  $e^{-x^2}$  are well known. They are called *Hermite polynomials*,  $H_n(x)$ , and they can be shown to be given by (see Problem 3.18 in Chapter 5)

$$(19a) \quad H_n(x) = (-1)^n e^{x^2} \frac{d^n}{dx^n} (e^{-x^2}).$$

It is not difficult to deduce that

$$(19b) \quad H_{n+1}(x) = 2xH_n(x) - \frac{d}{dx} H_n(x),$$

and hence by induction, since  $H_0(x) = 1$ , that

$$(19c) \quad H_n(x) = 2^n x^n + \dots$$

By repeatedly using integration by parts we can show that the normalized Hermite polynomials are

$$(19d) \quad \frac{H_n(x)}{\sqrt{2^n n! \pi^{1/2}}}.$$

The formulae based on the  $H_n(x)$  are called *Gauss-Hermite quadrature formulae* and are used to approximate integrals of the general form

$$\int_{-\infty}^{\infty} e^{-x^2} g(x) dx.$$

For integrals over  $[0, \infty]$ , the *Laguerre polynomials*,  $L_n(x)$ , defined as

$$(20a) \quad L_n(x) = (-1)^n e^x \frac{d^n}{dx^n} (x^n e^{-x}),$$

are sometimes useful. They are orthogonal over  $[0, \infty]$  with respect to the weight  $e^{-x}$ . It can be shown that

$$(20b) \quad L_n(x) = x^n + \dots,$$

and that the normalized Laguerre polynomials are  $(1/n!)L_n(x)$ . The *Gauss-Laguerre quadrature formulae* are based on these polynomials and are used to approximate integrals of the form

$$\int_0^\infty e^{-x} g(x) dx.$$

## PROBLEMS, SECTION 6

### 1. Evaluate

$$\int_0^1 \sqrt{1 - x^2} dx$$

by Method II, using the composite Simpson's rule, and obtain four-decimal-place accuracy. What is the largest interval  $h$  that is permissible?

### 2. Use Method III and the composite Simpson's rule to evaluate

$$\int_0^1 \sqrt{1 - x^2} dx$$

correctly to four decimal places. What is the largest interval  $h$  that is permissible?

### 3. Substitute the new variable $x = \cos \theta$ and use the composite Simpson's rule to evaluate

$$\int_0^1 \sqrt{1 - x^2} dx$$

correctly to four decimal places. What is the largest permissible interval  $h = \Delta\theta$ ?

### 4. Verify the properties of the Hermite and Laguerre polynomials given in the text [see equations (19) and (20)].

## 7. MULTIPLE INTEGRALS

The problem of efficiently approximating multiple integrals numerically has not been completely solved. An obvious source of complexity is the variety of domains of integration in higher dimensions compared to just intervals in our study of one dimensional integrals. However, even if the domain is restricted, say to the unit cube, then the resulting problem is still not in a satisfactory state. A fundamental difficulty is essentially the great degree of freedom in locating the nodes or equivalently in the large

number of, say, uniformly spaced nodes required to get reasonable accuracy.

One of the basic methods for approximating multiple integrals is, as in the one dimensional case, to integrate a polynomial approximation of the integrand. But since interpolation theory in higher dimensions is not well developed, we again have difficulty in devising practical schemes. Generalizations of the method of undetermined coefficients offer many possibilities, but only a few of these have been exploited for multiple integrals. Finally, we point out that the difficulties increase as the dimension of the domain of integration increases. This seems related to the fact that the ratio of the surface area to the volume, for an  $n$ -dimensional unit cube increases with  $n$ . We shall consider numerical methods for evaluating double integrals for the most part. Many of the procedures extend in an obvious way to higher dimensions, with perhaps a subsequent loss in efficiency or accuracy. Approximation methods for double integrals are frequently called *cubature formulae* since they approximate the *volume* associated with the integrand.

In general, the problem is to approximate an integral of the form

$$(1) \quad J\{f\} \equiv \int_D f(\mathbf{x}) d\mathbf{x},$$

where  $\mathbf{x} = (x_1, \dots, x_p)$  and  $d\mathbf{x} = dx_1 \cdots dx_p$  are a point and a volume element in the  $p$ -dimensional space, respectively, and  $D$  is a domain in this space. The approximations considered are all to be of the form

$$(2) \quad J_N\{f\} \equiv \sum_{v=1}^N A_v f(\mathbf{x}_v).$$

Here the  $N$  points,  $\mathbf{x}_v$ , are the nodes of the formula and the  $A_v$  are the coefficients. We say that formula (2) has degree of precision  $m$  as an approximation to the integral (1) if  $J\{P(\mathbf{x})\} = J_N\{P(\mathbf{x})\}$  for all polynomials,  $P(\mathbf{x})$ , in  $\mathbf{x}$  of degree† at most  $m$  but not for some polynomial of degree  $m + 1$ .

We cannot proceed as in Section 1 to study general interpolatory schemes since the general interpolation problem is not well posed in higher dimensions. However, if the nodes are specially chosen, say as in Section 6 of Chapter 6, then interpolation can be used and we consider such cases first.

† We say  $P(\mathbf{x})$  is of degree at most  $m$  in  $\mathbf{x}$ , if  $P(\mathbf{x})$  is a polynomial in  $(x_1, \dots, x_p)$  of the form

$$P(\mathbf{x}) \equiv \sum_{0 \leq j_1 + j_2 + \dots + j_p \leq m} C_{j_1 j_2 \dots j_p} x_1^{j_1} x_2^{j_2} \dots x_p^{j_p}.$$

### 7.1. The Use of Interpolation Polynomials

Let the integral in (1) be over a plane domain, say

$$(3) \quad J\{f\} \equiv \iint_D f(x, y) dx dy.$$

Let us pick as nodes the  $N = (m + 1)(n + 1)$  distinct points:  $(x_i, y_j)$ ,  $i = 0, 1, \dots, m$ ;  $j = 0, 1, \dots, n$ ; where the  $m + 1$  distinct numbers  $\{x_i\}$  and  $n + 1$  distinct numbers  $\{y_j\}$  are, at present, arbitrary. Then a polynomial  $P(x, y)$ , of degree  $m$  in  $x$  and  $n$  in  $y$  which is equal to  $f(x, y)$  at these  $N$  nodes is given by (6.3) in Chapter 6. We use this polynomial to define the cubature formula

$$(4a) \quad \begin{aligned} J_N\{f\} &\equiv \iint_D P(x, y) dx dy \\ &= \sum_{i=0}^m \sum_{j=0}^n A_{ij} f(x_i, y_j). \end{aligned}$$

Here we have introduced the coefficients,  $A_{ij}$ , by the definitions

$$(4b) \quad A_{ij} = \iint_D X_{m,i}(x) Y_{n,i}(y) dx dy, \quad i = 0, 1, \dots, m, j = 0, 1, \dots, n;$$

and the Lagrange interpolation coefficients  $X_{m,i}(x)$  and  $Y_{n,i}(y)$  are defined in (6.2) of Chapter 6.

While this procedure is formally valid for very general domains,  $D$ , it is only practical when the integrals in (4b) can be evaluated explicitly. A particularly simple and important special case is that of a rectangular domain,

$$D: \{x, y \mid a \leq x \leq b; c \leq y \leq d\}.$$

In this case we have

$$(5) \quad A_{ij} = \alpha_i \beta_j; \quad \alpha_i \equiv \int_a^b X_{m,i}(x) dx, \quad \beta_j \equiv \int_c^d Y_{n,i}(y) dy,$$

and the quantities  $\alpha_i$  and  $\beta_j$  are just the coefficients for appropriate one dimensional interpolatory quadrature formulae. Furthermore, if the numbers  $x_i$  are equally spaced in  $[a, b]$ , and the  $y_j$  are equally spaced in  $[c, d]$ , then the  $\alpha_i$  and  $\beta_j$  are the coefficients in the  $(m + 1)$ -point and  $(n + 1)$ -point Newton-Cotes quadrature formulae respectively (see Problem 1).

The error in the cubature formula (4) as an approximation to the integral

(3) is, upon recalling the expression (6.7c) of Chapter 6 for the error in the interpolation polynomial,

$$\begin{aligned}
 (6) \quad E_N\{f\} &\equiv \iint_D [f(x, y) - P(x, y)] dx dy \\
 &= \iint_D R(x, y) dx dy \\
 &= \iint_D \left\{ \frac{\omega_m(x)}{(m+1)!} \left( \frac{\partial}{\partial x} \right)^{m+1} f(\xi(x), y) \right. \\
 &\quad + \frac{\omega_n(y)}{(n+1)!} \left( \frac{\partial}{\partial y} \right)^{n+1} f(x, \eta(y)) \\
 &\quad - \frac{\omega_m(x)\omega_n(y)}{(m+1)!(n+1)!} \left( \frac{\partial}{\partial x} \right)^{m+1} \left( \frac{\partial}{\partial y} \right)^{n+1} \\
 &\quad \times \left. f(\xi'(x), \eta'(y)) \right\} dx dy.
 \end{aligned}$$

We deduce from this that the formula (4) has degree of precision at least  $\min(m, n)$ . For instance, if as is frequently the case, we take  $m = n$ , then by using  $N = (n+1)^2$  nodes, a formula with degree of precision at least  $n$  is obtained.

However, a formula using only  $M = \frac{1}{2}(n+1)(n+2)$  nodes can be devised which also has degree of precision at least  $n$ . For this purpose, we integrate the interpolation polynomial  $P_n(x, y)$ , given by (6.10) of Chapter 6, over  $D$ . The general result is somewhat cumbersome to write down in the form (2). First, divided differences of the type  $f[x_0, \dots, x_k; y_0, \dots, y_j]$  must be expanded as linear combinations of the function values,  $f(x_v, y_u)$ , and then all terms containing such function values must be combined to determine the coefficients,  $B_{vu}$ . For small values of  $n$ , say  $n \leq 3$ , this is easily done (see Problem 2). However, for equally spaced  $x_k$  and  $y_j$ , difference operators may be employed to simplify the notation and even the calculations. We indicate the general formula obtained in this manner as

$$\begin{aligned}
 (7) \quad K_M\{f\} &= \iint_D P_n(x, y) dx dy \\
 &= \sum_{k=0}^n \sum_{j=0}^{n-k} f[x_0, \dots, x_k; y_0, \dots, y_j] \\
 &\quad \times \iint_D \omega_{k-1}(x) \omega_{j-1}(y) dx dy \\
 &= \sum_{v=0}^n \sum_{\mu=0}^{n-v} B_{vu} f(x_v, y_\mu).
 \end{aligned}$$

$y_4 \odot$	•	•	•	•
$y_3 \odot$	◎	•	•	•
$y_2 \odot$	◎	◎	•	•
$y_1 \odot$	◎	◎	◎	•
$y_0 \odot$	◎	◎	◎	◎
$x_0$	$x_1$	$x_2$	$x_3$	$x_4$

Figure 1a

$y_1 \odot$	•	◎	◎	◎
$y_3 \odot$	•	•	•	◎
$y_2 \odot$	•	◎	•	◎
$y_4 \odot$	•	•	•	•
$y_0 \odot$	◎	◎	◎	◎
$x_0$	$x_4$	$x_2$	$x_3$	$x_1$

Figure 1b

The error in this formula is easily obtained from (6.11) of Chapter 6. We only wish to observe that it implies that (7) has degree of precision at least  $n$ .

The nodes that enter into this formula are a subset of the rectangular array of  $(n + 1)^2$  points  $(x_i, y_j)$ ,  $i, j = 0, 1, \dots, n$ . If the numbers  $x_i$  and  $y_j$  are monotonically ordered, say  $x_0 < x_1 < x_2 < \dots$ , then the nodes in (7) are those on and below the main diagonal of a schematic  $n + 1$  by  $n + 1$  matrix of dots (see Figure 1a). However, any other selection of points obtained by permuting rows or columns could also be employed. This just corresponds to a renumbering of the  $x_i$  and  $y_j$ , say as in Figure 1b.

Both of the interpolatory cubature formulae (4) and (7) are easily extended to integrals in higher dimensions. As the dimension increases, there is a greater saving in number of nodes in extensions of  $K_M\{f\}$  formulae compared to  $J_N\{f\}$  formulae while maintaining precision of degree at least  $n$ . Thus, in the plane,  $J_N\{f\}$  requires  $N = (n + 1)^2$  nodes and  $K_M\{f\}$  requires  $M = \frac{1}{2}(n + 1)(n + 2)$  nodes for each to have degree of precision at least  $n$ . The ratio of the number of nodes required is

$$\frac{M}{N} = \frac{n + 2}{2(n + 1)} \approx \frac{1}{2}, \quad \text{for large } n.$$

In three dimensions the ratio becomes

$$\frac{M}{N} = \frac{(n + 1)(n^2 + 5n + 6)/6}{(n + 1)^3} \approx \frac{1}{6}, \quad \text{for large } n.$$

## 7.2. Undetermined Coefficients (and Nodes)

The general formula (2) can be written for double integrals as

$$(8) \quad J_N\{f\} = \sum_{v=1}^N A_v f(x_v, y_v).$$

We note that there are  $3N$  parameters which determine such a scheme; the  $N$  coefficients,  $A_v$ , and the  $2N$  coordinates of the nodes,  $(x_v, y_v)$ . As an approximation to the integral (3) the cubature formula (8) will have degree of precision at least  $n$  if for non-negative integers  $r$  and  $q$ :

$$(9) \quad \sum_{v=1}^N A_v x_v^r y_v^q = \iint_D x^r y^q dx dy, \quad r + q = 0, 1, \dots, n.$$

There are  $\frac{1}{2}(n+1)(n+2)$  conditions imposed in (9). Hence, there are at least as many free parameters in (8) as there are conditions in (9) if

$$(10a) \quad N \geq \frac{(n+1)(n+2)}{6} \approx \frac{n^2}{6}$$

or

$$(10b) \quad n \leq \frac{1}{2}(\sqrt{1+24N}-3) \approx \sqrt{6N}.$$

We note from (10a) that the number of nodes for which a degree of precision  $n$  might possibly be obtained is about  $\frac{1}{3}$  the number used in the cubature formula (7) and about  $\frac{1}{6}$  the number used in (4).

This procedure is practical only if the integrals on the right-hand side of the equations in (9) can be evaluated explicitly (or perhaps if they can be accurately approximated with ease). This is, of course, the case if  $D$  is a rectangle or, in fact, any polygonal domain. However, the resulting system of  $\frac{1}{2}(n+1)(n+2)$  non-linear equations in  $3N$  unknowns must also have a real solution with nodes in  $D$ . There are many special cases in which the procedure can be employed successfully. Let us consider, for example, the simple case of one node,  $N = 1$ . Then from (10) we find that  $n = 1$  and there are only three equations in (9), namely,

$$A_1 = \iint_D dx dy, \quad A_1 x_1 = \iint_D x dx dy, \quad A_1 y_1 = \iint_D y dx dy.$$

Thus we find that the coefficient,  $A_1$ , is the area of the domain  $D$  and that the node,  $(x_1, y_1)$ , is at the centroid of  $D$ . The resulting cubature formula

$$J_1\{f\} = A_1 f(x_1, y_1)$$

is exact for all linear integrands in (3). This derivation and formula trivially generalize to any number of dimensions.

The next simplest case of only two nodes,  $N = 2$ , yields  $n = 2$  by (10), and hence a system of six non-linear algebraic equations of the form (9) must be solved. However, it is easy to show that this system does not always have a solution (see Problem 5). Thus we cannot, in general, determine a two point cubature formula with degree of precision two.

For domains which have symmetry about the  $x$ - and  $y$ -axes, the analysis of the system (9) can be simplified if the nodes are required to be symmetrically placed and have equal coefficients at corresponding locations. In this way various cubature schemes for integration over rectangles and circles may easily be derived.

When a formula of the form (8) satisfies the conditions (9), and hence has degree of precision  $n$  or more, an expression for the error can be derived by analogy with the proof of Theorem 0.1. For this purpose we require that the integrand function,  $f(x, y)$ , have continuous partial derivatives of all orders up to at least the  $(n + 1)$ st. Then we can expand the integrand about some point  $(x_0, y_0)$  into a finite Taylor's series with remainder in the form

$$f(x, y) = T_n(x, y) + R_n(x, y).$$

Here  $T_n(x, y)$  is a polynomial of degree at most  $n$  and  $R_n(x, y)$  is the known remainder which can be written symbolically as

$$R_n(x, y) = \frac{1}{(n + 1)!} \left[ (x - x_0) \frac{\partial}{\partial x} + (y - y_0) \frac{\partial}{\partial y} \right]^{n+1} f(\xi, \eta).$$

The error in the cubature formula is then

$$\begin{aligned} E_N\{f\} &\equiv J\{f\} - J_N\{f\} \\ &= J\{T_n\} + J\{R_n\} - J_N\{T_n\} - J_N\{R_n\} \\ &= J\{R_n\} - J_N\{R_n\}. \end{aligned}$$

Here we have used the fact that since the degree of precision is  $n$ ,  $J\{T_n\} = J_N\{T_n\}$ . In somewhat expanded form this error expression is

$$(11) \quad \begin{aligned} E_N\{f\} &= \frac{1}{(n + 1)!} \left\{ \iint_D \left[ (x - x_0) \frac{\partial}{\partial x} + (y - y_0) \frac{\partial}{\partial y} \right]^{n+1} \right. \\ &\quad \times f(\xi, \eta) dx dy - \sum_{v=1}^N A_v \\ &\quad \times \left. \left[ (x_v - x_0) \frac{\partial}{\partial x} + (y_v - y_0) \frac{\partial}{\partial y} \right]^{n+1} f(\xi_v, \eta_v) \right\}. \end{aligned}$$

The integrand in (11) is, of course, just symbolic since  $(\xi, \eta)$  depends upon  $(x, y)$  for purposes of the integration but not for the differentiations. Note that if the *maximum distance from  $(x_0, y_0)$  to any point in  $D$  or any node  $(x_v, y_v)$  is  $h$*  then the error satisfies  $E_N\{f\} = \mathcal{O}(h^{n+1})$ . In particular, if the coefficients  $A_v$  are all non-negative, so that

$$\sum_{v=1}^N A_v = \iint_D dx dy,$$

and all  $(n + 1)$ st order derivatives of  $f(x, y)$  are bounded by  $M_{n+1}$ , say, then with  $h$  as above we deduce from (11) that

$$(12) \quad |E_N\{f\}| \leq \frac{(2h)^{n+1}}{(n+1)!} 2M_{n+1} \iint_D dx dy.$$

*This estimate holds for any cubature formula which has non-negative coefficients and degree of precision  $n \geq 0$ , provided that the integrand has the appropriate smoothness.*

### 7.3. Separation of Variables

Perhaps the most obvious way to devise approximations for multiple integrals is by the repeated use of one dimensional quadrature formulae. The domain,  $D$ , must be somewhat special, or else it must be the union of special subdomains, in order for us to apply this method of separation of variables. In two dimensions the restriction is that vertical (or horizontal) lines have at most one segment in common with  $D$ . Integrals of the form (3) can then be written as

$$(13) \quad J\{f\} = \int_a^b \int_{y_1(x)}^{y_2(x)} f(x, y) dy dx,$$

where the segment  $y_1(x) \leq y \leq y_2(x)$  is in  $D$  for all  $x$  in  $[a, b]$ . If we introduce for each  $f(x, y)$  a function of the single variable  $x$  by the definition

$$(14a) \quad G(f; x) = \int_{y_1(x)}^{y_2(x)} f(x, y) dy, \quad a \leq x \leq b,$$

then the double integral (13) becomes

$$(14b) \quad J\{f\} = K\{G\} \equiv \int_a^b G(f; x) dx.$$

Now let us approximate the integral  $K\{G\}$  by some  $n$ -point quadrature formula with coefficients  $\alpha_j$  and *nodes*,  $x_j$ , which all lie in  $[a, b]$ , say

$$(15) \quad K_n\{G\} = \sum_{j=1}^n \alpha_j G(f; x_j).$$

The numbers  $G(f; x_j)$  which are required to evaluate this formula are given in (14a) as single integrals and hence can be approximated by applying other one dimensional quadrature formulae. For ease of presentation we use an  $m$ -point formula for each  $j$  and write the approximations to the  $G(f; x_j)$  as

$$(16) \quad G_m(f; x_j) \equiv \sum_{k=1}^m \beta_{jk} f(x_j, y_{jk}), \quad j = 1, 2, \dots, n.$$

Here the coefficients  $\beta_{jk}$  and nodes  $y_{jk}$  must, in general, depend upon the value  $x$ , since the interval of integration,  $[y_1(x), y_2(x)]$  in (14a) depends upon  $x$ . By using (16) the value  $K_n\{G\}$  of (15) finally yields the cubature formula

$$(17) \quad J_{mn}\{f\} = \sum_{j=1}^n \sum_{k=1}^m \alpha_j \beta_{jk} f(x_j, y_{jk}).$$

This formula employs  $mn$  nodes and is somewhat similar to that given by (4a) with coefficients (5). If the domain were a rectangle, then the same  $m$ -point formula could reasonably be used in (16) for all  $j$ . Then in (17) we could replace  $\beta_{jk}$  and  $y_{jk}$  by  $\beta_k$  and  $y_k$ , respectively, to get formal agreement with (4a) and (5).

The error in the cubature formula (17) is defined as

$$E_{mn}\{f\} = J\{f\} - J_{mn}\{f\}.$$

Let us introduce for the quadrature errors in (15) and (16) the notation:

$$(18) \quad \begin{aligned} e_n\{G\} &= K\{G\} - K_n\{G\}, \\ e_{mj}\{f\} &= G(f; x_j) - G_m(f; x_j), \quad j = 1, 2, \dots, n. \end{aligned}$$

Then from (14b) we have

$$(19) \quad \begin{aligned} E_{mn}\{f\} &= K\{G\} - K_n\{G\} + K_n\{G\} - J_{mn}\{f\} \\ &= e_n\{G\} + \sum_{j=1}^n \alpha_j [G(f; x_j) - G_m(f; x_j)] \\ &= e_n\{G\} + \sum_{j=1}^n \alpha_j e_{mj}\{f\}. \end{aligned}$$

It is interesting to note that when the degrees of precision of the quadrature formulae (15) and (16) are known we do not, in general, know the degree of precision of the cubature formula (17). This is because  $G(f; x)$  is not generally a polynomial in  $x$  when  $f(x, y)$  is a polynomial. In fact, it is easy to see that  $J_{mn}\{f\}$  may not even be exact for constant integrands when the quadrature formulae employed have arbitrarily high degrees of precision.

If the bounding curves  $y_1(x)$  and  $y_2(x)$  of  $D$  are polynomials of degree at most  $s \geq 0$ , then lower bounds can be given for the degree of precision. If  $f(x, y)$  is a polynomial of degree  $p$  then, by (14a),  $G(f; x)$  is a polynomial of degree at most  $s(p + 1)$ . So if (15) has degree of precision  $s(p + 1)$  and (16) has degree of precision  $p$  then  $J_{mn}\{f\}$  has degree of precision at least  $p$ . Of course, if the domain is a rectangle, i.e.,  $s = 0$ , then  $J_{mn}\{f\}$

has a degree of precision which is at least the minimum of those for (15) and (16).

From this result it follows that cubature formulae of high degree of precision with relatively few nodes may be devised if the domain of integration is a rectangle. To get degree of precision  $n$  in such a case we use  $[(n + 1)/2]$ -point Gaussian quadrature formulae as the two relevant schemes for (15) and (16). It is here assumed that  $n$  is odd and the total number of nodes required is then only  $\frac{1}{4}(n + 1)^2$ . For large values of  $n$  this is about half the number of points that were required in the efficient interpolation quadrature scheme (7) with the same degree of precision (and about  $\frac{2}{3}$  the number required in Subsection 7.2 by the method of undetermined coefficients). However, none of the nodes in the Gaussian scheme can be on the boundary of the domain and hence its usefulness in composite cubature formulae is reduced.

The extension to higher dimensions of the method of separation of variables is fairly clear. The restrictions on the domain are somewhat complicated, but, for instance, it is sufficient for the domain to be convex. In particular, for rectangular parallelopipeds, only a single one dimensional quadrature formula need be specified for each dimension. If the appropriate Gaussian schemes are used in this case, we obtain degree of precision  $n$  (odd) by using only  $[(n + 1)/2]^p$  nodes in  $p$  dimensions.

#### 7.4. Composite Formulae for Multiple Integrals

Just as in the case of one dimensional integrals, it may be necessary to decompose the integral (1) into a sum of integrals over smaller non-overlapping domains. That is, if  $D_i \cap D_j$  has no inner points for  $i \neq j$  and

$$D = D_1 \cup D_2 \cup \dots \cup D_M,$$

then

$$(20) \quad J\{f\} \equiv \int_D f(\mathbf{x}) d\mathbf{x} = \sum_{i=1}^M \int_{D_i} f(\mathbf{x}) d\mathbf{x}.$$

If  $N$  nodes are used to calculate the integral over each of the primitive domains  $D_i$  say

$$J_N\{f, D_i\} \equiv \sum_{j=1}^N \alpha_{ij} f(\mathbf{x}_{ij}),$$

then *at most*  $MN$  evaluations of  $f(\mathbf{x})$  are used in (20). We say *at most* because a node  $\mathbf{x}$  of  $D_i$  may also be a node of  $D_j$  but  $f(\mathbf{x})$  need only be found once for such a node.

If the region  $D$  is a  $p$ -dimensional rectangular parallelopiped, then a corner (or vertex) node of  $D_i$  may also occur in as many as  $2^p - 1$

adjoining cells  $D_j$ . Hence the amount of work necessary to evaluate the integrand may be minimized by selecting as many nodes as possible to be vertex nodes, then edge nodes, then face nodes, etc. For example, in two dimensions, the scheme of Figure 1b is more efficient than the scheme of Figure 1a.

If the cubature formula used in  $D_i$ ,  $J_N\{f; D_i\}$ , has degree of precision  $n$ , and has non-negative weights  $\{\alpha_{ij}\}$ , then just as in the derivation of (12),

$$(21) \quad \left| \int_{D_i} f(\mathbf{x}) d\mathbf{x} - \sum_{j=1}^N \alpha_{ij} f(\mathbf{x}_{ij}) \right| \leq \frac{(ph)^{n+1}}{(n+1)!} 2M_{n+1} \int_{D_i} d\mathbf{x},$$

when  $D_i$  is contained in a cube of side  $2h$ , and

$$\left| \frac{\partial^{n+1}}{\partial x_1^{j_1} \dots \partial x_p^{j_p}} f(\mathbf{x}) \right| \leq M_{n+1}$$

for all  $\mathbf{x}$  in  $D$  and all  $\{j_k\}$  satisfying  $\sum_{k=1}^p j_k = n+1$ . With these conventions,

it is then a simple matter to derive the fundamental estimate of the error in the composite cubature formula,

$$(22) \quad \left| J\{f\} - \sum_{i=1}^M J_N\{f; D_i\} \right| \leq \frac{(ph)^{n+1}}{(n+1)!} 2M_{n+1} \int_D d\mathbf{x}.$$

## PROBLEMS, SECTION 7

**1.** Devise the cubature schemes indicated by (4a) and (5) for equally spaced nodes when (1)  $m = n = 0$ ; (2)  $m = 0, n = 1$ ; (3)  $m = n = 1$ ; (4)  $m = n = 2$ . Case (4) is the generalization of Simpson's rule and (3) is the generalization of the trapezoidal rule to integrals over rectangles.

**2.** Determine the general cubature schemes for  $n = 0, 1, 2$  determined by integrating  $P_n(x, y)$ , given in (6.10) of Chapter 6, over an arbitrary domain  $D$ . Specialize these results for a rectangle  $a \leq x \leq b, c \leq y \leq d$ . Take uniform spacing in this rectangle in each case to simplify further. (Note: These schemes are not uniquely determined; see Figure 1.)

**3.** Compare the schemes (3) and (4) of Problem 1 to those with  $n = 1$  and  $n = 2$ , respectively, in Problem 2 by approximating the integral

$$\int_0^1 \int_1^2 x \sqrt{9 - y^2} dx dy.$$

Try at least two of the nodal schemes for each case of the methods of Problem 2.

**4.** Determine the ratio  $M/N$  for the number of nodes required in four and five dimensions, to extend the formulae  $J_N\{f\}$  and  $K_M\{f\}$  of Subsection 7.1 which have a degree of precision at least  $n$ .

**5.** Consider the case  $N = n = 2$  in the equations (9). Show that the resulting system does not have a solution in general by considering the special case

$$\iint_D x \, dx \, dy = \iint_D y \, dx \, dy = \iint_D xy \, dx \, dy = 0.$$

[Hint: Introduce the notation

$$A_1 x_1 = \xi, \quad A_1 y_1 = \eta, \quad A_1 = \zeta, \quad \iint_D dx \, dy = \zeta_0$$

and show that the system reduces to

$$\xi \eta \left( \frac{1}{\zeta} + \frac{1}{\zeta_0 - \zeta} \right) = 0, \quad \xi^2 \left( \frac{1}{\zeta} + \frac{1}{\zeta_0 - \zeta} \right) = \iint_D x^2 \, dx \, dy,$$

$$\eta^2 \left( \frac{1}{\zeta} + \frac{1}{\zeta_0 - \zeta} \right) = \iint_D y^2 \, dx \, dy.]$$

**6.** Give the proof of (22) in detail.

# 8

## Numerical Solution of Ordinary Differential Equations

### 0. INTRODUCTION

In order to study the effectiveness of various methods for the numerical solution of differential equation problems, we illustrate the theory for the case of the general first order ordinary differential equation

$$(1a) \quad \frac{dy}{dx} = f(x, y),$$

subject to the initial condition

$$(1b) \quad y(a) = y_0.$$

It is required to find a solution,  $y = y(x)$ , of the problem (1) in some interval, say  $a \leq x \leq b$ . Under suitable restrictions† on the function  $f(x, y)$ , it is well known that a unique solution exists.

The class of methods to be discussed uses a subdivision of the interval  $I \equiv [a, b]$ , by a finite set of distinct points

$$(2) \quad I_\Delta: x_0 = a, \quad x_{i+1} = x_i + \Delta x_i, \quad i = 0, 1, \dots, N.$$

Finer subdivisions also play a role and are denoted by the same generic symbol  $I_\Delta$ . In the present context, the set of points defining a subdivision is frequently called a *net*, *grid*, *lattice*, or *mesh*. The quantities  $\Delta x_i$  are called the *net spacings* or *mesh widths*. Corresponding to each point of the net we seek a quantity, say  $u_i$ , which is to approximate  $y_i \equiv y(x_i)$ , the exact

† For instance, existence and uniqueness of the solution are assured if  $f(x, y)$  is bounded, continuous in  $x$ , and Lipschitz continuous with respect to  $y$  in some sufficiently large rectangle  $R_C$ :  $[a \leq x \leq b, |y - y_0| \leq C]$ , see equation (1.5).

solution at the corresponding net point. The set of values  $\{u_i\}$  is called a *net function*. Clearly, the values  $\{y_i\}$  also form a net function on  $I_\Delta$ . We use the generic symbol  $\{u_i\}$  to denote a net function on any subdivision.

For most of the methods treated in this chapter, the quantities  $\{u_i\}$  are to be determined from a set of (usually non-algebraic) equations which in *some sense* approximate the system (1); these approximating equations are called *difference equations*. The natural requirements for the approximating difference equations are that for any function  $f(x, y)$  (in some class of sufficiently often differentiable functions):

- (a) They have a unique solution.
- (b) Their solution, at least for “sufficiently small” net spacings, should be “close” to the exact solution of (1).
- (c) Their solution should be “effectively computable.”

Property (a) is trivially satisfied by many of the difference equations to be studied, the so-called *explicit schemes*. Whether or not the *implicit schemes* satisfy condition (a) is determined by a study of the roots of a sequence of equations (or systems), of the form  $z = g(z)$  (see Section 2). In general, if  $\Delta x_i$  is small enough, the implicit equations have a unique solution.

Property (b) is related to the question of the convergence, as  $\max_i \Delta x_i \rightarrow 0$ , of  $\{u_i\}$  to  $\{y_i\}$ . The study of such *convergence properties* of the difference solution shall occupy a considerable part of this chapter. In Sections 1 through 3 we examine separately the convergence of each of several special methods. In Sections 5 and 6 we give a general treatment of convergence which includes the previous cases.

The vaguely formulated property (c) involves two important considerations: (i) the number of single precision computations required; (ii) the growth of roundoff errors in the computed difference “solution.” Of course, these two points are related since having to compensate for rounding errors by using more significant figures usually entails additional computations. A trivial first approximation of (i) is based on the operational count for infinite precision arithmetic. The growth of the roundoff error is related to the notion of *stability* of difference equations. The stability theory of difference equations treated in Section 5 is based on the study of difference equations with constant coefficients developed in Section 4. We establish the main general theorem of this chapter in Section 5 (i.e., stability is equivalent to convergence for consistent methods).

There are a number of systematic ways in which one can “derive” or rather generate difference equations that approximate or are consistent with (1). That is, these difference equations seem to be discrete models for the continuous problem (1). But, no matter how reasonable the

derivation, the efficacy of such difference equations can only be determined by checking conditions (a)–(c). In fact, in Subsection 1.4 we derive a discrete model which seems quite reasonable, but is absolutely useless since the growth of the roundoff error cannot be controlled (i.e., it is unstable).

Later, in Section 5, a simple criterion is developed for recognizing when a finite difference scheme is stable and convergent.

It should be recalled that some of the numerical methods for approximating solutions of ordinary differential equations, and systems of them, also have important theoretical applications. In fact, one of the basic existence and uniqueness proofs uses the *Euler-Cauchy difference method* of the next section. We resist the temptation to present such a proof here. Rather, we will assume that Problem (1) is “well-posed,” i.e., it has a unique solution with a certain number of continuous derivatives and furthermore, the solution depends differentiably on the initial data. As indicated in the footnote on page 364, we can guarantee the well-posedness of Problem (1) for a wide class of functions  $f(x, y)$ . We will be interested in showing that certain difference methods have properties (a)–(c) for such a class of functions,  $f(x, y)$ .

At the present time there seems to be no general way of formulating an “ideal method” for solving (1). An “ideal method” is one which requires the least amount of work (number of single precision computations) to produce an approximate solution of (1) accurate to within a given  $\epsilon > 0$ .

In the following sections a number of inequalities are derived with the use of two simple lemmas:

**LEMMA 1.** *For all real numbers  $z$ :*

$$(3) \quad 1 + z \leq e^z,$$

(where the equality holds only for  $z = 0$ ).

*Proof.* Since the function  $e^z$  has continuous derivatives of all orders we have by Taylor’s theorem

$$e^z = 1 + z + \frac{z^2}{2} e^{\theta z}, \quad 0 < \theta < 1.$$

But the last term on the right-hand side is non-negative and vanishes only when  $z = 0$  and so the lemma follows. ■

A simple corollary of this result is contained in

**LEMMA 1'.** *For all  $z$  such that  $1 + z \geq 0$ ,*

$$(4) \quad 0 \leq (1 + z)^n \leq e^{nz}, \quad n \geq 0.$$

*Proof.* Obvious from Lemma 1. ■

## 1. METHODS BASED ON APPROXIMATING THE DERIVATIVE: EULER-CAUCHY METHOD

To illustrate the basic concepts we consider the simple difference approximation to (0.1a) which results from approximating the derivative by a forward difference quotient,

$$(1a) \quad \frac{u_{j+1} - u_j}{h} = f(x_j, u_j), \quad j = 0, 1, \dots, N - 1.$$

Here, for simplicity only, we have chosen a *uniform net*

$$(2) \quad I_h: x_0 = a; \quad x_j = x_0 + jh, \quad j = 0, 1, \dots, N; \quad h = \frac{b - a}{N}.$$

The initial condition (0.1b) is replaced by

$$(1b) \quad u_0 = y_0 + e_0,$$

where we have intentionally permitted the introduction of an initial error,  $e_0$ . Equations (1) are the difference equations of the Euler-Cauchy method. This method is also called the *polygon method*, where the polygon is constructed by joining successive points  $(x_j, u_j)$  with straight line segments. Each segment has the slope given by the value of  $f$  at the left endpoint.

The existence of a unique solution  $u_j$  of the difference equations follows from writing (1a) as

$$(3) \quad u_{j+1} = u_j + hf(x_j, u_j), \quad j = 0, 1, \dots, N - 1.$$

Then with  $u_0$ , given in (1b), the above yields recursively  $u_1, u_2, \dots, u_N$  provided only that  $f(x_j, u_j)$  is defined.

The present analysis of (3) is based on using infinite precision arithmetic. That is, the numbers that would be calculated in finite precision arithmetic satisfy  $U_{j+1} = [U_j + hf(x_j, U_j)] + \rho_{j+1}$  where  $\rho_{j+1}$  is the rounding error made in evaluating the term in brackets. Later on, in Subsection 2, we study the error,  $U_j - y_j$ , as  $h \rightarrow 0$ .

We now turn to a consideration of the error  $\{e_j\}$  defined by

$$(4) \quad e_j = u_j - y_j, \quad j = 0, 1, \dots, N.$$

For this study we require that, in some region  $S$  to be specified later,  $f(x, y)$  be continuous in  $x$  and satisfy a uniform Lipschitz condition in the variable  $y$ :

$$(5) \quad |f(x, y) - f(x, y')| \leq K|y - y'|, \quad \text{for some constant } K > 0; \\ \text{all } (x, y) \text{ and } (x, y') \in S.$$

[If  $K = 0$ , then  $f(x, y)$  is independent of  $y$  and the problem (0.1) is a simple problem of quadrature treated in Chapter 7.] In addition, we will need a

measure of the error by which the exact solution of (0.1) fails to satisfy the difference relation (1a). This error is called the *local truncation error* or discretization error and is defined by:

$$\tau_{j+1} = \frac{y_{j+1} - y_j}{h} - f(x_j, y_j), \quad j = 0, 1, \dots, N - 1.$$

This relation is frequently written as

$$(6) \quad y_{j+1} = y_j + hf(x_j, y_j) + h\tau_{j+1}, \quad j = 0, 1, \dots, N - 1.$$

An explicit representation of the  $\tau_j$  will be derived shortly [under additional conditions on  $f(x, y)$ ]. If the  $\tau_j$  vanish as  $h \rightarrow 0$ , we say that the difference equations are *consistent* with the differential equation. But as will be seen in Subsection 2, for another consistent scheme, the corresponding difference solution  $\{u_j\}$  can diverge as  $h \rightarrow 0$ , from the exact solution  $\{y_j\}$  even though  $e_0$  is small. However, in the present case, we have

**THEOREM 1.** *Let  $\{u_j\}$  be the solution of (1) and  $y(x)$  the solution of (0.1) where  $f(x, y)$  satisfies (5) in the strip  $S: [a \leq x \leq b, |y| < \infty]$ . Then, with the definitions (6) of  $\{\tau_j\}$ :*

$$(7) \quad |u_j - y(x_j)| \leq e^{K(x_j - a)} \left[ |e_0| + \frac{\tau}{K} \right], \quad j = 0, 1, \dots, N,$$

where

$$\tau \equiv \max_j |\tau_j|.$$

*Proof.* The subtraction of (6) from (3) yields

$$e_{j+1} = e_j + h[f(x_j, u_j) - f(x_j, y_j)] - h\tau_{j+1}, \quad j = 0, 1, \dots, N - 1.$$

By means of the Lipschitz condition we deduce that

$$|e_{j+1}| \leq (1 + hK)|e_j| + h\tau.$$

This inequality yields recursively

$$\begin{aligned} |e_{j+1}| &\leq (1 + hK)^2 |e_{j-1}| + [1 + (1 + hK)]h\tau, \\ &\leq (1 + hK)^3 |e_{j-2}| + [1 + (1 + hK) + (1 + hK)^2]h\tau, \\ &\vdots \\ &\leq (1 + hK)^{j+1} |e_0| + \left[ \frac{(1 + hK)^{j+1} - 1}{K} \right] \tau, \end{aligned}$$

where we have summed the geometric progression. Since  $K > 0$ , we may apply Lemma 0.1' in the form

$$(1 + hK)^{j+1} \leq e^{(j+1)hK} = e^{K(x_{j+1} - x_0)}.$$

Hence for all  $j \leq N - 1$ ,

$$(8) \quad |e_{j+1}| \leq e^{K(x_{j+1} - x_0)} \left( |e_0| + \frac{\tau}{K} \right),$$

and the theorem follows. ■

The simple bound of Theorem 1 shows that the error at any net point,  $x_j$ , will be small if both the initial error,  $e_0$ , and the maximum local truncation error,  $\tau$ , are small. Now, the value of  $|e_0|$  is determined by the accuracy with which the number  $y_0$ , the initial condition, is approximated. On the other hand, we may guarantee that  $\tau$  can be made arbitrarily small by picking  $h$  sufficiently small, if  $d^2y/dx^2$  is continuous in  $[a, b]$ . That is, from Taylor's theorem,

$$y(x_j + h) = y(x_j) + h \frac{dy(x_j)}{dx} + \frac{h^2}{2} \frac{d^2y(x_j + \theta_j h)}{dx^2},$$

$$0 < \theta_j < 1; \quad j = 0, 1, \dots, N - 1.$$

However, since  $y(x)$  is the solution of (0.1),  $dy(x_i)/dx = f(x_i, y_i)$  and a comparison of the above with (6) yields

$$(9) \quad \tau_{j+1} = \frac{h}{2} \frac{d^2y(x_j + \theta_j h)}{dx^2}, \quad 0 < \theta_j < 1, \quad j = 0, 1, \dots, N - 1.$$

Using this representation of  $\tau_j$  in Theorem 1 and the formula obtained from (0.1a) by differentiation

$$\frac{d^2y}{dx^2} = f_x(x, y) + f_y(x, y) \frac{dy}{dx},$$

we obtain a result which may be summarized as the

**COROLLARY.** *If, in addition to the hypothesis of Theorem 1,  $f_x(x, y)$  and  $f_y(x, y)$  are continuous in  $S$ , then*

$$|e_j| \leq e^{K(x_j - a)} \left( |e_0| + h \frac{M_2}{2K} \right) \leq e^{K(b - a)} \left( |e_0| + h \frac{M_2}{2K} \right)$$

where†

$$M_2 = \max_{a \leq x \leq b} \left| \frac{d^2y(x)}{dx^2} \right|. \quad ■$$

If  $e_0 = 0$  or  $|e_0| \leq \alpha h$  for some constant  $\alpha$ , then as a consequence of the corollary,  $\lim_{h \rightarrow 0} e_j = 0$ , or more precisely, the maximum norm of the error  $\{e_j\}$  is at most  $\mathcal{O}(h)$  and converges uniformly to zero since the rightmost

† If in  $S$ ,  $|f_x| \leq P$ ,  $|f_y| \leq Q$ , and  $|f| \leq R$ , then  $\left| \frac{d^2y}{dx^2} \right| \leq P + QR$ . Hence for such a class of functions  $f(x, y)$ , we find the a priori bound  $M_2 \leq P + QR$ .

bound is independent of  $j$ . Note that since  $f_y$  is assumed continuous in the corollary, the condition (5) need not be postulated but can, in fact, be deduced if  $K \equiv \sup_s |f_y|$  is finite.

In general, the bounds on the error are usually tremendous overestimates. It is possible, however, to obtain more precise expressions for the error, essentially under the conditions of the corollary. These expressions are in turn not practical since they cannot be evaluated explicitly. But since they do have analytical significance we present

**THEOREM 2.** *If  $\{e_j\}$  and  $\{\tau_j\}$  are defined by (4) and (6) respectively, and  $f_y(x, y)$  is continuous in  $S$ , then there exist numbers  $\phi_i$  in  $0 < \phi_i < 1$  such that*

$$(10) \quad e_i = A_{i,0}e_0 - h \sum_{j=1}^i A_{i,j}\tau_j, \quad i = 1, 2, \dots, N;$$

where  $A_{0,0} \equiv 1$  and

$$(11) \quad A_{i,j} \equiv \begin{cases} 0 & j \geq i+1 \\ 1 & j = i, \\ \alpha_{i-1}A_{i-1,j} & j < i, \alpha_i = 1 + hf_y(x_i, y_i + \phi_i e_i). \end{cases}$$

*Proof.* The proof is similar to that of Theorem 1 but now the mean value theorem is used in place of the Lipschitz condition. Thus, from (6) and (3),

$$\begin{aligned} e_{i+1} &= e_i + h[f(x_i, u_i) - f(x_i, y_i)] - h\tau_{i+1} \\ &= \alpha_i e_i - h\tau_{i+1}. \end{aligned}$$

To show that the algebraic manipulations, in the recursive application of the above, yield quantities of the form (10) and (11), we proceed by induction. Then with  $i = 0$  in the above

$$\begin{aligned} e_1 &= \alpha_0 e_0 - h\tau_1 \\ &= A_{1,0}e_0 - hA_{1,1}\tau_1. \end{aligned}$$

Now we assume (10) to be valid and use (11) to obtain

$$\begin{aligned} e_{i+1} &= \alpha_i A_{i,0}e_0 - h \sum_{j=1}^i \alpha_i A_{i,j}\tau_j - h\tau_{i+1} \\ &= A_{i+1,0}e_0 - h \sum_{j=1}^i A_{i+1,j}\tau_j - hA_{i+1,i+1}\tau_{i+1} \\ &= A_{i+1,0}e_0 - h \sum_{j=1}^{i+1} A_{i+1,j}\tau_j. \end{aligned}$$

The induction is thus complete and the theorem follows. ■

By restricting  $h$  to be sufficiently small the exact error expression (10) can be reduced to a simple form which has practical significance. We state this result as

**COROLLARY 1.** *Under the hypothesis of Theorem 2 let*

$$d \equiv \inf_s f_y(x, y), \quad D \equiv \sup_s f_y(x, y)$$

*be finite and restrict  $h$  so that*

$$(12) \quad 1 + hd \geq 0.$$

*Then for each  $i = 1, 2, \dots, N$ , there exist three numbers  $p_i$ ,  $q_i$ , and  $t_i$  in the intervals*

$$(13) \quad d \leq p_i \leq D, \quad d \leq q_i \leq D, \quad \min_{1 \leq j \leq i} \tau_j \leq t_i \leq \max_{1 \leq j \leq i} \tau_j,$$

*such that*

$$(14) \quad e_i = (1 + hp_i)^i e_0 - \left[ \frac{(1 + hq_i)^i - 1}{q_i} \right] t_i; \quad i = 1, 2, \dots, N.$$

*Proof.* We note that from (11) and (12) it follows that

$$0 \leq 1 + hd \leq \alpha_j \leq 1 + hD, \quad j = 0, 1, \dots, N - 1.$$

Then

$$(1 + hd)^i \leq A_{i,0} = \alpha_0 \alpha_1 \cdots \alpha_{i-1} \leq (1 + hD)^i,$$

and hence there is a number  $p_i$  in the interval  $[d, D]$  such that

$$A_{i,0} = (1 + hp_i)^i.$$

Now define the quantities

$$(15) \quad S_i = \sum_{j=1}^i A_{i,j}, \quad t_i = \sum_{j=1}^i \left( \frac{A_{i,j}}{S_i} \right) \tau_j.$$

The  $A_{i,j}$  are non-negative as a result of condition (12). Hence  $t_i$ , which is an average with non-negative weights of the  $\tau_j$ , must satisfy condition (13) (see Lemma 1.1 of Chapter 7 which can be used to prove this assertion). We also note, using (11), that

$$0 \leq (1 + hd)^{i-j} \leq A_{i,j} = \alpha_j \alpha_{j+1} \cdots \alpha_{i-1} \leq (1 + hD)^{i-j}, \quad j < i;$$

$$A_{i,i} = 1.$$

Then from the definition of  $S_i$

$$\begin{aligned} 1 + (1 + hd) + \cdots + (1 + hd)^{i-1} &\leq S_i \\ &\leq 1 + (1 + hD) + \cdots + (1 + hD)^{i-1}, \end{aligned}$$

or by summing the progressions

$$\frac{(1 + hd)^i - 1}{hd} \leq S_i \leq \frac{(1 + hD)^i - 1}{hD}; \quad i = 1, 2, \dots, N.$$

But the functions  $[(1 + z)^i - 1]/z$  are continuous functions of  $z$  and hence there exist numbers  $q_i$ , in the interval  $[d, D]$ , such that

$$S_i = \frac{(1 + hq_i)^i - 1}{hq_i}; \quad i = 1, 2, \dots, N.$$

The corollary now follows by using the expressions for  $A_{i,0}$ ,  $S_i$ , and  $t_i$  in (10). ■

The form of the error given in (14) can be used to derive practical information in many cases. Clearly, if  $d$  and  $D$  are known, or can be estimated, we can obtain upper and *lower* bounds on the factors which multiply the two error terms ( $e_0$  and  $t_i$ ). A more striking application occurs, however, when  $f(x, y)$  is such that  $D < 0$  (i.e.,  $f_y < 0$ ). Now clearly,  $p_i < 0$ ,  $q_i < 0$ , and by the condition (12) imposed on  $h$ :

$$0 \leq (1 + hp_i) < 1, \quad 0 \leq (1 + hq_i) < 1.$$

Then by Lemma 0.1'

$$(1 + hp_i)^i < e^{ihp_i} = e^{(x_i - a)p_i},$$

or since  $p_i < 0$ , this may be written as

$$(1 + hp_i)^i < e^{-|p_i(x_i - a)|} < 1.$$

Similarly,  $0 \leq (1 + hq_i)^i < 1$ , and by taking absolute values in (14), we find

**COROLLARY 2.** *If  $f_y < 0$ , then the hypotheses of Theorem 2 and its Corollary 1 imply*

$$(16) \quad |e_i| \leq e^{-|p_i(x_i - a)|}|e_0| + \left| \frac{t_i}{D} \right|, \quad i = 1, 2, \dots, N. \quad \blacksquare$$

This result shows that the initial error cannot grow if  $f_y < 0$  and further, that the local truncation errors in this case contribute at most an amount  $\tau/|D|$ .

## 1.1. Improving the Accuracy of the Numerical Solution

We now improve upon the corollary to Theorem 1 by characterizing the  $\mathcal{O}(h)$  term in the error  $\{e_j\}$ .

**THEOREM 3.** Let the solution  $y = y(x)$  of (0.1) have three continuous and bounded derivatives: let  $f_{yy}(x, y)$  be continuous and bounded, and let the initial error  $e_0$  in the difference solution  $\{u_i\}$  of (1) be

$$e_0 = \xi_0 h,$$

where  $\xi_0$  is independent of  $h$ . Then

$$e_j = h\xi(x_j) + \mathcal{O}(h^2), \quad j = 0, 1, \dots, N,$$

where  $\xi(x)$  is the solution† of the linear problem

$$\frac{d\xi}{dx} = f_y(x, y(x))\xi - \frac{1}{2}y''(x),$$

$$\xi(a) = \xi_0.$$

*Proof.* As in the proof of Theorem 2,

$$e_{i+1} = \alpha_i e_i - h\tau_{i+1}.$$

But now in (9) and (11) we use the extra differentiability properties to obtain

$$\begin{aligned} \alpha_i &= 1 + hf_y(x_i, y_i) + hf_{yy}(x_i, y_i + \phi_i' e_i)\phi_i e_i \\ \tau_{i+1} &= \frac{h}{2} y''(x_i) + \frac{h}{2} y'''(x_i + \theta_i' h)\theta_i h, \quad 0 < \phi_i, \phi_i', \theta_i, \theta_i' < 1. \end{aligned}$$

Then from this,

$$e_{i+1} = [1 + hf_y(x_i, y_i)]e_i - \frac{h^2}{2} y''(x_i) + h[\mathcal{O}(e_i^2) + \mathcal{O}(h^2)].$$

By using the differential equation which defines  $\xi(x)$ , Taylor's expansion yields

$$\begin{aligned} \xi(x_{i+1}) &= \xi(x_i) + h\xi'(x_i) + \frac{h^2}{2} \xi''(x_i + \psi_i h) \\ &= [1 + hf_y(x_i, y_i)]\xi(x_i) - \frac{h}{2} y''(x_i) + \mathcal{O}(h^2). \end{aligned}$$

We now form the quantities

$$\delta_j \equiv e_j - h\xi_j$$

and find

$$\delta_{i+1} = [1 + hf_y(x_i, y_i)]\delta_i + h[\mathcal{O}(e_i^2) + \mathcal{O}(h^2)], \quad i = 1, 2, \dots$$

† Under the hypothesis,  $\xi(x)$  exists and has a continuous second derivative. In fact,  $\xi(x)$  can be explicitly represented by quadratures.

Now we observe, as remarked after the corollary to Theorem 1, that  $|e_i| = \mathcal{O}(h)$ . Hence we may delete the term  $\mathcal{O}(e_i^2)$ . But, from the specific initial conditions chosen for  $\xi(x)$  we have  $\delta_0 = 0$ , and hence a recursive application of the formulae for  $\delta_i$  yields, as in the derivation of (7),

$$|\delta_j| \leq e^{K(b-a)} \frac{[\mathcal{O}(h^2)]}{K} = \mathcal{O}(h^2)$$

and the theorem follows. ■

To apply Theorem 3, we introduce the notation  $u_j(h)$  to indicate the dependence of the numerical solution on the net spacing. Then the theorem states that with, say  $x_j = z$ ,

$$u_j(h) = y(z) + h\xi(z) + \mathcal{O}(h^2).$$

Similarly, with the net spacing  $h/2$  and  $x_{2j} = z$ , we have

$$u_{2j}\left(\frac{h}{2}\right) = y(z) + \frac{h}{2}\xi(z) + \mathcal{O}(h^2).$$

Then

$$2u_{2j}\left(\frac{h}{2}\right) - u_j(h) = y(z) + \mathcal{O}(h^2),$$

and an extra order of magnitude in accuracy is obtained if we use as the difference approximation, at any point  $x_j = z$  of the net with spacing  $h$ , the quantity

$$\bar{u}_j = 2u_{2j}\left(\frac{h}{2}\right) - u_j(h).$$

This requires computations with two nets of spacings  $h$  and  $h/2$  respectively. It should be observed that the formula for  $\bar{u}_j$  is similar to the formula which arises in Aitken's  $\delta^2$ -process in the iterative solution of arbitrary equations (see Subsection 2.4 of Chapter 3). In the present context this procedure is called *Richardson's deferred approach to the limit*, or *extrapolation to zero mesh width*. This extrapolation may be applied, in an appropriately modified form, to many of the numerical methods to be considered here.

## 1.2. Roundoff Errors

In actually performing the calculations required to evaluate (1), round-off errors will, in general, be introduced. Thus the numbers actually obtained will not be the set  $\{u_i\}$  but, say, some quantities  $\{U_i\}$ . These numbers satisfy equations of the form

$$(17a) \quad U_{i+1} = U_i + hf(x_i, U_i) + \rho_{i+1}, \quad i = 0, 1, \dots, N-1;$$

where  $\rho_{i+1}$  represents the error introduced by inexact evaluation of the quantity  $U_i + hf(x_i, U_i)$ . The  $\rho_i$  are called the *local roundoff errors*. If we let  $\rho_0$  be the initial roundoff error committed in evaluating  $y_0$ , then the initial condition becomes

$$(17b) \quad U_0 = y_0 + \rho_0.$$

Let the errors between the  $U_i$ , the actual numbers obtained in the computation, and the  $u_i$ , the exact solution of the difference equations, be denoted by

$$(18) \quad \epsilon_i \equiv U_i - u_i, \quad i = 0, 1, \dots, N,$$

Then from (1) and (17) we obtain

$$(19) \quad \begin{aligned} \epsilon_0 &= \rho_0 - e_0; \\ \epsilon_{i+1} &= \epsilon_i + h[f(x_i, U_i) - f(x_i, u_i)] + \rho_{i+1}, \quad i = 0, 1, \dots, N-1. \end{aligned}$$

It is clear that these equations for  $\epsilon_i$  are formally similar to those which determine the quantities  $e_i$ . In fact, the previous theorems and corollaries can be restated in an obvious way to give bounds and representations of the errors  $\epsilon_i$ . We shall return to the study of the growth of the  $\epsilon_i$  in Section 5. But, we now consider the more important total errors

$$(20) \quad E_i \equiv U_i - y(x_i) = e_i + \epsilon_i, \quad i = 0, 1, \dots, N,$$

between the actual numerical solution,  $U_i$ , and the exact solution of the differential equation,  $y(x_i)$ . In an obvious manner, we find that

$$(21) \quad \begin{aligned} E_0 &= \rho_0; \\ E_{i+1} &= E_i + h[f(x_i, U_i) - f(x_i, y(x_i))] - (h\tau_{i+1} - \rho_{i+1}), \\ &\quad i = 0, 1, \dots, N-1. \end{aligned}$$

Again we may prove the analogs of the previous results.

**THEOREM 4.** *Under the conditions of the corollary to Theorem 1, we find that the error (20) satisfies*

$$(22a) \quad |E_j| \leq e^{K(b-a)} \left[ |\rho_0| + \frac{1}{K} \left( \frac{hM_2}{2} + \frac{\rho}{h} \right) \right], \quad \text{for } j = 0, 1, \dots, N,$$

where the roundoff errors  $\rho_i$  are defined in (17) and

$$(22b) \quad \rho = \max_{1 \leq i \leq N} |\rho_i|, \quad M_2 = \max_{a \leq x \leq b} \left| \frac{d^2y(x)}{dx^2} \right|. \quad \blacksquare$$

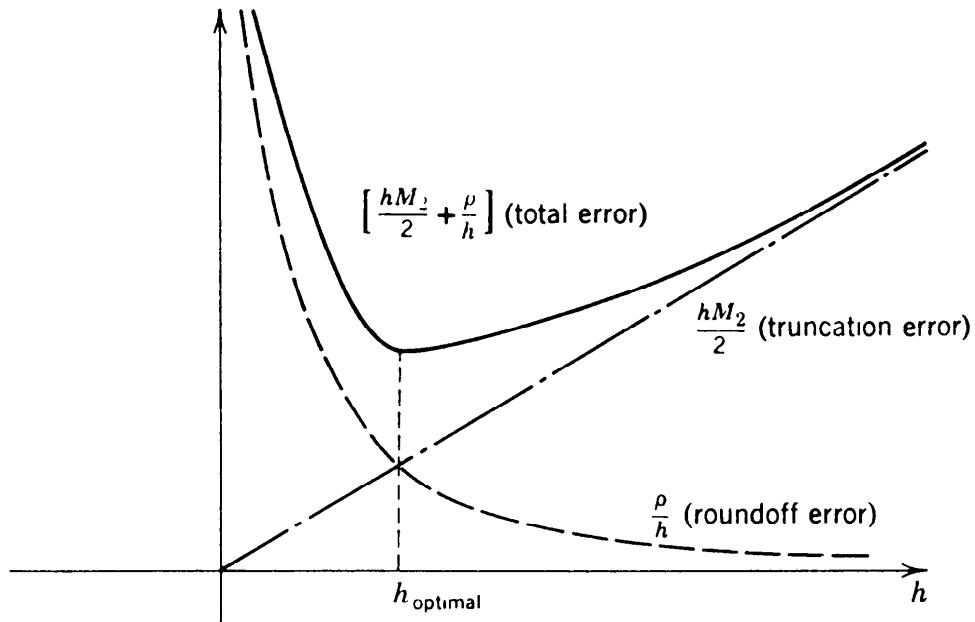


Figure 1. Comparison of truncation and roundoff error bounds as functions of  $h$ .

The dependence of this error bound on the net spacing,  $h$ , is illustrated in Figure 1.

Clearly, the choice of  $h$  for which the bound in (22) is a minimum is obtained when

$$(23) \quad \frac{hM_2}{2} = \frac{\rho}{h}, \quad \text{or} \quad h = \sqrt{2\rho/M_2}.$$

For this optimal value of  $h$ ,

$$\frac{hM_2}{2} + \frac{\rho}{h} = \sqrt{2M_2\rho}.$$

In many calculations performed on electronic computers  $\rho \ll M_2$ , and so the “optimal” value for  $h$  will be unnecessarily small and need not be employed. Furthermore, the bound (22) indicates that for fixed  $h$  no greater accuracy is obtained by reducing the roundoff error so that

$$\rho < \frac{h^2 M_2}{2};$$

in fact, any extra labor required for such computational precision is essentially wasted. If the relation (23) is approximately satisfied there might be some fortuitous cancellation of local roundoff and truncation errors.

On the other hand, we remark that (22) establishes the *convergence* as  $h \rightarrow 0$  of the Euler-Cauchy method (17) if the rounding error satisfies

$$|\rho_i| \leq \rho = \mathcal{O}(h^2), \quad \text{for } i = 1, 2, \dots, N$$

while initially

$$|E_0| = |\rho_0| = \xi_0 h.$$

In fact, under these circumstances

$$|E_j| = \mathcal{O}(h); \quad j = 0, 1, \dots, N.$$

We leave to Problem 2 the proof of the validity of the Richardson extrapolation to zero mesh width provided  $\rho = \mathcal{O}(h^3)$ .

### 1.3. Centered Difference Method

To obtain greater accuracy with a fixed mesh size we seek difference approximations with smaller local truncation errors. One such modification of (1a) is suggested by attempting to approximate the derivative at  $x_i$  by a more accurate expression than the forward difference quotient. We shall briefly examine here the use of the *centered* formula

$$(24a) \quad \frac{u_{i+1} - u_{i-1}}{2h} = f(x_i, u_i), \quad i = 1, 2, \dots, N-1.$$

However, in order to use these difference equations to compute  $\{u_i\}$ , two starting values are required, say

$$(24b) \quad u_0 = y_0 + e_0, \quad u_1 = y_1 + e_1.$$

The first value  $u_0$  is again the approximation of the exact initial data, while  $u_1$  should be determined such that  $|e_1|$  is “small.” This could be done by employing the Euler-Cauchy method in the interval  $0 \leq x \leq h$  with some smaller spacing, say  $h' = h/N'$  with  $N' \geq 1$ , or by developing the Taylor’s series

$$u_1 = y_0 + hy_0' + \frac{h^2}{2} y_0'',$$

with

$$y_0' = f(x_0, y_0), \quad y_0'' = f_x(x_0, y_0) + y_0' f_y(x_0, y_0).$$

However, this problem or similar ones will occur again and shall be discussed in more detail later.

The truncation error in (24a) is now defined by

$$(25) \quad y_{i+1} = y_{i-1} + 2hf(x_i, y_i) + 2h\tau_{i+1}, \quad i = 1, 2, \dots, N-1.$$

Let us make sure that  $y(x)$  has three continuous derivatives by assuming that  $f(x, y)$  has continuous second derivatives. Then by Taylor’s theorem

$$y_{i+1} = y(x_i + h) = y_i + hy_i' + \frac{h^2}{2} y_i'' + \frac{h^3}{3!} y'''(\xi_{i+1}), \quad x_i < \xi_{i+1} < x_{i+1};$$

$$y_{i-1} = y(x_i - h) = y_i - hy_i' + \frac{h^2}{2} y_i'' - \frac{h^3}{3!} y'''(\xi_{i-1}), \quad x_{i-1} < \xi_{i-1} < x_i.$$

These equations imply

$$y_{i+1} = y_{i-1} + 2hy_i' + \frac{h^3}{3!} [y'''(\xi_{i+1}) + y'''(\xi_{i-1})],$$

and since  $y_i' = f(x_i, y_i)$ , a comparison with (25) yields

$$(26) \quad \begin{aligned} \tau_{i+1} &= \frac{h^2}{6} \frac{y'''(\xi_{i+1}) + y'''(\xi_{i-1})}{2}, \\ &= \frac{h^2}{6} y'''(\xi_i), \quad i = 1, 2, \dots, N-1. \end{aligned}$$

Here we have used the continuity of  $y''(x)$  to replace the average of third derivatives by an intermediate value at  $\xi_i$ . Hence the local truncation error of the centered scheme is smaller than the truncation error (9) of the Euler-Cauchy method as  $h \rightarrow 0$ . The present  $\tau_i$  vanish to second order in  $h$  and we thus call the centered scheme (24a) a second order method.

To show that these effects are indeed relevant for the convergence of the finite difference solution, we again consider the errors  $e_i \equiv u_i - y(x_i)$  and find from (24a) and (25)

$$\begin{aligned} e_{i+1} &= e_{i-1} + 2h[f(x_i, u_i) - f(x_i, y(x_i))] - 2h\tau_{i+1} \\ &= e_{i-1} + 2h \frac{\partial f(x_i, y(x_i)) + \theta_i e_i}{\partial y} e_i - 2h\tau_{i+1}, \quad i = 1, 2, \dots, N-1. \end{aligned}$$

To demonstrate convergence, let us introduce the bounds

$$(27) \quad K \geq 2 \left| \frac{\partial f}{\partial y} \right|, \quad M_3 \geq |y'''(x)|, \quad \tau = \frac{h^2}{3} M_3 \geq 2|\tau_i|.$$

Hence by taking absolute values

$$(28) \quad |e_{i+1}| \leq hK|e_i| + |e_{i-1}| + h\tau, \quad i = 1, 2, \dots, N-1.$$

To obtain bounds on the  $|e_i|$  we introduce a comparison or majorizing set of quantities  $\{a_i\}$  defined by

$$(29) \quad \begin{aligned} a_0 &\equiv \max(|e_0|, |e_1|), \\ a_{i+1} &= (1 + hK)a_i + h\tau, \quad i = 0, 1, \dots, N-1. \end{aligned}$$

From the definition it is clear that  $a_0 \geq |e_0|$ . We will show by induction that  $a_j \geq |e_j|$ ,  $j = 0, 1, \dots, N$ . Assume that  $a_j \geq |e_j|$ ,  $j = 0, 1, \dots, i$ . Equation (28) yields

$$\begin{aligned} |e_{i+1}| &\leq hKa_i + a_{i-1} + h\tau \\ &\leq (1 + hK)a_i + h\tau \\ &= a_{i+1}. \end{aligned}$$

Here we have employed (29) and the obvious relation  $a_i \geq a_{i-1}$ . Hence the induction proof is complete and  $a_j \geq |e_j|$  for  $j = 0, 1, \dots, N$ . However, by the usual recursive application of (29) we now obtain

$$\begin{aligned}
 (30) \quad |e_N| &\leq a_N \\
 &= (1 + hK)^N a_0 + \frac{(1 + hK)^N - 1}{K} \tau \\
 &\leq (1 + hK)^N \left( a_0 + \frac{\tau}{K} \right) \\
 &\leq e^{K(b-a)} \left[ \max(|e_0|, |e_1|) + h^2 \frac{M_3}{3K} \right].
 \end{aligned}$$

By comparing (30) with the result in the corollary to Theorem 1, we now find that the error is of order  $h^2$  if the initial errors,  $e_0$  and  $e_1$ , are proportional to  $h^2$ . So in the present centered scheme, the higher order local truncation error (26) is reflected in faster convergence of  $\{u_i\}$  to  $\{y(x_i)\}$  as  $h \rightarrow 0$ . We might naturally expect this to be the case in general, and so seek difference equations with local truncation errors of arbitrarily high order in  $h$ . However, as is demonstrated in the next subsection, this expectation is not always realized.

It should be mentioned that roundoff effects can also be included in the study of the present centered scheme.

**THEOREM 5.** *Let the roundoffs  $\rho_i$  satisfy*

$$U_0 = y_0 + \rho_0, \quad U_1 = y_1 + \rho_1,$$

where

$$U_{i+1} = U_{i-1} + 2hf(x_i, U_i) + \rho_{i+1}, \quad \text{for } i = 1, 2, \dots, N-1,$$

$$\max_{2 \leq i \leq N} |\rho_i| = \rho.$$

Then  $E_i = U_i - y_i$  can be bounded by

$$|E_i| \leq e^{K(b-a)} \left\{ \max(|\rho_0|, |\rho_1|) + \left[ h^2 \frac{M_3}{3K} + \frac{\rho}{hK} \right] \right\}, \quad i = 0, 1, \dots, N,$$

provided (27) holds. ■

Now the error is at least of order  $h^2$ , if the maximum roundoff error satisfies  $\rho = \mathcal{O}(h^3)$  as  $h \rightarrow 0$ , while  $\rho_0 = \mathcal{O}(h^2)$ ,  $\rho_1 = \mathcal{O}(h^2)$ .

#### 1.4. A Divergent Method with Higher Order Truncation Error

To demonstrate the care which must be taken in generating difference schemes, we consider a case with third order local truncation error but

which is completely unsuitable for computation. The basis for this scheme is an attempt to approximate better the derivative at  $x_i$ , and thus to obtain a local truncation error which is higher order in  $h$ . For this purpose we consider a difference equation of the form

$$(31) \quad a_1 u_{i+1} + a_2 u_i + a_3 u_{i-1} + a_4 u_{i-2} = f(x_i, u_i),$$

$$i = 2, 3, \dots, N - 1;$$

and seek coefficients  $a_1, \dots, a_4$  such that the local truncation error is of as high an order, in  $h$ , as possible. This is essentially the *method of undetermined coefficients* applied to the problem of approximating derivatives (see Subsection 5.1 of Chapter 6). That is, recalling  $y'(x_i) = f(x_i, y(x_i))$ , we define  $\tau_i$  by

$$(32) \quad a_1 y(x_{i+1}) + a_2 y(x_i) + a_3 y(x_{i-1}) + a_4 y(x_{i-2}) - y'(x_i) = \tau_{i+1}.$$

Then if  $y(x)$  has four continuous derivatives, Taylor's theorem yields

$$\begin{aligned} y_{i+1} &= y(x_i + h) = y_i + hy'(x_i) + h^2y''(x_i)/2 + h^3y'''(x_i)/3! \\ &\quad + h^4y^{iv}(\xi_{i+1})/4! \end{aligned}$$

$$(33) \quad \begin{aligned} y_{i-1} &= y(x_i - h) = y_i - hy'(x_i) + h^2y''(x_i)/2 - h^3y'''(x_i)/3! \\ &\quad + h^4y^{iv}(\xi_{i-1})/4! \end{aligned}$$

$$\begin{aligned} y_{i-2} &= y(x_i - 2h) = y_i - 2hy'(x_i) + 4h^2y''(x_i)/2 - 8h^3y'''(x_i)/3! \\ &\quad + 16h^4y^{iv}(\xi_{i-2})/4! \end{aligned}$$

Forming the sum indicated in (32) and requiring as many terms as possible to vanish, we find

$$\begin{aligned} a_1 + a_2 + a_3 + a_4 &= 0 \\ (a_1 + 0 - a_3 - 2a_4)h &= 1 \\ (a_1 + 0 + a_3 + 4a_4)h^2 &= 0 \\ (a_1 + 0 - a_3 - 8a_4)h^3 &= 0. \end{aligned}$$

This system of linear equations has the unique solution

$$(34) \quad a_1 = \frac{1}{3h}, \quad a_2 = \frac{1}{2h}, \quad a_3 = -\frac{1}{h}, \quad a_4 = \frac{1}{6h}.$$

Now (32) can be written as

$$(35) \quad \begin{aligned} y(x_{i+1}) &= -\frac{3}{2}y(x_i) + 3y(x_{i-1}) - \frac{1}{2}y(x_{i-2}) \\ &\quad + 3hf(x_i, y(x_i)) + 3h\tau_{i+1} \end{aligned}$$

where the local truncation error is from (33) and (34) in (32)

$$(36) \quad \tau_{i+1} = \frac{h^3}{4!} [\frac{1}{3}y^{iv}(\xi_{i+1}) - y^{iv}(\xi_{i-1}) + \frac{8}{3}y^{iv}(\xi_{i-2})].$$

The difference equations (31) become

$$(37a) \quad u_{i+1} = -\frac{3}{2}u_i + 3u_{i-1} - \frac{1}{2}u_{i-2} + 3hf(x_i, u_i), \\ i = 2, 3, \dots, N-1;$$

and to these must be adjoined starting values, say,

$$(37b) \quad u_0 = y_0 + e_0, \quad u_1 = y_1 + e_1, \quad u_2 = y_2 + e_2.$$

As before, the “extra” values  $u_1$  and  $u_2$  would have to be obtained by some other procedure. However, we proceed to show that this scheme, which has third order truncation errors, does not converge in general.

Let us first consider a case in which  $f_y$  is continuous and

$$\left| \frac{\partial f}{\partial y} \right| \leq K, \quad \tau = h^3 M_4 \geq \max_i |\tau_i|.$$

Then with  $e_i = u_i - y(x_i)$  we obtain, from (35) and (37), in the usual way

$$|e_{i+1}| \leq (\frac{3}{2} + 3hK)|e_i| + 3|e_{i-1}| + \frac{1}{2}|e_{i-2}| + 3h\tau.$$

If we introduce a majorizing set  $\{a_i\}$ , analogous to (29), we find that

$$\begin{aligned} |e_N| &\leq (5 + 3hK)^N \left[ \max(|e_0|, |e_1|, |e_2|) + \frac{3h^4 M_4}{4 + 3hK} \right], \\ &\leq 5^N e^{3K(b-a)/5} \left[ \max(|e_0|, |e_1|, |e_2|) + \frac{3h^4 M_4}{4 + 3hK} \right], \\ &= 5^{(b-a)/h} e^{3K(b-a)/5} \left[ \max(|e_0|, |e_1|, |e_2|) + \frac{3h^4 M_4}{4 + 3hK} \right]. \end{aligned}$$

However, as  $h \rightarrow 0$  this bound becomes infinite. While this does not prove divergence, we strongly suspect it.

To actually show that the third order scheme (37) cannot converge in general, we apply it to the special case where  $f(x, y) \equiv -y$ ; i.e., to the equation

$$\frac{dy}{dx} = -y,$$

whose solution  $y = y_0 e^{-(x-a)}$  satisfies  $y(a) = y_0$ .

Now (37a) can be written as

$$(38) \quad u_{i+1} + (\frac{3}{2} + 3h)u_i - 3u_{i-1} + \frac{1}{2}u_{i-2} = 0, \quad i = 2, 3, \dots$$

This is a linear difference equation with constant coefficients (see Section 4) and it can be solved exactly. That is, we seek a solution of the form

$$u_j = \alpha^j, \quad j = 0, 1, \dots$$

But then from (38),

$$[\alpha^3 + (\frac{3}{2} + 3h)\alpha^2 - 3\alpha + \frac{1}{2}]\alpha^{i-2} = 0, \quad i = 2, 3, \dots$$

Thus, in addition to the trivial solution,  $u_j \equiv 0$ , we have three solutions of the form  $u_j = \alpha_v^j$ , where  $\alpha_v$  for  $v = 1, 2, 3$  are the roots of

$$(39) \quad \alpha^3 + (\frac{3}{2} + 3h)\alpha^2 - 3\alpha + \frac{1}{2} = 0.$$

It is easily verified that, for  $h$  sufficiently small, these three roots are distinct.

It is easy to check that a linear combination

$$(40) \quad u_j = A_1\alpha_1^j + A_2\alpha_2^j + A_3\alpha_3^j,$$

is also a solution. The coefficients  $A_v$  are determined from the assumed known data for  $j = 0, 1, 2$ , by satisfying

$$\begin{aligned} A_1 + A_2 + A_3 &= u_0 \\ A_1\alpha_1 + A_2\alpha_2 + A_3\alpha_3 &= u_1 \\ A_1\alpha_1^2 + A_2\alpha_2^2 + A_3\alpha_3^2 &= u_2. \end{aligned}$$

Since the coefficient determinant is a Vandermonde determinant and the  $\alpha_v$  are distinct, the  $A_v$  are uniquely determined. Then the  $u_j$  for  $j \geq 3$  are also uniquely determined by (40).

Let us write

$$p(\alpha, h) \equiv \alpha^3 + (\frac{3}{2} + 3h)\alpha^2 - 3\alpha + \frac{1}{2},$$

and denote the roots of  $p(\alpha, h) = 0$  by  $\alpha_v(h)$ . Since

$$p(\alpha, 0) \equiv (\alpha - 1)(\alpha^2 + \frac{5}{2}\alpha - \frac{1}{2})$$

we have, with the ordering  $\alpha_1 \leq \alpha_2 \leq \alpha_3$ ,

$$\alpha_1(0) = -\frac{5 + \sqrt{33}}{4}, \quad \alpha_2(0) = \frac{\sqrt{33} - 5}{4}, \quad \alpha_3(0) = 1.$$

For  $h$  sufficiently small,  $|\alpha_v(h) - \alpha_v(0)|$  can be made arbitrarily small and  $\alpha_1(h) < -2$ . So the solution (40), for large  $j$ , behaves like

$$u_j \approx A_1[\alpha_1(h)]^j$$

or in particular for  $x_N = b$ ,

$$u_N \approx A_1[\alpha_1(h)]^{(b-a)/h}.$$

Thus as  $h \rightarrow 0$  the difference solution becomes exponentially unbounded at any point  $x_N = b > a$ . Furthermore, notice that  $\alpha_1(h)$  is negative, and hence  $u_j$  oscillates. This behavior is typical of “unstable” schemes (see

Section 5). Of course, we have assumed here that the initial data is such that  $A_1 \neq 0$ . In fact,

$$A_1 = \frac{\det \begin{vmatrix} u_0 & 1 & 1 \\ u_1 & \alpha_2 & \alpha_3 \\ u_2 & \alpha_2^2 & \alpha_3^2 \\ 1 & 1 & 1 \end{vmatrix}}{\det \begin{vmatrix} \alpha_1 & \alpha_2 & \alpha_3 \\ \alpha_1^2 & \alpha_2^2 & \alpha_3^2 \end{vmatrix}}.$$

Hence if

$$u_1 = u_0 + ch, \quad u_2 = u_0 + dh,$$

then in general, it follows that

$$A_1 = h\beta + \mathcal{O}(h^2), \quad \beta \neq 0.$$

This is based on the fact that  $\alpha_i(h)$  can be developed in the form

$$\alpha_i(h) = \alpha_i(0) + h\alpha'_i(0) + \mathcal{O}(h^2).$$

(For example, in the exceptional case

$$u_1 = \alpha_3 u_0 \quad \text{and} \quad u_2 = \alpha_3^2 u_0,$$

the quantity  $A_1 = 0$ .)

For the actual calculations of the quantities  $U_j$ , the local roundoff error at any net point  $x_j$  will set off such an exponentially growing term. Hence this method is divergent!

## PROBLEMS, SECTION 1

1. If  $f(x, y)$  is independent of  $y$ , i.e., the Lipschitz constant  $K = 0$ , show that the error estimates of Theorem 1 and its corollary are respectively

- (a)  $|e_j| \leq |e_0| + |x_j - x_0|\tau, \quad j \geq 0;$
- (b)  $|e_j| \leq |e_0| + h|x_j - x_0|M_2/2, \quad j \geq 0.$

2. Carry out the proof of the validity of the Richardson extrapolation to zero mesh width for the Euler-Cauchy method defined in (17) with rounding errors  $\rho_0 = \xi_0 h$  and for  $i = 1, 2, \dots$ ,  $|\rho_i| \leq \rho = \mathcal{O}(h^3)$ .

3. Find the coefficient  $\alpha'_i(0)$  in the expansion

$$\alpha_i(h) = \alpha_i(0) + h\alpha'_i(0) + \mathcal{O}(h^2)$$

by formally substituting this expression into (39) and setting the coefficient of  $h$  equal to zero.

## 2. MULTISTEP METHODS BASED ON QUADRATURE FORMULAE

The study of the divergent scheme introduced in (1.37) shows that more accurate approximations of the derivative do not lead to more accurate numerical methods. But a way to determine convergent schemes with an arbitrarily high order of accuracy is *suggested* by converting the original differential equation into an equivalent integral equation. Thus by integrating (0.1a) over the interval  $[a, x]$  and using (0.1b) we obtain

$$(1) \quad y(x) = y_0 + \int_a^x f(\xi, y(\xi)) d\xi.$$

Clearly, any solution of (0.1) satisfies this integral equation and, by differentiation, we find that any solution† of (1) also satisfies (0.1a) and (0.1b). With the subscript notation,  $y_i \equiv y(x_i)$ , the solution of (0.1) or (1) also satisfies

$$(2) \quad y_{i+1} = y_{i-p} + \int_{x_{i-p}}^{x_{i+1}} f(x, y(x)) dx.$$

This is obtained by integrating (0.1a) over  $[x_{i-p}, x_{i+1}]$  for any  $i = 0, 1, \dots$  and any  $p = 0, 1, \dots, i$ . For a given choice of  $p$  a variety of approximations are suggested by applying quadrature formulae to evaluate the integral in (2). The number of schemes suggested in this manner is great, but in practice only relatively few of them are ever used.

We shall limit our study to the case of uniformly spaced net points and interpolatory quadrature formulae. In order to classify these methods in a fairly general manner we distinguish two types of quadrature formulae:

**TYPE A.** *Closed on the right*, i.e., with  $n + 1$  nodes

$$x_{i+1}, x_i, x_{i-1}, \dots, x_{i+1-n};$$

or else

**TYPE B.** *Open on the right*, i.e., with  $n + 1$  nodes

$$x_i, x_{i-1}, \dots, x_{i-n}.$$

The difference equations suggested by these two classes of methods can be written as

$$(3a) \quad u_{i+1} = u_{i-p} + h \sum_{j=0}^n \alpha_j f(x_{i+1-j}, u_{i+1-j});$$

$$(3b) \quad u_{i+1} = u_{i-p} + h \sum_{j=0}^n \beta_j f(x_{i-j}, u_{i-j}).$$

† It is easy to see that any continuous solution  $y(x)$  of (1) is differentiable. This follows since the right-hand side of (1) is differentiable, if  $f(x, y)$  and  $y(x)$  are continuous.

We have denoted the coefficients of the quadrature formulae by  $h\alpha_j$  and  $h\beta_j$  and note that they are independent of  $i$ . If the net spacing were not uniform, the coefficients would depend on  $i$ , in general, and the schemes would require the storage of more than  $n + 1$  coefficients. This is one of the main reasons for the choice of uniform spacing in methods based on quadrature formulae.

When the integers  $p$  and  $n$  are specified, the coefficients are determined as in Subsection 1.2 of Chapter 7. We see from (1.15) of Chapter 7 that the quantities  $\alpha_j$  and  $\beta_j$  are independent of  $h$ , the net spacing. It also follows from Theorem 1.3 of Chapter 7 that the interpolatory quadrature formulae in (3) have the maximum degree of precision possible with the specified nodes; this is reflected in their having the smallest “truncation error,” defined later in equation (8a and b).

In order to compute with a method of type A (closed on the right) we must have available the quantities  $u_i, \dots, u_{i+1-n}$  and  $u_{i-p}$ . Thus the points  $x_{i+1}$  for which (3a) may be used satisfy

$$(4a) \quad i \geq \max(n - 1, p) = t,$$

provided that  $u_0, u_1, \dots, u_t$  are given. Similarly, method B requires  $u_i, \dots, u_{i-n}$  and  $u_{i-p}$ , so that the points  $x_{i+1}$  for which (3b) can be used satisfy

$$(4b) \quad i \geq \max(n, p) = r,$$

provided that  $u_0, u_1, \dots, u_r$  are given. Special procedures are required to obtain these starting values in either case (see Section 3).

The fundamental difference between the open and closed methods is the ease with which the difference equations can be solved. There is no difficulty in solving the equations based on the open formulae. In fact, since formula (3b) is an explicit expression for  $u_{i+1}$  these are called *explicit methods*. But, the closed formulae define *implicit methods* since the equation for the determination of  $u_{i+1}$  is implicit.† That is, (3a) is of the form

$$(5a) \quad u_{i+1} = g_i(u_{i+1}),$$

where

$$g_i(z) \equiv c_i + h\alpha_0 f(x_{i+1}, z),$$

with

$$c_i \equiv u_{i-p} + h \sum_{j=1}^n \alpha_j f(x_{i+1-j}, u_{i+1-j})$$

† In the special case that  $f(x, y)$  is linear in  $y$ , i.e.,  $f(x, y) \equiv a(x)y + b(x)$ , the implicit equation (3a) is easily solved explicitly for  $u_{i+1}$  if  $1 - h\alpha_0 a(x_{i+1}) \neq 0$ .

and so

$$(5b) \quad \frac{dg_i(z)}{dz} = h\alpha_0 \frac{\partial f(x_{i+1}, z)}{\partial z}.$$

Clearly the method of functional iteration is natural for solving (5a). By Theorem 1.2 of Chapter 3 we know that, if a “sufficiently close” initial estimate  $u_{i+1}^{(0)}$  of the root is given, the iterations

$$u_{i+1}^{(\nu+1)} = g_i(u_{i+1}^{(\nu)}), \quad \nu = 0, 1, \dots,$$

will converge provided  $h$  is sufficiently small, e.g.,

$$(6) \quad h < \frac{1}{|\alpha_0 K|}, \quad K \equiv \max \left| \frac{\partial f(x, y)}{\partial y} \right|.$$

On the other hand, we may apply Theorem 1.1 of Chapter 3 to show existence of a unique root in the interval  $[c_i - \rho, c_i + \rho]$ . That is, by selecting  $u_{i+1}^{(0)} = c_i$ , we have

$$|u_{i+1}^{(1)} - u_{i+1}^{(0)}| \leq h\alpha_0 M$$

where  $M \equiv \max |f(x, y)|$ . Hence for any

$$\rho \geq \frac{h\alpha_0 M}{1 - h\alpha_0 K}$$

we have an interval in which a unique root of (5) exists. We shall see that it is not necessary to find the root of (5) in order to preserve the high accuracy of the method. In fact, the *predictor-corrector* technique, which we next study, uses only one iteration of (5) without a loss in accuracy.

The *predictor-corrector method* is defined by

$$(7a) \quad u_{i+1}^* = u_{i-q} + h \sum_{j=0}^m \beta_j f(x_{i-j}, u_{i-j});$$

$$(7b) \quad u_{i+1} = u_{i-p} + h \sum_{k=1}^n \alpha_k f(x_{i+1-k}, u_{i+1-k}) + h\alpha_0 f(x_{i+1}, u_{i+1}^*);$$

$$(7c) \quad i \geq s \equiv \max [p, q, m, n - 1].$$

Here, an  $m + 1$  point quadrature formula open on the right has been used [in the *predictor* (7a)] to approximate an integral over  $[x_{i-q}, x_{i+1}]$ . The closed formula (7b) (called the *corrector*) is similar to that of (3a) but  $u_{i+1}^*$  has been used in place of  $u_{i+1}$  in the right-hand side. Thus as previously indicated, the corrector is the first iteration step in solving the implicit equation (5) with the initial guess furnished by the predictor. Hence *only two evaluations* of the function  $f(x, y)$  are required for each step of the predictor-corrector method; i.e.,  $f(x_i, u_i)$  and  $f(x_{i+1}, u_{i+1}^*)$ .

The procedure (7) can only be employed after the values  $u_0, u_1, \dots, u_s$  have been determined. Here  $s$ , defined by (7c), is determined from the open and closed formulae (7a and b). To compute the  $u_i$ ,  $i < s$ , we refer to the procedures of Section 3.

It is clear, from the analysis given in the remainder of this section, that the predictor-corrector method (7) has the same order of accuracy as the implicit method of type A defined by the corrector (3a), provided that the explicit predictor is sufficiently accurate. In other words, we avoid the necessity of repeatedly iterating the corrector, as in (5), by using a good initial approximation. We shall first develop estimates of the error in solving (1) by the predictor-corrector method (7). Next, we shall show how to modify the error estimates to cover the case of finite precision arithmetic, i.e., with rounding errors. Finally, we indicate briefly how the error estimates may be derived for the methods (3a) or (3b).

The predictor-corrector method has the advantage of permitting the detection of an isolated numerical error through the comparison of  $u_{i+1}^*$  with  $u_{i+1}$  (or  $U_{i+1}^*$  with  $U_{i+1}$  defined later).

The truncation error of (7) is obtained as follows. We define  $\sigma_{i+1}^*$  and  $\sigma_{i+1}$  in terms of the exact solution  $y = y(x)$  by

$$(8a) \quad y_{i+1} = y_{i-p} + h \sum_{j=0}^m \beta_j f(x_{i-j}, y_{i-j}) + h\sigma_{i+1}^*;$$

$$(8b) \quad y_{i+1} = y_{i-p} + h \sum_{k=0}^n \alpha_k f(x_{i+1-k}, y_{i+1-k}) + h\sigma_{i+1}.$$

Then with the definition

$$(8c) \quad y_{i+1}^* \equiv y_{i+1} - h\sigma_{i+1}^*,$$

the *local truncation error*,  $\tau_{i+1}$ , of the predictor-corrector method (7) is defined by

$$(9) \quad y_{i+1} = y_{i-p} + h \sum_{k=1}^n \alpha_k f(x_{i+1-k}, y_{i+1-k}) \\ + h\alpha_0 f(x_{i+1}, y_{i+1}^*) + h\tau_{i+1}, \quad i \geq s.$$

To obtain a more explicit expression for this error we subtract (9) from (8b) and use (8c) to get

$$(10) \quad \begin{aligned} \tau_{i+1} &= \sigma_{i+1} + \alpha_0 [f(x_{i+1}, y_{i+1}) - f(x_{i+1}, y_{i+1} - h\sigma_{i+1}^*)] \\ &= \sigma_{i+1} + h\sigma_{i+1}^* \alpha_0 \frac{\partial f}{\partial y}. \end{aligned}$$

Here we have assumed  $f_y$  to be continuous in  $y$  and used the mean value theorem;  $\bar{f}_y$  is a value of  $f_y$  at some intermediate point in the obvious

interval. The quantities  $h\sigma_{i+1}^*$  and  $h\sigma_i$  are the errors in the  $(m + 1)$ - and  $(n + 1)$ -point quadrature formulae, (8a and b). Explicit expressions for these quadrature errors have been obtained for various cases in Section 1 of Chapter 7. It should be noted from (10) that the order as  $h \rightarrow 0$  of the truncation error in the predictor-corrector method is the same as the order for the corresponding closed formula used alone, provided that the order of  $h\sigma^*$  is not less than the order of  $\sigma$ . Table 1 has a brief listing of commonly used predictor-corrector schemes.

**Table 1****Table 1 Some Common Predictor-Corrector Methods**

Associated Name	$m, q; n, p$	Weights					$\sigma^*$	$\sigma$
		$j =$	0	1	2	3		
Modified Euler	0, 0; 1, 0	$\beta_j \equiv 1$					$\frac{1}{2}hy^{(2)}$	$-\frac{1}{12}h^2y^{(3)}$
		$\alpha_j \equiv \frac{1}{2}$		$\frac{1}{2}$				
Milne's Method (3 points)	2, 3; 2, 1	$\beta_j \equiv \frac{8}{3}$	$-\frac{4}{3}$	$\frac{8}{3}$			$\frac{28}{90}h^4y^{(5)}$	$-\frac{1}{90}h^4y^{(5)}$
		$\alpha_j \equiv \frac{1}{3}$	$\frac{4}{3}$	$\frac{1}{3}$				
Improved Adams, or Moulton's Method (4 points)	3, 0; 3, 0	$\beta_j \equiv \frac{55}{24}$	$-\frac{59}{24}$	$\frac{37}{24}$	$-\frac{9}{24}$		$\frac{251}{720}h^4y^{(5)}$	$-\frac{19}{720}h^4y^{(5)}$
		$\alpha_j \equiv \frac{9}{24}$	$\frac{19}{24}$	$-\frac{5}{24}$	$\frac{1}{24}$			
Milne's Method (5 points)	4, 5; 4, 3	$\beta_j \equiv \frac{33}{10}$	$-\frac{21}{5}$	$\frac{39}{5}$	$-\frac{21}{5}$	$\frac{33}{10}$	$\frac{41}{140}h^6y^{(7)}$	$-\frac{8}{945}h^6y^{(7)}$
		$\alpha_j \equiv \frac{14}{45}$	$\frac{64}{45}$	$\frac{24}{45}$	$\frac{64}{45}$	$\frac{14}{45}$		

## 2.1. Error Estimates in Predictor-Corrector Methods

To examine convergence of the scheme (7) we introduce the *errors*

$$(11) \quad e_i \equiv u_i - y_i, \quad e_i^* \equiv u_i^* - y_i^*.$$

Then subtraction of (9) from (7) yields, if  $f_y$  is continuous,

$$e_{i+1} = e_{i-p} + h \sum_{k=1}^n \alpha_k g_{i+1-k} e_{i+1-k} + h \alpha_0 g_{i+1} e_{i+1}^* - h \tau_{i+1}.$$

Here we have used the mean value theorem to introduce

$$g_j \equiv \frac{\partial f}{\partial y} (x_j, \bar{y}_j), \quad \bar{y}_j \in (y_j, u_j) \quad \text{for } j \neq i + 1, \quad \text{while}$$

$$\bar{y}_{i+1} \in (y_{i+1}^*, u_{i+1}^*).$$

However, from (7a) and (8a and c) we obtain

$$e_{i+1}^* = e_{i-q} + h \sum_{j=0}^m \beta_j g_{i-j} e_{i-j},$$

and using this in the above implies finally

$$(12) \quad e_{i+1} = e_{i-p} + h \sum_{k=1}^n \alpha_k g_{i+1-k} e_{i+1-k} + h \alpha_0 g_{i+1} e_{i-q} \\ + h^2 \alpha_0 g_{i+1} \sum_{j=0}^m \beta_j g_{i-j} e_{i-j} - h \tau_{i+1}, \quad i \geq s.$$

To estimate these errors we introduce

$$(13) \quad K \equiv \max \left| \frac{\partial f}{\partial y} \right|; \quad \tau \equiv \max |\tau_j|; \\ A \equiv \sum_{k=0}^n |\alpha_k|; \quad B \equiv \sum_{j=0}^m |\beta_j|.$$

Then by taking the absolute value of both sides of (12), we have

$$(14) \quad |e_{i+1}| \leq |e_{i-p}| + hK(|\alpha_0| |e_{i-q}| + \sum_{k=1}^n |\alpha_k| |e_{i+1-k}|) \\ + h^2 K^2 |\alpha_0| \sum_{j=0}^m |\beta_j| |e_{i-j}| + h|\tau_{i+1}|; \quad i \geq s.$$

We again introduce a comparison or majorizing set,  $\{a_i\}$ , defined by

$$(15a) \quad a_0 \equiv \max (|e_0|, |e_1|, \dots, |e_s|),$$

$$(15b) \quad a_{i+1} = (1 + hKA + h^2 K^2 |\alpha_0| B) a_i + h\tau;$$

and claim that

$$|e_j| \leq a_j; \quad j = 0, 1, \dots, N.$$

The proof of this inequality is easily given by induction. From (15),  $\{a_i\}$  is a non-decreasing sequence. Therefore, (15a) establishes the inequality

for  $j \leq s$ . Now assume the inequality holds for all  $j \leq i$  where  $i \geq s$ . Then (14) implies

$$\begin{aligned}
|e_{i+1}| &\leq a_{i-p} + hK \left( |\alpha_0| a_{i-q} + \sum_{k=1}^n |\alpha_k| a_{i+1-k} \right) \\
&\quad + h^2 K^2 |\alpha_0| \sum_{j=0}^m |\beta_j| a_{i-j} + h\tau, \\
&\leq \left[ 1 + hK \left( |\alpha_0| + \sum_{k=1}^n |\alpha_k| \right) + h^2 K^2 |\alpha_0| \sum_{j=0}^m |\beta_j| \right] a_i + h\tau \\
&= (1 + hKA + h^2 K^2 |\alpha_0| B) a_i + h\tau \\
&= a_{i+1}.
\end{aligned}$$

Note that from the recursive expression for  $e_{i+1}^*$ , we have the single estimate  $|e_{i+1}^*| \leq (1 + hBK)|a_i|$ . The application of (15b) recursively, in the by now familiar manner, yields the final result which can be summarized as

**THEOREM 1.** *Let the predictor-corrector method (7) be applied to solve (0.1) or (1) in  $a \leq x \leq b$  with “initial” values,  $u_i$ , satisfying*

$$(16a) \quad |u_i - y(x_i)| \leq a_0, \quad i = 0, 1, \dots, s.$$

*Let  $f_y(x, y)$  be continuous and bounded in  $S$ :  $\{(x, y) \mid a \leq x \leq b; |y| < \infty\}$ . Then with the definitions (9) and (13) the errors in the numerical solution satisfy, for  $a \leq x_j \leq b$ ,*

$$\begin{aligned}
(16b) \quad |u_j - y(x_j)| &\leq \left[ a_0 + \frac{\tau}{K(A + hK|\alpha_0|B)} \right] \\
&\quad \times \exp [(x_j - a)K(A + hK|\alpha_0|B)],
\end{aligned}$$

$$\begin{aligned}
(16c) \quad |u_j^* - y^*(x_j)| &\leq \left[ a_0 + \frac{\tau}{K(A + hK|\alpha_0|B)} \right] \\
&\quad \times \exp [(x_j - a)K(A + hK|\alpha_0|B) + hBK]. \quad \blacksquare
\end{aligned}$$

From this theorem it follows that  $u_j$  converges to  $y(x_j)$  as  $h \rightarrow 0$  if  $a_0 \rightarrow 0$  and  $\tau \rightarrow 0$ . The order in  $h$  of the estimate (16) is the minimum of the orders in  $h$  of  $a_0$  and  $\tau$ . We say that the methods for selecting the initial data and method (7) are *balanced* if  $a_0$  and  $\tau$  vanish to the same order in  $h$ . If they do not, then some “extra accuracy” has been wasted.

If the exact solution,  $y(x)$ , has sufficiently many continuous derivatives,<sup>†</sup> then the local truncation error,  $\tau$ , can be simply expressed by using the

<sup>†</sup> If all partial derivatives of order  $p$  of  $f(x, y)$  are continuous, then  $y(x)$  has a continuous derivative of order  $p + 1$ . This results by differentiating (1) often enough.

methods of Section 1, Chapter 7. For instance, from the crude estimates of the form (1.8) of Chapter 7 and (10) we find

$$\tau = \mathcal{O}(h^{m+2}) + \mathcal{O}(h^{n+1}).$$

From these estimates, we see that the predictor and corrector formulae are balanced if  $m + 1 = n$ . The method is said to be of order  $t \equiv \min(m + 2, n + 1)$ . [In the special case  $p = n - 1, q = m + 1$ , with  $m$  and  $n$  both even,

$$\tau = \mathcal{O}(h^{m+3}) + \mathcal{O}(h^{n+2}),$$

which results from the estimate of error in the Newton-Cotes formulae (1.11) of Chapter 7. Parity makes  $m = n$  optimal.]

Of course, roundoff errors are committed when the formulae in (7) are evaluated. If we call  $U_j$  and  $U_j^*$  the numbers actually obtained in these evaluations, then we can write

$$(17a) \quad U_{i+1}^* = U_{i-q} + h \sum_{j=0}^m \beta_j f(x_{i-j}, U_{i-j}) + \rho_{i+1}^*,$$

$$(17b) \quad U_{i+1} = U_{i-p} + h \sum_{k=1}^n \alpha_k f(x_{i+1-k}, U_{i+1-k}) \\ + h\alpha_0 f(x_{i+1}, U_{i+1}^*) + \rho_{i+1}, \quad i \geq s.$$

Here  $\rho_j^*$  and  $\rho_j$  are the roundoff errors introduced into each of the indicated computations. Now we define the errors

$$E_j \equiv U_j - y(x_j), \quad E_j^* \equiv U_j^* - y_j^*;$$

and obtain from (8), (9), and (17), as in the derivation of (12)

$$(18) \quad E_{i+1} = E_{i-p} + h \sum_{k=1}^n \alpha_k g_{i+1-k} E_{i+1-k} + h\alpha_0 g_{i+1} E_{i-q} \\ + h^2 \alpha_0 g_{i+1} \sum_{j=0}^m \beta_j g_{i-j} E_{i-j} - h\tau_{i+1} + \rho_{i+1} \\ + h\alpha_0 g_{i+1} \rho_{i+1}^*, \quad i \geq s,$$

with

$$g_j = f_y(x_j, \bar{y}_j) \quad \text{and} \quad \bar{y}_j \in (y_j, U_j) \quad \text{for } j \neq i + 1,$$

while

$$\bar{y}_{i+1} \in (y_{i+1}^*, U_{i+1}^*).$$

By applying the previous method of analysis to this system we find the total error bound in

**THEOREM 2.** *Under the hypothesis of Theorem 1, with the notation (13) and (17),*

$$(19) \quad |U_j - y(x_j)| \leq \left[ b_0 + \frac{\tau + |\alpha_0|K\rho^* + (1/h)\rho}{K(A + hK|\alpha_0|B)} \right] \times \exp [(x_j - a)K(A + hK|\alpha_0|B)],$$

$$a \leq x_j \leq b,$$

where

$$(20) \quad \begin{aligned} \rho &= \max_{s < j < N} |\rho_j|, & \rho^* &= \max_{s < j < N} |\rho_j^*|, \\ b_0 &= \max_{0 < j < s} |U_j - y(x_j)|. \end{aligned} \quad \blacksquare$$

It is of interest to note that in the bound (19) the corrector roundoff enters in the form  $\rho/h$  while that from the predictor has a coefficient independent of  $h$ . However, it is unlikely that any special measures in actual computation could be adopted to balance these different orders in  $h$  of the roundoff. If in Theorem 2 we know that

$$b_0 = \mathcal{O}(h), \quad \rho^* = \mathcal{O}(h), \quad \tau = \mathcal{O}(h), \quad \text{and} \quad \rho = \mathcal{O}(h^2),$$

then (19) yields  $|E_j| \leq Vh$  for  $a \leq x_j \leq b$  for a constant  $V$  independent of  $h$ .

It should be observed that if  $\alpha_0 \equiv 0$  in (7), the predictor is never used. The corrector in this case is an open formula, and the above error analysis then applies to the method based on the use of a *single open formula*. The corresponding result for the method based on the use of a *single closed formula* (i.e., the *implicit method*) is obtained by a slight modification of the above technique (see Problem 1). Now if, in the predictor-corrector method, more than one iteration is employed, the estimates (16) and (19) no longer apply. But a comparison of the error bounds of the predictor-corrector method and the corresponding implicit corrector method shows that there is no great gain to be expected in using the corrector more than once, provided  $h\sigma_{i+1}^*$  and  $\sigma_{i+1}$  are of the same order in  $h$ .

We can remark further that in Theorem 2 the requirement that  $f(x, y)$  and  $f_y(x, y)$  be bounded and continuous for  $|y| < \infty$  can be replaced by the milder restriction that  $f(x, y)$  and  $f_y(x, y)$  be bounded and continuous in the strip

$$S': \{(x, y) \mid a \leq x \leq b, |y - y(x)| \leq d\} \quad \text{for some } d > 0,$$

provided that:

- (a)  $h$  is sufficiently small,
- (b)  $b_0 = \mathcal{O}(h)$ ,  $\rho^* = \mathcal{O}(h)$ ,  $\tau = \mathcal{O}(h)$  and,
- (c)  $\rho = \mathcal{O}(h^2)$ .

To show that the estimate (19) holds, we now could show inductively that the values  $U_{i+1}^*$  and  $U_{i+1}$  exist and are in the strip  $S'$ , and therefore (19) is satisfied for  $j = i + 1$ . The constant  $K \equiv \max_{S'} |f_y|$  replaces the previous definition of  $K$  in (19).

## 2.2. Change of Net Spacing

During the course of a computation based on a predictor-corrector method, we should keep track of the “measure of error,”

$$U_{i+1} - U_{i+1}^* \equiv \eta_{i+1}.$$

That is,

$$(20) \quad |\eta_{i+1}| = |U_{i+1} - U_{i+1}^*| = |E_{i+1} - E_{i+1}^* + h\sigma_{i+1}^*| \\ \leq |E_{i+1}| + |E_{i+1}^*| + h|\sigma_{i+1}^*|.$$

Hence if  $|\eta_{i+1}|$  is large, we know that the actual error is probably large.

An isolated mistake in computation may be responsible for a large  $|\eta_{i+1}|$ . But, if the computation is correct, then the obvious way to reduce  $|\eta_{i+1}|$  is to reduce the interval size  $h$ . In practice, this is usually done by successively halving  $h$ . Alternatively if the estimate (20) becomes very small,  $h$  may be increased, say, by doubling it.

Doubling the interval size offers no difficulty if at least  $2s$  points have been computed with the net spacing  $h$ . We merely discard the data at every other net point, replace  $h$  by  $2h$ , and continue the calculations.

On the other hand, in order to halve  $h$ , we require data at  $s/2$  new intermediate points, say

$$x_i - h/2, x_{i-1} - h/2, \dots, x_{i+1-(s/2)} - h/2.$$

These values can be determined by the application of an interpolation procedure which uses the known data at appropriate net points  $x_j = x_0 + jh$ . However, the accuracy of the interpolation formula must be consistent with that of the predictor-corrector formula being used. That is, *the interpolation error must be at least of the same order in  $h$  as the truncation error,  $\tau$* , given in (10). Otherwise, as Theorem 1 indicates, the accuracy of the numerical solution will not be greater than that determined by the interpolation. The caution required in order to reduce the net spacing without sacrificing accuracy is one of the disadvantages of predictor-corrector methods when compared to single-step methods of the next section. Sometimes it is feasible to actually restart the integration at  $x_i$ , by using the method employed at  $x_0$ , but with the net size  $h/2$ .

## PROBLEMS, SECTION 2

1. Verify the entries in Table I labeled Modified Euler and Milne's method (3 points).
2. Verify the entries in Table I labeled Improved Adams and Milne's Method (5 points).
3. For each of the methods of Table I, with what interval size  $h$ , and how many decimal places should the equation

$$y' = y \quad y(0) = 1$$

be solved in  $0 \leq x \leq 5$  in order that the error satisfy

$$|E_t| \equiv |U_t - y_t| \leq 10^{-4}?$$

4. Assume that  $f(x, y)$  and  $f_y(x, y)$  are bounded and continuous in  $S$ :

$$\{(x, y) \mid a \leq x \leq b, |y| < \infty\}.$$

Then if  $h < 1/|\alpha_0 K|$  where  $K \equiv \max |f_y|$ , the implicit scheme (3a) can be used to find the  $\{u_i\}$ , given  $u_0, u_1, \dots, u_r$ , with  $r = \max(n, p)$ . [See discussion after equation (6).] Estimate the total error,  $E_t \equiv U_t - y_t$ , in solving (1) by the implicit scheme based on (3a).

[Hint: If we stop the iterative process described in (5), when the equation (3a) is satisfied with the error  $\rho_{t+1}$ , we will obtain a sequence  $\{U_t\}$  with  $U_0 = u_0$ ,  $U_1 = u_1, \dots, U_r = u_r$ , that satisfies

$$U_{t+1} = U_{t-p} + h \sum_{j=0}^n \alpha_j f(x_{t+1-j}, U_{t+1-j}) + \rho_{t+1}.$$

Equation (8b) defines the corresponding truncation error  $\sigma_{t+1}$ . Show that  $E_t$  satisfies

$$\begin{aligned} E_{t+1} &= E_{t-p} + h \sum_{j=1}^n \alpha_{t+1-j} g_{t+1-j} E_{t+1-j} \\ &\quad + h \alpha_0 g_{t+1} E_{t+1} + \rho_{t+1} - h \sigma_{t+1}, \end{aligned}$$

where  $g_j = f_y(x_j, \bar{y}_j)$  at some suitable point  $\bar{y}_j$ . Hence show that  $|E_t| \leq a_t$ , where the sequence  $\{a_t\}$  is defined by

$$\begin{aligned} a_0 &= \max(|E_0|, \dots, |E_r|) \\ a_{t+1} &= a_t \left( \frac{1 + hAK}{1 - h|\alpha_0|K} \right) + \left( \frac{\rho + h\sigma}{1 - h|\alpha_0|K} \right) \end{aligned}$$

where  $\rho = \max |\rho_t|$ ,  $\sigma = \max |\sigma_t|$ . Then show that

$$a_{t+1} = a_t(1 + hQ) + R$$

where

$$Q = \frac{K(A + |\alpha_0|)}{1 - h|\alpha_0|K}$$

$$R = \frac{\rho + h\sigma}{1 - h|\alpha_0|K}.$$

### 3. ONE-STEP METHODS

The higher accuracy predictor-corrector methods of Section 2 all require special procedures for starting the calculations. That is, some approximate solution,  $u_j$ , must first be computed for  $j = 0, 1, \dots, s$ . In addition, if the interval size,  $h$ , is reduced during the course of the calculation, care must be taken to preserve the accuracy of the method. The single-step methods, which we now consider, require none of these special measures. In fact, they can be used to determine the starting values and to change the net spacing in other methods. The price paid for these advantages is, in general, the requirement of a greater number of evaluations of the function  $f(x, y)$  (or functions related to it) for each step in the solution.

Again we consider the initial value problem

$$(1) \quad \frac{dy}{dx} = f(x, y), \quad y(a) = y_0.$$

By *single-step* we mean that only data at  $x = x_0$  are to be employed in obtaining the approximation to  $y(x)$  at  $x = x_1$ . Obviously such a procedure could then be employed at  $x_1$ , and so forth, to extend the solution with arbitrary step sizes. However, for convenience in exposition we shall consider calculations on a uniform net

$$x_j = a + jh, \quad h = \frac{b - a}{N}.$$

Any single-step method for approximating the solution of (1) in  $[a, b]$  can be indicated by the general form

$$(2a) \quad u_0 = y_0 + e_0,$$

$$(2b) \quad u_{j+1} = u_j + hF\{h, x_j, u_j; f\}, \quad j = 0, 1, \dots, N - 1.$$

Here we denote by  $F\{h, x_j, u_j; f\}$  some quantity whose value is uniquely determined by the value of  $(h, x_j, u_j)$  and the function  $f$ . For example, the Euler-Cauchy scheme in (1.1) is a single-step method in which  $F\{h, x, u; f\} \equiv f(x, u)$ . We shall see that a variety of different choices for  $F$  is determined by using Taylor's theorem or quadrature formulae.

It is a simple matter to obtain estimates for the error in a very general class of single-step methods. To do this we first define the *local truncation errors*,  $\tau_{j+1}$ , by writing

$$(3) \quad y(x_{j+1}) = y(x_j) + hF\{h, x_j, y(x_j); f\} + h\tau_{j+1}, \\ j = 0, 1, \dots, N - 1;$$

where  $y(x)$  is the solution of (1). The largest integer  $p$  such that  $|\tau_j| = \mathcal{O}(h^p)$  is called the order† of the method. As usual, the errors in the numerical solution are defined by

$$(4) \quad e_j \equiv u_j - y(x_j), \quad j = 0, 1, \dots, N.$$

Now in analogy with Theorem 1.1 we have

**THEOREM 1.** *Let  $u_j$  be the numerical solution defined in (2) where  $F\{h, x, u; f\}$  satisfies*

$$(5) \quad |F\{h, x, u; f\} - F\{h, x, v; f\}| \leq K|u - v|$$

*for all  $(x, u)$  and  $(x, v)$  in the strip  $S$ :  $\{(x, y) \mid a \leq x \leq b, |y| < \infty\}$ . Then if  $y(x)$  is the solution of (1), and  $\{\tau_j\}$  is defined by (3),*

$$(6) \quad |u_j - y(x_j)| \leq e^{K(x_j - a)} \left( |e_0| + \frac{\tau}{K} \right), \quad j = 0, 1, \dots, N;$$

where

$$\tau \equiv \max_j |\tau_j|.$$

*Proof.* Subtract (3) from (2b) and use (5) to find

$$\begin{aligned} |e_{j+1}| &= |e_j + h[F\{h, x_j, u_j; f\} - F\{h, x_j, y(x_j); f\}] - h\tau_{j+1}| \\ &\leq (1 + hK)|e_j| + h\tau, \quad j = 0, 1, \dots, N - 1. \end{aligned}$$

The remainder of the proof follows exactly as in Theorem 1.1. ■

If the function  $f(x, y)$  and the scheme, determined by  $F\{h, x, u; f\}$ , have special smoothness properties it may be possible to replace (5) by a type of mean value *equality*, say,

$$(7) \quad F\{h, x, u; f\} - F\{h, x, v; f\} = G\{h, x, u, v; f\}(u - v).$$

Here  $G$  is determined by the value of  $(h, x, u, v)$  and the function  $f$ . Again in the Euler-Cauchy scheme if  $\partial f / \partial y$  is continuous, then (7) holds with

$$G\{h, x, u, v; f\} = \frac{\partial f(x, \theta u + (1 - \theta)v)}{\partial y}, \quad \text{for some } \theta \text{ in } 0 < \theta < 1.$$

When this mean value property is satisfied, we can prove exact analogs of Theorem 1.2 and its corollaries.

The roundoff errors in single-step methods can be treated very much as

† We assume that  $f(x, y)$  has enough continuous derivatives so that  $p$  may be determined by a Taylor's series expansion of  $y(x_j + h) - y(x_j) - hF\{h, x_j, y(x_j); f\}$  in powers of  $h$ .

in Subsection 1.2. Thus the numbers actually obtained, say  $\{U_j\}$ , in trying to evaluate the set  $\{u_j\}$  from (2) will, in general, have errors due to the finite precision arithmetic. These numbers will satisfy equations of the form

$$(8a) \quad U_0 = y_0 + \rho_0$$

$$(8b) \quad U_{j+1} = U_j + hF\{h, x_j, U_j; f\} + \rho_{j+1}, \quad j = 0, 1, \dots, N-1.$$

Then, if (5) is satisfied, we deduce in an obvious manner

$$(9) \quad |U_j - y(x_j)| \leq e^{K(x_j - a)} \left( |\rho_0| + \frac{\tau + \frac{\rho}{h}}{K} \right) \quad j = 0, 1, \dots, N,$$

where  $\rho = \max_{1 \leq j} |\rho_j|$ . Again we see that as  $h \rightarrow 0$ , while  $x_j - a = jh \equiv c$  is fixed, the roundoff error may become arbitrarily large if the computing accuracy remains unchanged. This effect is due to the fact that infinitely many computations are required to get to the finite point  $x = c$  as  $h \rightarrow 0$ . If the single-step scheme is of order  $p$ , then  $\tau$  can be bounded by a term of the form  $Mh^p$ . For numerical balance then,  $|\rho_0| = \mathcal{O}(h^p)$  and  $\rho = \mathcal{O}(h^{p+1})$  are reasonable requirements for the magnitude of the rounding error.

### 3.1. Finite Taylor's Series

If the solution  $y(x)$ , of (1) has continuous derivatives of order  $r + 1$  in  $[a, b]$ , then by Taylor's theorem:

$$(10a) \quad \begin{aligned} y(x_{j+1}) &= y(x_j) + hy^{(1)}(x_j) + \cdots + \frac{h^r}{r!} y^{(r)}(x_j) \\ &\quad + \frac{h^{r+1}}{(r+1)!} y^{(r+1)}(x_j + \theta_j h), \\ 0 < \theta_j &< 1; \quad j = 0, 1, \dots, N-1. \end{aligned}$$

From the differential equation it follows that the higher order derivatives of  $y(x)$  can be expressed as

$$(10b) \quad \begin{aligned} y^{(1)}(x) &= f(x, y), \\ y^{(2)}(x) &= f_x(x, y) + f_y(x, y)y^{(1)}(x), \\ y^{(3)}(x) &= f_{xx}(x, y) + 2f_{xy}(x, y)y^{(1)}(x) \\ &\quad + f_{yy}(x, y)[y^{(1)}(x)]^2 + f_y(x, y)y^{(2)}(x), \\ &\quad \vdots \end{aligned}$$

or in general,

$$(10b') \quad y^{(\nu)}(x) = \frac{d^{\nu-1}}{dx^{\nu-1}} f(x, y(x)); \quad \nu = 1, 2, \dots$$

Thus given the value of  $y(x)$  at a point, we may determine its derivatives if we can evaluate the partial derivatives of  $f(x, y)$ . We use these observations in the *finite Taylor's series method* for approximating solutions of (1).

Equation (10a) suggests the scheme

$$(11a) \quad u_{j+1} = u_j + hu_j^{(1)} + \cdots + \frac{h^r}{r!} u_j^{(r)}, \quad j = 0, 1, \dots, N - 1,$$

where in analogy with (10b) we have defined, for given  $(x_j, u_j)$ ,

$$(11b) \quad \begin{aligned} u_j^{(1)} &= f(x_j, u_j), \\ u_j^{(2)} &= f_x(x_j, u_j) + f_y(x_j, u_j)u_j^{(1)}, \\ u_j^{(3)} &= f_{xx}(x_j, u_j) + 2f_{xy}(x_j, u_j)u_j^{(1)} \\ &\quad + f_{yy}(x_j, u_j)(u_j^{(1)})^2 + f_y(x_j, u_j)u_j^{(2)}, \\ &\vdots \end{aligned}$$

These formulae are easily deduced from the compact symbolic formula obtained from (10b')

$$(11b') \quad u_j^{(v)} = \left[ \left( \frac{\partial}{\partial x} + f(x, y) \frac{\partial}{\partial y} \right)^{v-1} f(x, y) \right]_{x=x_j, y=u_j}.$$

The initial value is, allowing for an error in obtaining  $y_0$ ,

$$(11c) \quad u_0 = y_0 + e_0.$$

The formulation of the method is complete and the approximation  $\{u_i\}$  can be computed by recursive application of (11a) through (11c).

To write the Taylor's series method in the form (2) we need only define the operator

$$(12) \quad F\{h, x_j, u_j; f\} \equiv u_j^{(1)} + \frac{h}{2!} u_j^{(2)} + \cdots + \frac{h^{r-1}}{r!} u_j^{(r)}$$

where the  $u_j^{(v)}$  are defined in (11). Then from the expansion (10a) and the definition (3) of the truncation error for a one-step method, we obtain

$$(13) \quad \tau_{j+1} = \frac{h^r}{(r+1)!} y^{(r+1)}(x_j + \theta_j h),$$

$$0 < \theta_j < 1, \quad j = 0, 1, \dots, N - 1.$$

The method is thus of order  $r$  when the first  $r + 1$  terms in the Taylor expansion are used. To verify condition (5) we use Taylor's theorem in (11b), after eliminating all  $u^{(v)}$  and  $v^{(v)}$ , to get

$$\begin{aligned} u^{(1)} - v^{(1)} &= (u - v)[f_y] \\ u^{(2)} - v^{(2)} &= (u - v)[f_{xy} + f_y^2 + f_{yy}f] \\ &\vdots \end{aligned}$$

The arguments of  $f(x, y)$  and its derivatives which occur in the brackets are all of the form  $(x, \theta u + (1 - \theta)v)$  with different values of  $\theta$ , in  $0 < \theta < 1$ , in different brackets. From the representation (11b') we find in a straightforward manner that

$$(14) \quad u^{(v)} - v^{(v)} = (u - v) \left[ \frac{\partial}{\partial y} \left\{ \left( \frac{\partial}{\partial x} + f(x, y) \frac{\partial}{\partial y} \right)^{v-1} f(x, y) \right\} \right] \Big|_{y=\theta_v u + (1-\theta_v)v} \\ v = 1, 2, \dots$$

Hence we can conclude that if  $f(x, y)$  has sufficiently many continuous and bounded partial derivatives for  $(x, y) \in S$  then  $F\{h, x, u; f\}$  defined in (12) satisfies (7). Thus (5) is also satisfied, say, for all  $h \leq h_0$ , where  $h_0$  is some fixed spacing. The constant  $K$  entering into (5) can be written in the form

$$(15) \quad K = M_1 + \frac{h_0}{2!} M_2 + \dots + \frac{h_0^{r-1}}{r!} M_r,$$

where  $M_k$  is a bound on the appropriate bracket in the  $k$ th equation in (14); i.e.,

$$M_k \equiv \sup_s \left| \frac{\partial}{\partial y} \left( \frac{\partial}{\partial x} + f(x, y) \frac{\partial}{\partial y} \right)^{k-1} f(x, y) \right|, \quad k = 1, 2, \dots, r.$$

By applying Theorem 1 and equation (13), we find that for all  $h \leq h_0$

$$(16) \quad |u_j - y(x_j)| \leq e^{K(x_j - a)} \left[ |e_0| + \frac{M_{r+1} h^r}{(r+1)! K} \right], \quad j = 0, 1, \dots, N,$$

where

$$M_{r+1} \equiv \sup_{[a, b]} |y^{(r+1)}(x)|.$$

If we neglect the initial error, i.e., set  $e_0 = 0$ , then the error is at most  $\mathcal{O}(h^r)$ . Thus the Taylor series method can be used to generate starting data which is consistent with any order predictor-corrector method provided only that  $f(x, y)$  is sufficiently smooth. However, many different function evaluations, as in (11b), are required and so this method is not very efficient.

Let us assume that

$$-C^2 \leq f_y(x, y) \leq -B^2 < 0 \quad \text{for all } x \in [a, b] \text{ and } |y| < \infty.$$

We show that in this case the bound (16) can be improved upon. Pick  $h_0$  such that

$$K' = -B^2 + \frac{h_0}{2!} M_2 + \dots + \frac{h_0^{r-1}}{r!} M_r < 0 \quad \text{and} \quad 1 + h_0 K'' > 0,$$

where

$$K'' = -C^2 - B^2 - K'.$$

Now we find, by retracing the proof of Theorem 1 with a little care that for all  $h \leq h_0$

$$(17) \quad |u_j - y(x_j)| \leq e^{K'(x_j - a)} \left[ |e_0| + \frac{M_{r+1} h^r}{(r+1)! |K'|} \right], \quad j = 0, 1, \dots, N.$$

We note that the exponential here is a decreasing function of  $x_j$ .

### 3.2. One-Step Methods Based on Quadrature Formulae

By integrating the differential equation (1) over  $[x_j, x_{j+1}]$  we get

$$(18a) \quad y(x_{j+1}) = y(x_j) + \int_{x_j}^{x_j+h} f(x, y(x)) dx.$$

Hence, we see that various forms for  $hF\{h, x, u; f\}$  are naturally suggested by quadrature formulae. However, as we are considering one-step methods the appropriate quadrature formulae should only employ nodes in  $[x_j, x_{j+1}]$ , say, for example, the  $n + 1$  points  $\xi_\nu$  satisfying

$$x_j \leq \xi_0 < \xi_1 < \dots < \xi_n \leq x_{j+1}.$$

But the integrand or an approximation to it must be known at these nodes and so we require approximations to  $y(\xi_\nu)$ ,  $\nu = 0, 1, \dots, n$ . We have for the exact solution at these points,

$$(18b) \quad y(\xi_\nu) = y(x_j) + \int_{x_j}^{\xi_\nu} f(x, y(x)) dx, \quad \nu = 0, 1, \dots, n.$$

Thus we could use a sequence of quadrature formulae to estimate successively the values  $y(\xi_\nu)$  and ultimately  $y(x_{j+1})$ .

A general class of one-step methods based upon these observations is given by using (2) with

$$(19a) \quad hF\{h, x_j, u_j; f\} = h \sum_{\mu=0}^n \alpha_\mu f(\xi_\mu, \eta_\mu);$$

$$(19b) \quad \xi_0 = x_j; \quad \xi_\nu = \xi_0 + \theta_\nu h, \quad \nu = 1, 2, \dots, n.$$

$$(19c) \quad \eta_0 = u_j; \quad \eta_\nu = \eta_0 + h \sum_{k=0}^{\nu-1} \alpha_{\nu k} f(\xi_k, \eta_k)$$

If in (19) we regard  $u_\nu$  as an approximation to  $y(\xi_\nu)$ , then the sums in (19a) and (19c) can be regarded as approximations to the integrals, respectively, in (18a) and (18b). In fact, these considerations suggest that we require

$$(20a) \quad 0 = \theta_0 \leq \theta_1 \leq \theta_2 \leq \dots \leq \theta_n \leq 1;$$

$$(20b) \quad \sum_{\nu=0}^n \alpha_\nu = 1;$$

$$(20c) \quad \sum_{k=0}^{\nu-1} \alpha_{\nu k} = \theta_\nu, \quad \nu = 1, 2, \dots, n.$$

Condition (20a) requires that all the nodes,  $\xi_v$ , lie in  $[x_j, x_{j+1}]$  and form a non-decreasing set; condition (20b) implies that the sum in (19a) has degree of precision at least 0 as a quadrature formula over  $[\xi_0, \xi_n]$ ; conditions (20c) imply similar results for the sums in (19c) as quadrature formulae over  $[\xi_0, \xi_v]$ . If, in addition to (20b), we require that

$$(21) \quad \sum_{v=0}^n \alpha_v (\theta_v)^p = \frac{1}{p+1}, \quad p = 1, 2, \dots, m;$$

then the basic quadrature scheme in (19a) has degree of precision at least  $m$ .

These considerations suggest many choices for the parameters in (19), some of which have been examined in the literature (i.e., Gaussian quadrature, equal coefficient formulae, etc.). In fact, in practice, the parameters are determined by the reasonable requirement that the local truncation error for a fixed choice of  $n$ , be of as high an order in  $h$  as possible. From (3), we determine the local truncation error,  $\tau_{j+1}$ , for the one-step method defined by (19) from

$$(22a) \quad y(x_j + h) = y(x_j) + h \sum_{v=0}^n \alpha_v f(\xi_v, y_{vj}) + h\tau_{j+1},$$

where

$$(22b) \quad y_{0j} = y(x_j); \quad y_{vj} = y_{0j} + h \sum_{k=0}^{v-1} \alpha_{vk} f(\xi_k, y_{kj}), \quad v = 1, 2, \dots, n.$$

If the parameters are given and  $f(x, y)$  has sufficiently many continuous derivatives, then  $y(x_j + h)$  and the  $f(\xi_k, y_{kj})$  can be expanded in powers of  $h$  [about  $x_j$ ,  $y(x_j)$ ]. Equation (22) then yields, upon equating coefficients of like powers of  $h$  in (22a), an expression for  $\tau_{j+1}$ . Obviously, this procedure can be used to determine the parameters in (19) such that  $\tau_{j+1}$  has the highest possible order in  $h$ . The use of Taylor's theorem here is similar to its use in Section 5 of Chapter 5 to determine high order approximations to derivatives, but is now much more complicated. We do not repeat here any of these lengthy calculations, but present in Table 2 some sets of parameter values for one-step methods of indicated order. It is found, in fact, that for  $n = 0$  and 1 the maximum orders are 1 and 2, respectively, and the conditions imposed are just those in (20) and (21) with  $m = 1$  or 2. For  $n = 2$  an order of 3 can be obtained, if, in addition to (20) and (21) with  $m = 3$  one additional relation is satisfied; namely,

$$\sum_{v=0}^n \alpha_v \sum_{k=0}^{v-1} \alpha_{vk} \theta_k = \frac{1}{6}.$$

(This relation can be explained as the result of requiring the coefficients  $\alpha_v$  and  $\alpha_{vk}$  to come from quadrature formulae with respective degrees of precision 2 and 1, at least.)

**Table 2** Some Standard Single-Step Difference Methods

Associated Name	n	Coefficients and Nodes				Order in $h$ of $\tau$
		$\nu$ or $j$	0	1	2	
Modified Euler	1	$\alpha_\nu =$	$\frac{1}{2}$	$\frac{1}{2}$		$\mathcal{O}(h^2)$
		$\theta_j =$	0	1		
		$\alpha_{1j} =$	1			
Heun	2	$\alpha_\nu =$	$\frac{1}{4}$	0	$\frac{3}{4}$	$\mathcal{O}(h^2)$
		$\theta_j =$	0	$\frac{1}{3}$	$\frac{2}{3}$	
		$\alpha_{1j} =$	$\frac{1}{3}$			
		$\alpha_{2j} =$	0	$\frac{2}{3}$		
Kutta	2	$\alpha_\nu =$	$\frac{1}{6}$	$\frac{2}{3}$	$\frac{1}{6}$	$\mathcal{O}(h^3)$
		$\theta_j =$	0	$\frac{1}{2}$	1	
		$\alpha_{1j} =$	$\frac{1}{2}$			
		$\alpha_{2j} =$	-1	2		
Runge-Kutta	3	$\alpha_\nu =$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\mathcal{O}(h^4)$
		$\theta_j =$	0	$\frac{1}{2}$	$\frac{1}{2}$	
		$\alpha_{1j} =$	$\frac{1}{2}$			
		$\alpha_{2j} =$	0	$\frac{1}{2}$		
		$\alpha_{3j} =$	0	0	1	
Runge-Kutta	3	$\alpha_\nu =$	$\frac{1}{8}$	$\frac{3}{8}$	$\frac{3}{8}$	$\mathcal{O}(h^4)$
		$\theta_j =$	0	$\frac{1}{3}$	$\frac{2}{3}$	
		$\alpha_{1j} =$	$\frac{1}{3}$			
		$\alpha_{2j} =$	$-\frac{1}{3}$	1		
		$\alpha_{3j} =$	1	-1	1	

We shall now show that all the schemes included in (19) satisfy the mean value property (7), if  $f_y(x, y)$  is continuous in  $S$ . Let the quantities  $\zeta_\nu$  be defined as the  $\eta_\nu$  are in (19) but with  $u_j$  replaced by  $v_j$ . Now we introduce the notation  $g(x, y) \equiv f_y(x, y)$  and use the continuity of  $g$  to deduce

$$(23a) \quad (\eta_\nu - \zeta_\nu) = (\eta_0 - \zeta_0) + h \sum_{k=0}^{\nu-1} \alpha_{\nu k} [f(\xi_k, \eta_k) - f(\xi_k, \zeta_k)] \\ = (\eta_0 - \zeta_0) + h \sum_{k=0}^{\nu-1} \alpha_{\nu k} g_k (\eta_k - \zeta_k); \quad \nu = 1, 2, \dots, n.$$

Here we have used

$$(23b) \quad g_k = g(\xi_k, \phi_k \eta_k + (1 - \phi_k) \zeta_k), \quad 0 < \phi_k < 1, \quad k = 0, 1, \dots, n.$$

By applying the equations in (23) recursively, we can determine expressions for the  $(\eta_v - \zeta_v)$  in terms of  $(\eta_0 - \zeta_0)$ . However, this procedure is rather complicated and so we will just present the result and verify it by induction.

Let us define the quantities  $B_{vj}$  as follows:  $B_{0j} \equiv 1$ ;

$$(24) \quad \begin{aligned} B_{vj} &= 1 & j = 0 \\ B_{vj} &= \sum_{k=j-1}^{v-1} \alpha_{vk} g_k B_{k,j-1} & j = 1, 2, \dots, v & v = 1, 2, \dots, n. \\ B_{vj} &= 0 & j \geq v + 1 \end{aligned}$$

Then we have

$$(25) \quad (\eta_v - \zeta_v) = (\eta_0 - \zeta_0)(B_{v0} + hB_{v1} + h^2B_{v2} + \dots + h^v B_{vv}),$$

$$v = 1, 2, \dots, n.$$

To verify (25) by induction, we note that from (23) and (24) with  $v = 1$ ,

$$(\eta_1 - \zeta_1) = (\eta_0 - \zeta_0)(1 + h\alpha_{10}g_0) = (\eta_0 - \zeta_0)(B_{10} + hB_{11}).$$

Thus (25) is valid for  $v = 1$ . We now assume (25) to be valid up to  $v - 1$  and use it in (23a) to obtain

$$\begin{aligned} (\eta_v - \zeta_v) &= (\eta_0 - \zeta_0) \left( 1 + h \sum_{k=0}^{v-1} \alpha_{vk} g_k \sum_{m=0}^k h^m B_{km} \right), \\ &= (\eta_0 - \zeta_0) \left( 1 + \sum_{m=0}^{v-1} h^{m+1} \sum_{k=m}^{v-1} \alpha_{vk} g_k B_{km} \right), \\ &= (\eta_0 - \zeta_0) \left( 1 + \sum_{m=0}^{v-1} h^{m+1} B_{v,m+1} \right). \end{aligned}$$

The induction is thus concluded and (25) is established.

We now obtain, from the mean value theorem and (25),

$$\begin{aligned} F\{h, x_j, u_j; f\} - F\{h, x_j, v_j; f\} &= \sum_{v=0}^n \alpha_v g_v (\eta_v - \zeta_v) \\ &= (u_j - v_j) \sum_{v=0}^n \alpha_v g_v \sum_{k=0}^v h^k B_{vk}, \end{aligned}$$

since  $(\eta_0 - \zeta_0) = (u_j - v_j)$ . That is, we have established

**LEMMA 1.** *Under the assumption that  $f_y(x, y)$  is continuous every one-step scheme defined by (19) satisfies the generalized mean value property (7) with*

$$(26) \quad G\{h, x, u, v; f\} \equiv \sum_{v=0}^n \left( \alpha_v g_v \sum_{k=0}^v h^k B_{vk} \right).$$

■

Here the  $B_{vk}$  are defined in (24) and the  $g_v$  and  $g_k$  are values of  $g(x, y) = f_y(x, y)$  at appropriate points of the form  $(x, \phi u + (1 - \phi)v)$ ,  $0 < \phi < 1$ . Again it should be observed that if  $f_y < 0$  in  $S$ , then by taking  $h$  so small that  $\alpha_v \geq 0$ , (26) implies  $G < 0$ . Hence in such cases, the initial error in one-step methods will decay exponentially with distance.

Of course, the indicated class of single-step methods satisfies a Lipschitz condition of the form (5). To obtain a suitable constant  $K_0$  we could use, from (26),

$$K_0 = \sup_{S, h \leq h_0} \left| \sum_{v=0}^n \alpha_v g_v \sum_{k=0}^v h^k B_{vk} \right|.$$

However, this is not readily calculable and so we shall determine an upper bound for it in terms of the parameters of the scheme and

$$M \equiv \sup_s \left| \frac{\partial f(x, y)}{\partial y} \right|.$$

We note that  $|g_v| \leq M$  for all  $v$ . Now define

$$\Phi \equiv \max_v \left( \sum_{k=0}^{v-1} |\alpha_{vk}| \right) \quad \text{and} \quad \beta_j \equiv \max_v |B_{vj}|,$$

and from (24) with  $j$  in  $1 \leq j \leq v$

$$\begin{aligned} |B_{vj}| &\leq M \sum_{k=j-1}^{v-1} |\alpha_{vk}| \cdot |B_{k,j-1}| \\ &\leq M \cdot \max_{j \leq t \leq v} |B_{t-1,t-1}| \cdot \sum_{k=j-1}^{v-1} |\alpha_{vk}| \\ &\leq M \Phi \beta_{j-1}. \end{aligned}$$

Since the right-hand side is independent of  $v$  we conclude that

$$\beta_j \leq M \Phi \beta_{j-1},$$

and by recursion using  $\beta_0 = 1$

$$\beta_j \leq (M \Phi)^j, \quad j = 0, 1, \dots, v.$$

Since  $|B_{vj}| \leq \beta_j$  for all  $v$  we have

$$\begin{aligned} (27) \quad K_0 &\leq \sum_{v=0}^n |\alpha_v| M \left[ 1 + \sum_{k=1}^v h^k (\Phi M)^k \right] \\ &\leq M \sum_{v=0}^n |\alpha_v| \frac{1 - (h \Phi M)^{v+1}}{1 - (h \Phi M)}. \end{aligned}$$

If we require that the coefficients  $\alpha_v$  be non-negative and satisfy at least (20b), then

$$\sum_{v=0}^n |\alpha_v| = \sum_{v=0}^n \alpha_v = 1.$$

If further,  $h$  is chosen such that  $h \leq h_0$ , where  $h_0 \Phi M < 1$ , then the above bound simplifies to

$$(28) \quad K_0 \leq \frac{M}{1 - h\Phi M}.$$

We also note that if the  $\alpha_{vk}$  are non-negative for all  $v$  and all  $k = 0, 1, \dots, v-1$  and if (20) is satisfied, then  $\Phi = \theta_n$  and hence  $0 \leq \Phi \leq 1$ . For sufficiently small  $h$ , in any event, the above bound can be made as close as we please to  $M$  which serves as the Lipschitz constant in the simple Euler method treated in Section 1.

### PROBLEMS, SECTION 3

1. Verify the entries in Table 2 under the name *Modified Euler*.
2. Verify the entries in Table 2 under the names *Heun* and *Kutta* with  $n = 2$ .
3. Verify the entries in Table 2 for both schemes under the name *Runge-Kutta* with  $n = 3$ .

### 4. LINEAR DIFFERENCE EQUATIONS

We recall that linear difference equations with constant coefficients have appeared previously in our study, for example, in Section 4 of Chapter 3 and in Subsection 1.4 of the present chapter. The theory of such difference equations will be sketched here because it will be used in the general treatment of difference methods given in the next section. The general linear difference equation with constant coefficients is a relation of the form

$$(1) \quad L(u_j) \equiv \sum_{s=0}^n a_s u_{j+s} = c_{j+n}, \quad j = j_0, j_0 + 1, \dots$$

Here the quantities  $a_s$  are the coefficients, the  $c_{j+n}$  are the inhomogeneous terms, and the sequence  $\{u_j\}$  is to be determined subject in general to additional conditions. Usually the sequence is desired starting from some initial index, say as indicated in (1) for  $j \geq j_0$ . The difference equation in (1) is said to be of order  $n$ , if  $a_n a_0 \neq 0$ , since then the indices on the  $u_j$  vary over  $n+1$  consecutive integers. We shall see that a solution of an  $n$ th order linear difference equation is, in general, determined by specifying  $n$

initial conditions. That is, if  $j_0$  is the initial index then we adjoin to (1) the conditions

$$(2) \quad u_{j_0} = v_0, \quad u_{j_0+1} = v_1, \dots, \quad u_{j_0+n-1} = v_{n-1}.$$

We now have

**THEOREM 1.** *If the difference equation (1) is of  $n$ th order then there is one and only one solution  $\{u_i\}$  satisfying the initial conditions (2).*

*Proof.* The existence of the solution follows trivially since  $a_n \neq 0$  implies from (1) that

$$(3) \quad u_{j+n} = -\frac{1}{a_n} \sum_{s=0}^{n-1} (a_s u_{j+s}) + \frac{c_{j+n}}{a_n}, \quad j = j_0, j_0 + 1, \dots$$

For uniqueness let there be two solutions,  $\{u'_i\}$  and  $\{u''_i\}$ . Then their difference  $\{u_i\} \equiv \{u'_i - u''_i\}$  satisfies (2) with  $v_0 = v_1 = \dots = v_{n-1} = 0$  and (3) with  $c_{j+n} \equiv 0$ . Thus we find that  $\{u_i\} \equiv \{0\}$  and the proof is complete. ■

We consider the  $n$ th order *homogeneous difference equations* corresponding to (1), namely:

$$(4) \quad L(u_i) = 0; \quad j = j_0, j_0 + 1, \dots$$

If the sequences  $\{u_i\}$  and  $\{v_i\}$  are solutions of (4) then, by the linearity of these equations, the sequence  $\{\alpha u_i + \beta v_i\}$  is also a solution. Here  $\alpha$  and  $\beta$  are arbitrary numbers. Thus we easily find that the set of all solutions of (4) forms a linear vector space. A set of solutions, say  $r$  of them

$$\{u_i^{(1)}\}, \{u_i^{(2)}\}, \dots, \{u_i^{(r)}\},$$

are *linearly independent* if only the *trivial* combination of the  $\{u_i^{(v)}\}$  vanishes identically; that is, if

$$\alpha_1 u_i^{(1)} + \alpha_2 u_i^{(2)} + \dots + \alpha_r u_i^{(r)} = 0, \quad \text{for } i = j_0, j_0 + 1, \dots,$$

implies that  $\alpha_1 = \alpha_2 = \dots = \alpha_r = 0$ . This is essentially the same notion as the linear independence of vectors. A set of  $n$  independent solutions of the  $n$ th order equations (4) is called a *fundamental set of solutions*.

A basic result now can be stated as

**THEOREM 2.** *Let  $\{u_i^{(v)}\}$ ,  $v = 1, 2, \dots, n$ , be a fundamental set of solutions of the homogeneous difference equations (4). Then any solution,  $\{v_i\}$ , of these equations can be expressed uniquely in the form*

$$\{v_i\} = \left\{ \sum_{v=1}^n \alpha_v u_i^{(v)} \right\}.$$

*Proof.* Since the set  $\{u_i^{(v)}\}$  is independent, Theorem 1 implies that the  $n$  vectors  $\mathbf{u}^{(v)}$ , where  $\mathbf{u}^{(v)} \equiv (u_i^{(v)})$  for  $i = j_0, j_0 + 1, \dots, j_0 + n - 1$ , are linearly independent. That is, according to Theorem 1, if the  $\mathbf{u}^{(v)}$  were linearly dependent, then the corresponding infinite sequences  $\{u_i^{(v)}\}$  would be linearly dependent. Hence the  $n$ th order matrix

$$A \equiv \begin{bmatrix} u_{j_0}^{(1)} & u_{j_0}^{(2)} & \cdots & u_{j_0}^{(n)} \\ u_{j_0+1}^{(1)} & u_{j_0+1}^{(2)} & \cdots & u_{j_0+1}^{(n)} \\ \vdots & \vdots & & \vdots \\ u_{j_0+n-1}^{(1)} & u_{j_0+n-1}^{(2)} & \cdots & u_{j_0+n-1}^{(n)} \end{bmatrix}$$

is non-singular. Thus given any  $n$  components, say,  $\mathbf{v} \equiv \{v_i\}$  for  $i = j_0, j_0 + 1, \dots, j_0 + n - 1$ , we can uniquely solve the  $n$ th order system

$$A\alpha = \mathbf{v}.$$

The components  $\alpha_v$  of  $\alpha$  are the coefficients to be used in the theorem. Since the first  $n$  components of any solution  $\{v_i\}$  can be expressed as a linear combination of the first  $n$  components of the fundamental set the theorem now follows by an application of Theorem 1. ■

We can, furthermore, find a fundamental set of solutions of (4). We try as a solution the powers of some scalar, say

$$u_i = \alpha x^i; \quad i = j_0, j_0 + 1, \dots$$

Then (4) yields

$$(a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0)(\alpha x^j) = 0.$$

If  $\alpha x^j = 0$  the corresponding solution is trivial and does not lead to a fundamental set. Hence, we only consider the roots of

$$(5) \quad p_n(x) \equiv a_n x^n + a_{n-1} x^{n-1} + \cdots + a_0 = 0.$$

The  $n$ th degree polynomial  $p_n(x)$  is called the *characteristic polynomial* of the difference equation (4). We easily find that if  $x$  is a root of (5) then  $\{u_i\} = \{x^i\}$  is a solution of the homogeneous difference equations. If the roots of the characteristic equation are distinct, say  $x_1, x_2, \dots, x_n$ , then a fundamental set of solutions is given by  $\{u_i^{(v)}\} = \{x_v^i\}$ ,  $v = 1, 2, \dots, n$ . Since  $a_n a_0 \neq 0$ , there is no zero root and the independence of the  $\{u_i^{(v)}\}$  follows from the independence of the first  $n$  components. That is, let us define the matrix

$$U \equiv (\mathbf{u}^{(1)}, \mathbf{u}^{(2)}, \dots, \mathbf{u}^{(n)}),$$

whose columns are vectors obtained from the first  $n$  components of the

set of solutions defined above. Then clearly,  $U$  is non-singular if the  $x_i$  are distinct, since

$$U = \begin{bmatrix} 1 & 1 & \cdots & 1 \\ x_1 & x_2 & \cdots & x_n \\ \vdots & \vdots & & \vdots \\ x_1^{n-1} & x_2^{n-1} & \cdots & x_n^{n-1} \end{bmatrix} \begin{bmatrix} x_1^{j_0} & & 0 \\ & x_2^{j_0} & \\ & & \ddots \\ 0 & & x_n^{j_0} \end{bmatrix}.$$

If the roots  $x_i$  are not distinct, we can still define a fundamental set of solutions. Let  $x_1$  be a root of multiplicity  $m_1 > 1$  of  $p_n(x) = 0$ . Then we use the powers to generate one solution and successive derivatives† with respect to  $x_1$ , up to order  $m_1 - 1$  to generate  $m_1 - 1$  additional solutions. Specifically, let  $u_j^{(1)} = x_1^j$ . Now try

$$\{u_j^{(2)}\} = \frac{d}{dx_1} \{u_j^{(1)}\}, \dots, \{u_j^{(m_1)}\} = \frac{d}{dx_1} \{u_j^{(m_1-1)}\}.$$

However, since any solution can be multiplied by a non-zero constant we multiply the resulting  $\{u_j^{(\nu)}\}$  by  $x_1^{\nu-1}$ , to retain the original powers of  $x_1$  in corresponding terms, that is, we introduce

$$\{v_j^{(\nu)}\} = x_1^{\nu-1} \{u_j^{(\nu)}\}, \quad \nu = 1, 2, \dots, m_1.$$

The elements of these sequences are found to be

$$\begin{aligned} v_j^{(1)} &= x_1^j, \\ v_j^{(2)} &= jx_1^j, \\ v_j^{(3)} &= j(j-1)x_1^j, \\ &\vdots \\ v_j^{(m_1)} &= j(j-1)\cdots(j-m_1+2)x_1^j, \quad \text{for } j = j_0, j_0 + 1, \dots \end{aligned}$$

We leave the verification that these form  $m_1$  solutions as Problem 1.

By forming linear combinations of the solutions  $\{v_j^{(\nu)}\}$  corresponding to a root,  $x_1$ , of multiplicity  $m_1$ , we find the simpler sequences  $\{w_j^{(\nu)}\}$

$$\begin{aligned} (6) \quad w_j^{(1)} &= x_1^j, \\ w_j^{(2)} &= jx_1^j, \\ w_j^{(3)} &= j^2x_1^j, \\ &\vdots \\ w_j^{(m_1)} &= j^{m_1-1}x_1^j, \quad j = j_0, j_0 + 1, \dots \end{aligned}$$

† To motivate this procedure, observe that if  $x_1$  and  $x_2 = x_1 + h$  are two roots of (5), then

$$u_i = h^{-1}[(x_1 + h)^i - x_1^i], \quad i = j_0, j_0 + 1, \dots,$$

is a solution of (4). But then

$$\lim_{h \rightarrow 0} u_i = ix_1^{i-1}$$

is also a solution.

These sequences are obviously linearly independent and are solutions because they can be obtained as linear combinations of the  $\{v_j^{(v)}\}$ .

Let us now consider the inhomogeneous  $n$ th order linear difference equations (1). If  $\{v_j\}$  is any particular solution of equation (1) and  $\{u_j\}$  is a solution of the homogeneous system (4), then  $\{u_j + v_j\}$  is a solution of (1). This solution of (1) can be made to satisfy any particular initial conditions by adjusting the  $\{u_j\}$ . We now develop a discrete analog of what is known as *Duhamel's principle* in the theory of differential equations (where integral representations of solutions are obtained).

**THEOREM 3.** *Let  $\{u_j^{(v)}\}$  be the fundamental set of solutions of the  $n$ th order homogeneous difference equation (4) which satisfy the initial conditions*

$$(7) \quad u_i^{(v)} = \delta_{iv}, \quad i = 0, 1, \dots, n - 1; \quad v = 0, 1, \dots, n - 1.$$

*Then the solution of (1) subject to the initial conditions (2) with  $j_0 \equiv 0$ , is given by*

$$(8) \quad u_j = \sum_{v=0}^{n-1} v \bar{u}_j^{(v)} + \frac{1}{a_n} \sum_{k=0}^{j-n} c_{k+n} \bar{u}_{j-k-1}^{(n-1)}, \quad j = 0, 1, \dots$$

(Here we define  $u_i^{(n-1)} \equiv 0$  for all  $i < 0$  and  $c_j \equiv 0$  for all  $j < n$ .)

*Proof.* The first sum in (8) satisfies the initial conditions (2) and the homogeneous difference equations. Thus we need only show that the second sum in (8) satisfies homogeneous initial conditions and the inhomogeneous difference equations (1), with  $j_0 = 0$ . Let us define

$$w_j \equiv \frac{1}{a_n} \sum_{k=0}^{j-n} c_{k+n} \bar{u}_{j-k-1}^{(n-1)}; \quad j = 0, 1, \dots$$

Then we have  $w_j = 0$  for  $j = 0, 1, \dots, n - 1$  by recalling that  $u_i^{(n-1)} = 0$  for  $i \leq n - 2$  and  $c_j = 0$  for  $j < n$ . In fact, for the same reason, we may write

$$w_j = \frac{1}{a_n} \sum_{k=-\infty}^{\infty} c_{k+n} \bar{u}_{j-k-1}^{(n-1)},$$

since the additional terms vanish. Then

$$\begin{aligned} L(w_j) &= \sum_{s=0}^n a_s w_{j+s} = \frac{1}{a_n} \sum_{s=0}^n a_s \sum_{k=-\infty}^{\infty} c_{k+n} \bar{u}_{j+s-k-1}^{(n-1)} \\ &= \frac{1}{a_n} \sum_{s=0}^n a_s \sum_{k=0}^j c_{k+n} \bar{u}_{j+s-k-1}^{(n-1)}, \end{aligned}$$

since the terms corresponding to other values of  $k$  vanish. Hence,

$$L(w_j) = \frac{1}{a_n} \sum_{k=0}^j c_{k+n} L(\bar{u}_{j-k-1}^{(n-1)}).$$

However, it is easily verified that

$$L(u_{j-k-1}^{(n-1)}) = a_n \delta_{jk}$$

and so finally,

$$L(w_j) = c_{j+n}. \quad \blacksquare$$

## PROBLEMS, SECTION 4

1. Verify that equation (6) defines  $m_1$  linearly independent solutions of (4).
2. Verify that a fundamental set of solutions of (4) is obtained from (6), if  $m_1$  is replaced by  $m_i$  and  $x_1$  by  $x_i$  for each root  $x_i$  of (5) of multiplicity  $m_i$ .
  
5. **CONSISTENCY, CONVERGENCE, AND STABILITY OF DIFFERENCE METHODS**

The numerical procedures which we have introduced in Sections 1 through 3 in order to approximate the solutions of differential equations may be called difference methods. In this section we study the convergence of a more general class of difference schemes. The analysis constitutes a uniform development for all of the commonly used methods that were treated separately in Sections 1 through 3.

The solution of the difference equations is what we try to compute, and this may have to be done for very fine meshes, i.e., for many net points. Thus, as a practical matter, it is important that these solutions should not be too sensitive to small errors in the computations (for example, roundoff errors). This sensitivity to errors is related to what is called the *stability of the difference equations*. We have already investigated such matters but without the introduction of this terminology. We shall see that for *consistent* methods, stability of the difference equations is equivalent to convergence of the difference equation solution to the solution of the differential equation problem.

As usual, we consider methods for approximating the solution,  $y(x)$ , of the initial value problem

$$(1) \quad \begin{aligned} y' &= f(x, y), & a \leq x \leq b, \\ y(a) &= y_0. \end{aligned}$$

We assume that  $f(x, y)$  is in the class  $\mathcal{F}$  of functions such that  $f_y(x, y)$ ,  $f_x(x, y)$ , and all partial derivatives of  $f$  of some finite order  $q \geq 1$  are continuous and uniformly bounded in  $S$ :  $\{(x, y) \mid a \leq x \leq b; |y| < \infty\}$ . For any fixed net spacing  $h = (b - a)/N$ , we use a uniform net  $x_j = a + jh$ ,  $j = 0, 1, \dots, N$ , and seek approximations  $u_j$  to  $y(x_j)$  on this net.

The approximations are defined as the solution of some difference problem, say

$$(2a) \quad a_n u_{j+n} + \cdots + a_0 u_j = h F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\} + h \rho_{j+n}, \\ j = m, m+1, \dots, N-n;$$

where  $\{a_i\}$  are real constants independent of  $h$  satisfying  $a_n a_0 \neq 0$ , and  $\rho_{j+n}$  is the *local rounding error* subject to

$$|\rho_{j+n}| \leq \rho(h), \quad \text{for } j \geq m.$$

The initial data are specified as, say

$$(2b) \quad u_0 = y_0 + \rho_0, u_1 = y_1 + \rho_1, \dots, u_{m+n-1} = y_{m+n-1} + \rho_{m+n-1}$$

where

$$|\rho_k| \leq r(h), \text{ for } 0 \leq k \leq m+n-1.$$

We shall later require that  $\rho(h) \rightarrow 0$  and  $r(h) \rightarrow 0$  as  $h \rightarrow 0$ .

By suitably defining  $F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\}$ , we may incorporate in (2) all of the schemes treated in the previous sections. On the other hand, the only properties that we need postulate for  $F$ , in order to make this general study of convergence, are easily seen to hold for all of the commonly used difference methods. That is, we require

$$(3a) \quad F\{h, x_j; u_{j-m}, \dots, u_{j+n}; 0\} \equiv 0;$$

$$(3b) \quad |F\{h, x_j; v_{j-m}, \dots, v_{j+n}; f\} - F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\}|$$

$$\leq C \sum_{k=-m}^n |v_{j+k} - u_{j+k}|,$$

where the constant  $C$  depends only on the bounds of  $f$  and a finite number of its partial derivatives in  $S$ . The *local truncation error*,  $\tau_{j+n}$ , is defined by

$$(4a) \quad \sum_{k=0}^n a_k y_{j+k} - h F\{h, x_j; y_{j-m}, \dots, y_{j+n}; f\} = h \tau_{j+n},$$

where  $y(x)$  is a solution of (1). We further require that

$$(4b) \quad |\tau_{j+n}| \leq \tau(h) \quad \text{for } m \leq j \leq N-n$$

and

$$(4c) \quad \lim_{h \rightarrow 0} \tau(h) = 0,$$

i.e., the truncation error tends to zero. Condition (4c) implies that the difference equation (2) is an “approximation” to (1), rather than some other equation. Strictly speaking, we say that (2) is *consistent* with (1).

if  $r(h) \rightarrow 0$  and  $\tau(h) \rightarrow 0$  as  $h \rightarrow 0$ . [For example, let  $m = 0$ ,  $n = 1$ ;  $a_0 = -1$ ;  $a_1 = 1$  and

$$F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\} \equiv f(x_{j+1}, u_{j+1}) + f(x_j, u_j).$$

Then (2) is not consistent with the equation (1). That is, by using Taylor series in (4a), for small  $h$ ,  $\tau_{j+1} \approx -f(x_j, y_j)$ . Hence  $\tau_j$  does not approach zero. In fact, this scheme is consistent with the equation  $y' = 2f(x, y)$ .] If, for all  $h \leq h_0$ ,

$$(5) \quad \tau_{j+n} \leq \tau(h) \equiv Mh^p, \quad m \leq j \leq N-n,$$

where  $M$  depends only on the bounds of  $f$  and a finite number of its derivatives in  $S$ , we say that the truncation error of the difference method is of *order p*.

We, of course, are interested in characterizing the convergent schemes and in obtaining an estimate of the error. For a fixed mesh width,  $h$ , we define the *pointwise error*

$$e_j \equiv u_j - y_j, \quad \text{where } y_j = y(x_j).$$

The method (2) is *convergent* if, for any  $f(x, y)$  in  $\mathcal{F}$ ,  $\max_{0 \leq j \leq N} |e_j| \rightarrow 0$  as  $h \rightarrow 0$ , provided that the rounding errors  $\rho(h)$  and  $r(h)$  tend to zero.

If scheme (2) is convergent for all  $f$  in  $\mathcal{F}$ , then it is convergent for the problem (1) with  $f(x, y) \equiv 0$ ,  $y_0 = 0$ . From this simple observation, we note that if (2) is convergent and  $F$  satisfies (3a), then the solution  $\{u_j\}$  of

$$(6) \quad \begin{aligned} a_n u_{j+n} + a_{n-1} u_{j+n-1} + \cdots + a_0 u_j &= 0 \\ u_0 = \rho_0, u_1 = \rho_1, \dots, u_{n-1} = \rho_{n-1}, \end{aligned}$$

must tend to the solution  $y(x) \equiv 0$ , for any set of initial errors  $\{\rho_k\}$  such that  $\max_k |\rho_k| \rightarrow 0$  as  $h \rightarrow 0$ .

We say that the difference method (2) satisfies the *root condition* if

$$(7a) \quad P(\zeta) \equiv a_n \zeta^n + a_{n-1} \zeta^{n-1} + \cdots + a_1 \zeta + a_0,$$

has only zeros  $\zeta_i$  such that

$$(7b) \quad |\zeta_i| \leq 1,$$

and the multiplicities,  $r_i$ , of the  $\zeta_i$  are such that if

$$(7c) \quad |\zeta_k| = 1, \quad \text{then } r_k = 1.$$

In other words, scheme (2) satisfies the root condition, if the zeros of  $P(\zeta)$  lie in the unit circle and only simple zeros may lie on the boundary of the unit circle. We can now establish a necessary condition that (2) be convergent; i.e., for the solution of (6) to tend to zero.

**THEOREM 1.** *If (2) is convergent and F satisfies (3a), then (2) satisfies the root condition (7).*

*Proof.* We show by contradiction that the root condition is necessary. That is, if  $|\zeta_i| > 1$  and  $\zeta_i$  is a complex root of  $P(\zeta)$ , define

$$(8a) \quad u_j \equiv h(\zeta_i^j + \bar{\zeta}_i^j), \quad j = 0, 1, \dots;$$

if  $\zeta_i$  is a real root set

$$(8b) \quad u_j \equiv h\zeta_i^j.$$

Clearly, (8) is a solution of (6) with  $\rho_k = u_k$  for  $k = 0, 1, \dots, n - 1$ , and  $\max |\rho_k| \rightarrow 0$  as  $h \rightarrow 0$ . On the other hand, for any  $c$  in  $a < c < b$  set  $j = [c/h]$ . But then  $|u_{[c/h]}| \rightarrow \infty$  as  $h \rightarrow 0$ . Hence such a scheme is not convergent.

If on the other hand,  $|\zeta_i| = 1$  and  $\zeta_i$  is a multiple root and complex set

$$(9a) \quad u_j = hj(\zeta_i^j + \bar{\zeta}_i^j), \quad j = 0, 1, \dots;$$

while if  $\zeta_i$  is real, set

$$(9b) \quad u_j = hj\zeta_i^j.$$

Now if  $j = [c/h]$ ,  $|u_{[c/h]}|$  does not approach zero as  $h \rightarrow 0$ . Hence such a scheme is not convergent. ■

The requirement that the solution  $\{u_i\}$  of (2) depend Lipschitz continuously on  $\{\rho_k\}$  is the definition of *stability*. That is, we say that (2) is *stable* if for any  $f$  in  $\mathcal{F}$ , there is an  $h_0$  and an  $M$ , such that for all  $0 < h \leq h_0$ , and  $N \equiv N(h) = (b - a)/h$

$$(10a) \quad |u_i - v_i| \leq M\epsilon, \quad \text{for } 0 \leq i \leq N$$

whenever  $\{v_i\}$  satisfies

$$(10b) \quad a_n v_{j+n} + \dots + a_0 v_j = hF\{h, x_j; v_{j-m}; \dots, v_{j+n}; f\} + h\sigma_{j+n}, \\ j \geq m,$$

$$(10c) \quad v_0 = y_0 + \sigma_0, \dots, v_{m+n-1} = y_{m+n-1} + \sigma_{m+n-1}$$

where

$$|\rho_k - \sigma_k| \leq \epsilon \quad \text{for } 0 \leq k \leq N.$$

It is then easy to show

**THEOREM 2.** *If the scheme (2) is stable and F satisfies (3a) then the root condition (7) is satisfied.*

*Proof.* The proof follows by contradiction as did the previous theorem. Merely verify that in the case  $f \equiv 0$  and  $\rho_{j+n} = 0$  for  $j \geq 0$ , the definitions

(8) or (9) with  $h$  replaced by  $\delta$  define a solution  $\{u_i\}$  of (2). On the other hand, set  $\sigma_k = 0$  and  $v_i = 0$ . Then it follows that

$$|\rho_k - \sigma_k| \leq \epsilon \quad \text{for } 0 \leq k \leq n - 1,$$

where  $\epsilon$  is proportional to  $\delta$ . But now, (10a) cannot be satisfied for any fixed  $M$  as  $h \rightarrow 0$ . ■

Next we have

**THEOREM 3.** *If (2) is consistent with (1), i.e. satisfies (4) and  $r(h) \rightarrow 0$  as  $h \rightarrow 0$ , and  $F$  satisfies (3), then (2) is a convergent scheme if and only if the root condition (7) is satisfied.*

*Proof.* In Theorem 1 we have shown that the root condition is necessary for convergence. We now assume that the root condition holds and prove that (2) is convergent. By subtracting equation (4a) from equation (2a), we obtain a difference equation satisfied by the pointwise error  $e_j = u_j - y_j$ ,

$$(11) \quad \sum_{s=0}^n a_s e_{j+s} = c_{j+n}, \quad \text{for } m \leq j \leq N - n,$$

where

$$\begin{aligned} c_{j+n} \equiv h[F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\} - F\{h, x_j; y_{j-m}, \dots, y_{j+n}; f\}] \\ + h\rho_{j+n} - h\tau_{j+n}. \end{aligned}$$

We may solve the inhomogeneous difference equation (11), by using Theorem 4.3, in the form

$$(12) \quad e_{j+m} = \sum_{k=0}^{n-1} e_{m+k} u_j^{(k)} + \frac{1}{a_n} \sum_{k=0}^{j-n} c_{k+m+n} u_{j-k-1}^{(n-1)},$$

for  $j = 0, 1, \dots, N - m$ .

[That is, define

$$E_j \equiv e_{j+m}, \quad C_j \equiv c_{j+m}, \quad \text{for } j = 0, 1, \dots$$

Then (11) holds, i.e.,

$$\sum_{s=0}^n a_s E_{j+s} = C_{j+n} \quad \text{for } 0 \leq j \leq N - m - n.$$

Hence equation (4.8) gives a representation for  $E_j$ , which reduces to (12).]

But because the root condition is satisfied, we know from Theorem 4.2 and equation (4.6) that the solutions  $\{u_j^{(k)}\}$  satisfy

$$(13) \quad |u_j^{(k)}| \leq Q$$

for some constant  $Q$  independent of  $k$  and  $j$ .

Furthermore, from the definitions of  $c_{j+n}$ ,  $\rho(h)$ ,  $\tau(h)$  that appear after equations (11), (2a), and (4a) respectively, and from (3b), we find

$$(14) \quad |c_{k+m+n}| \leq h \left[ C \sum_{r=0}^{m+n} |e_{k+r}| + \rho(h) + \tau(h) \right].$$

If we use the estimates (13) and (14) in (12), we have

$$\begin{aligned} |e_{j+m}| &\leq Qn \max_{0 \leq k \leq n-1} |e_{m+k}| \\ &+ \frac{Q}{a_n} (j-n+1)h[(m+n+1)C \max_{0 \leq r \leq j+m} |e_r| + \rho(h) + \tau(h)] \end{aligned}$$

This inequality simplifies, if we introduce

$$\omega_j = \max_{0 \leq k \leq j} |e_k|,$$

to read

$$\begin{aligned} (15) \quad |e_{j+m}| &\leq Qn\omega_{m+n-1} + \frac{Q(j-n+1)h}{a_n} \\ &\times [(m+n+1)C\omega_{j+m} + \rho(h) + \tau(h)], \\ &\text{for } j = 0, 1, \dots, N-m. \end{aligned}$$

Since  $\omega_{j+m}$  is equal to  $|e_k|$  for some index  $k \leq j+m$  and since  $n \geq 1$ , we find from (15) that a fortiori

$$(16) \quad \omega_{j+m} \leq \kappa j h \omega_{j+m} + Qn\omega_{m+n-1} + \frac{Qjh}{a_n} [\rho(h) + \tau(h)],$$

where

$$\kappa = \frac{QC(m+n+1)}{a_n}, \quad \text{for } j = 0, 1, \dots, N-m.$$

If we limit the range of  $j$ , as  $h$  tends to zero, so that

$$(17) \quad jh\kappa \leq \frac{1}{2},$$

then (16) yields

$$(18) \quad \omega_{j+m} \leq 2Q \left[ n\omega_{m+n-1} + \frac{\rho(h) + \tau(h)}{2\kappa a_n} \right] \quad \text{for } 0 \leq j \leq \frac{1}{2\kappa h}.$$

If we now employ the definition of  $\omega_j$ , (18) yields, using  $r(h)$  defined after (2b),

$$(19) \quad |e_{j+m}| \leq 2Q \left[ nr(h) + \frac{\rho(h) + \tau(h)}{2\kappa a_n} \right] \quad \text{for } 0 \leq jh \leq \frac{1}{2\kappa}.$$

Equation (19) bounds the pointwise error, for a finite interval  $(a, a + 1/(2\kappa))$ , in terms of the bound for the initial error,  $\omega_{m+n-1}$ , and

the bounds  $\rho(h)$  and  $\tau(h)$  of the rounding and truncation errors. The length of the interval of convergence,  $1/(2\kappa)$ , is independent of  $h$  and is defined after (16).

Hence we may repeat this argument by beginning with the  $m + n$  errors bounded by (19).

$$e_{[1/(2\kappa h)] - m - n + 1}, e_{[1/(2\kappa h)] - m - n + 2}, \dots, e_{[1/(2\kappa h)]}.$$

In this way, we may successively establish pointwise convergence as  $h \rightarrow 0$ , in the finite number of intervals

$$\left(a + \frac{1}{2\kappa}, a + \frac{2}{2\kappa}\right), \left(a + \frac{2}{2\kappa}, a + \frac{3}{2\kappa}\right), \dots, \left(a + \frac{R}{2\kappa}, a + (b - a)\right),$$

where  $R = [(b - a)/(2\kappa)]$ . The error estimates for successive intervals can then be seen to satisfy, in analogy with (19),

$$(20) \quad I_{p+1} \leq 2Q \left[ nI_p + \frac{\rho(h) + \tau(h)}{2\kappa a_n} \right], \quad \text{for } p = 1, 2, \dots, R,$$

where  $I_p$  is the pointwise error bound for the interval

$$\left(a + \frac{p-1}{2\kappa}, a + \frac{p}{2\kappa}\right).$$

From (20) it is then possible to recursively bound  $I_{R+1}$  and hence to bound  $|e_j|$  for  $0 \leq j \leq N$  by

$$(21) \quad |e_j| \leq (2Qn)^{R+1}r(h) + \frac{(2Qn)^{R+1} - 1}{2Qn - 1} \frac{Q(\rho(h) + \tau(h))}{\kappa a_n},$$

if  $2Qn \neq 1$ ;

$$\leq r(h) + \frac{(R+1)Q(\rho(h) + \tau(h))}{\kappa a_n}, \quad \text{if } 2Qn = 1.$$

Formula (21) not only establishes convergence of the finite difference scheme (2), but gives an upper bound for the error  $e_j$  in terms of the initial error, the rounding error and the truncation error. This bound is of the same general character as were the bounds that we derived earlier for the special methods treated in Sections 1 through 3. ■

By essentially the same arguments we could prove:

**THEOREM 4.** *If the  $F$  in (2) satisfies (3) then (2) is a stable scheme iff the root condition (7) is satisfied.*

We have therefore established the important consequence of Theorems 3 and 4;

**THEOREM 5.** *If the scheme (2) is consistent with (1) and F satisfies condition (3), then the necessary and sufficient condition that (2) be convergent is that it be stable.* ■

It is possible to strengthen Theorem 3 by noting that F need only satisfy the Lipschitz condition (3b) in a narrow strip about the solution  $y(x)$  given by  $S_d$ :  $\{(x, y) \mid a \leq x \leq b; |y - y(x)| \leq d\}$  for any fixed constant  $d > 0$ . That is, for  $h$  sufficiently small, the error estimate (21) shows that if the solution of the difference equation starts in the strip  $S_{d/2}$  then it remains in the strip  $S_d$ .

The special case with  $m = 0$  and

$$(22) \quad F\{h, x_j; u_{j-m}, \dots, u_{j+n}; f\} \equiv \sum_{s=0}^n b_s f(x_{j+s}, u_{j+s})$$

has been treated by Dahlquist. He found the surprising result that although by proper choice of the  $2n + 1$  independent parameters  $\{a_s/a_n\}$ ,  $\{b_s/a_n\}$ , it is possible to construct a scheme having a truncation error of order  $2n$ ; *only schemes with a truncation error of order at most  $n + 2$  may be convergent.* (In fact, if  $n$  is odd then only schemes for which the truncation error is of order at most  $n + 1$  may be convergent.) The implicit scheme of equation (2.3a) with  $p = n - 1$ , based on the Newton-Cotes quadrature formulae applied to (2.2) with  $p = n - 1$ , then has the maximum possible order of truncation error for convergent schemes of form (2) with F given by (22). Dahlquist's work finds other schemes having a truncation error of the same order, but shows that schemes which are both convergent and of greater accuracy do not exist.

## PROBLEMS, SECTION 5

1. Define F for the following schemes treated in Sections 1 through 3 given by

- (a) equation (1.1a)
- (b) equation (1.24a)
- (c) equation (1.37a)
- (d) equation (2.3a)
- (e) equation (2.3b)
- (f) equation (2.7a and b)
- (g) equation (3.11a and b)
- (h) equation (3.8) and (3.19a, b, and c)

and verify that conditions (3a and b) are satisfied. Which of these schemes do not satisfy the root condition?

2. If (2) is convergent, show that  $P(1) = \sum_{s=0}^n a_s = 0$ .

[Hint: Let  $f(x, y) \equiv 0$ ;  $y(a) = y_0 \neq 0$ ;  $\rho_k = 0$ .]

3. In the scheme (2) with  $F$  given by (22), show that with  $T(\zeta) \equiv \sum_{s=0}^n b_s \zeta^s$ ,  $P'(1) = T(1)$  implies that the truncation error is of order  $p \geq 1$ .  
 [Hint: Expand the left side of equation (3c) about  $(x_j, y_j)$  in powers of  $h$ . Observe that

$$P'(1) = \sum_{s=0}^n s a_s, \quad T(1) = \sum_{s=0}^n b_s.$$

## 6. HIGHER ORDER EQUATIONS AND SYSTEMS

Any  $r$ th order ordinary differential equation,

$$\frac{d^r z}{dx^r} = g\left(x, z, \frac{dz}{dx}, \dots, \frac{d^{r-1} z}{dx^{r-1}}\right),$$

can be replaced by an equivalent system of first order equations. There are a variety of ways in which this reduction can be performed; the most straightforward introduces the variables

$$y^{(1)}(x) \equiv z(x), y^{(2)}(x) \equiv \frac{dy^{(1)}(x)}{dx}, \dots, y^{(r)}(x) \equiv \frac{dy^{(r-1)}(x)}{dx}.$$

Then the differential equation can be written as

$$\begin{aligned} \frac{d}{dx} y^{(1)} &= y^{(2)}, \\ &\vdots \\ \frac{d}{dx} y^{(r-1)} &= y^{(r)} \\ \frac{d}{dx} y^{(r)} &= g(x, y^{(1)}, y^{(2)}, \dots, y^{(r)}). \end{aligned}$$

This is, of course, a special case of the general system

$$(1a) \quad \frac{d\mathbf{y}}{dx} = \mathbf{f}(x; \mathbf{y}).$$

Here we have introduced the  $r$ -dimensional column vectors  $\mathbf{y}$  and  $\mathbf{f}$  with components

$$y^{(\nu)}(x), f^{(\nu)}(x; \mathbf{y}) = f^{(\nu)}(x, y^{(1)}, y^{(2)}, \dots, y^{(r)}), \quad \nu = 1, 2, \dots, r.$$

We will study the difference methods appropriate for solving a system (1a). The initial data for such a system are assumed given in the form

$$(1b) \quad \mathbf{y}(a) = \mathbf{y}_0,$$

where we seek a solution of (1) in the interval  $a \leq x \leq b$ .

All of the difference methods previously proposed for a single first order equation have their direct analogs for the system (1). With (1a) in component form,

$$\frac{dy^{(v)}}{dx} = f^{(v)}(x, y^{(1)}(x), \dots, y^{(r)}(x)), \quad v = 1, 2, \dots, r;$$

it does not require much insight to write down the corresponding difference methods based on quadrature formulae or even the single-step methods. In fact, the general predictor-corrector becomes, in vector form,

$$(2a) \quad \mathbf{u}_{i+1}^* = \mathbf{u}_{i-q} + h \sum_{j=0}^m \beta_j \mathbf{f}(x_{i-j}; \mathbf{u}_{i-j});$$

$$(2b) \quad \mathbf{u}_{i+1} = \mathbf{u}_{i-p} + h \sum_{k=1}^n \alpha_k \mathbf{f}(x_{i+1-k}; \mathbf{u}_{i+1-k}) + h \alpha_0 \mathbf{f}(x_{i+1}; \mathbf{u}_{i+1}^*).$$

Similarly, the general one-step difference methods for the system (1) can be written as

$$(3) \quad \mathbf{u}_{j+1} = \mathbf{u}_j + h \mathbf{F}\{h, x_j; \mathbf{u}_j; \mathbf{f}\}.$$

For the class of methods defined in Subsection 3.2 we take (for systems)

$$(4a) \quad h \mathbf{F}\{h, x_j; \mathbf{u}_j; \mathbf{f}\} = h \sum_{\mu=0}^n \alpha_\mu \mathbf{f}(\xi_\mu; \eta_\mu);$$

$$(4b) \quad \xi_0 = x_j; \quad \xi_v \equiv \xi_0 + \theta_v h;$$

$$(4c) \quad \eta_0 = \mathbf{u}_j; \quad \eta_v \equiv \eta_0 + h \sum_{k=0}^{v-1} \alpha_{vk} \mathbf{f}(\xi_k; \eta_k),$$

$$v = 1, 2, \dots, n.$$

Here the quantities  $\alpha_\mu$ ,  $\theta_v$ , and  $\alpha_{vk}$  are defined as in (3.20) and (3.21).

As in Section 2 we define the truncation error in predictor-corrector methods applied to a system. That is, an  $r$ -dimensional vector,  $\tau_{i+1}$ , defined by

$$(5a) \quad \mathbf{y}_{i+1} = \mathbf{y}_{i-p} + h \sum_{k=1}^n \alpha_k \mathbf{f}(x_{i+1-k}; \mathbf{y}_{i+1-k})$$

$$+ h \alpha_0 \mathbf{f}(x_{i+1}; \mathbf{y}_{i+1}^*) + h \tau_{i+1}.$$

Here  $\mathbf{y}_{i+1}^* \equiv \mathbf{y}_{i+1} - h \sigma_{i+1}^*$  defines  $\sigma_{i+1}^*$ , and  $\mathbf{y}_{i+1}^*$  is defined by the right side of (2a) with  $\mathbf{u}$  replaced by  $\mathbf{y}$ . We find that

$$(5b) \quad \tau = \sigma + h \alpha_0 \bar{J} \sigma^*,$$

where  $\sigma^*$  and  $\sigma$  have components  $\sigma^{(v)*}$  and  $\sigma^{(v)}$  which are respectively the errors in applying the appropriate quadrature formulae to the integration

of  $f^{(v)}(x, y^{(1)}(x), \dots, y^{(r)}(x))$ . The elements of the matrix  $\bar{J}$  are found by evaluating the corresponding elements of the *Jacobian matrix*

$$(5c) \quad J \equiv (a_{\nu\mu}) = \left( \frac{\partial f^{(\nu)}}{\partial y^{(\mu)}} \right)$$

at appropriate intermediate points. The detailed derivation of (5) is left as an exercise. By using the matrix (5c) we find that the error vectors

$$(6) \quad \mathbf{e}_j \equiv \mathbf{u}_j - \mathbf{y}_j, \quad \mathbf{e}_j^* \equiv \mathbf{u}_j^* - \mathbf{y}_j^*$$

satisfy the systems

$$(7a) \quad \mathbf{e}_{i+1}^* = \mathbf{e}_{i-q} + h \sum_{j=0}^m \beta_j J_{i-j} \mathbf{e}_{i-j};$$

$$(7b) \quad \begin{aligned} \mathbf{e}_{i+1} &= \mathbf{e}_{i-p} + h \sum_{j=1}^n \alpha_j J_{i+1-j} \mathbf{e}_{i+1-j} \\ &\quad + h^2 \alpha_0 J_{i+1} \sum_{j=0}^m \beta_j J_{i-j} \mathbf{e}_{i-j} - h \tau_{i+1}. \end{aligned}$$

Here again the matrices  $J_i$  have as elements the  $a_{\nu\mu}$  of  $J$  evaluated at appropriate intermediate points (the elements in each row of  $J_i$  can be shown to be evaluated at the same point). A convergence proof can now be given exactly as in Subsection 2.1 (see Theorem 2.1) if we employ appropriate vector and matrix norms.

In fact, if the root condition (5.7) is satisfied, we may copy the proof of convergence and the error estimates given in Theorem 5.3 by replacing

$$y, f, u, v, e, F, \rho_k, \tau_k, c_k, E_k, C_k, | |$$

by the corresponding vector quantities

$$\mathbf{y}, \mathbf{f}, \mathbf{u}, \mathbf{v}, \mathbf{e}, \mathbf{F}, \boldsymbol{\rho}_k, \boldsymbol{\tau}_k, \mathbf{c}_k, \mathbf{E}_k, \mathbf{C}_k, \| \|_\infty$$

(i.e., absolute value is replaced by maximum absolute component), for the scheme

$$(8a) \quad \sum_{j=0}^n a_s \mathbf{u}_{j+s} = h \mathbf{F}\{h, x_j; \mathbf{u}_{j-m}, \dots, \mathbf{u}_{j+n}; \mathbf{f}\} + h \boldsymbol{\rho}_{j+n},$$

where  $m \leq j \leq N - n$ , with

$$(8b) \quad \mathbf{u}_0 = \mathbf{y}_0 + \boldsymbol{\rho}_0, \quad \mathbf{u}_1 = \mathbf{y}_1 + \boldsymbol{\rho}_1, \dots, \mathbf{u}_{m+n-1} = \mathbf{y}_{m+n-1} + \boldsymbol{\rho}_{m+n-1}.$$

## PROBLEMS, SECTION 6

- Verify the error estimate corresponding to equation (5.19), as indicated in the last sentence of Section 6, for the scheme defined by (8). That is, show that

$$\|\mathbf{e}_{j+m}\|_\infty \leq 2Q \left[ n \max_{0 \leq k \leq m+n-1} \|\mathbf{e}_k\|_\infty + \frac{\rho(h) + \tau(h)}{2\kappa a_n} \right] \quad \text{for } 0 \leq jh \leq \frac{1}{2\kappa},$$

where

$$\kappa = \frac{QC(m + n + 1)}{a_n}$$

and  $C$ , which appears in the vector analog of (5.3b), is a bound for the vector norm  $\| \cdot \|_\infty$  of  $\mathbf{f}$  and of all its partial derivatives with respect to  $x, y^{(1)}, y^{(2)}, \dots, y^{(r)}$  of some finite order, in the domain  $S: \{(x; \mathbf{y}) \mid a \leq x \leq b, \|\mathbf{y}\|_\infty < \infty\}$ .

2. Verify that if  $\mathbf{u} \equiv (u^{(k)})$ ,  $\mathbf{v} \equiv (v^{(k)})$ , and  $f(x; \mathbf{y})$  has a continuous derivative with respect to all variables, then

$$f(x; \mathbf{u}) - f(x; \mathbf{v}) = \sum_k (u^{(k)} - v^{(k)}) \frac{\partial f}{\partial y^{(k)}}(x; \mathbf{v} + \theta(\mathbf{u} - \mathbf{v}))$$

for some  $\theta$  such that  $0 < \theta < 1$ .

Hence, if  $f$  is replaced by a vector valued function  $\mathbf{f}$ , each component  $f^{(j)}$  of  $\mathbf{f}$  may have its own  $\theta_j$  satisfying  $0 < \theta_j < 1$ .

[Hint: Study  $g(t) \equiv f(x; \mathbf{v} + t(\mathbf{u} - \mathbf{v}))$ . Note that  $g(t) - g(0) = tg'(\theta t)$  for some  $\theta$  in  $0 < \theta < 1$ . Then evaluate

$$\frac{d}{dt} g(t) = \frac{d}{dt} f(x, \mathbf{v} + t(\mathbf{u} - \mathbf{v}))$$

and set  $t = 1$ .] This justifies the definition of  $J_k$  in (7) and  $\bar{J}$  in (5b).

## 7. BOUNDARY VALUE AND EIGENVALUE PROBLEMS

A boundary value problem for an ordinary differential equation (or system) is one in which the dependent variable is required to satisfy specified conditions at more than one point. Since an equation of  $n$ th order has a general solution depending upon  $n$  parameters, the total number of boundary conditions required to determine a unique solution is, in general,  $n$ . However, when the total of  $n$  boundary conditions is given at *more than one point*, it is possible for more than one solution to exist or for no solution to exist. Of course, if more than  $n$  conditions are imposed, even for the initial value problem, there will, in general, be no solution. A detailed study of the existence and uniqueness theory is beyond the scope of our book. However, for linear problems, the theory is well known and we shall indicate here the elements of this theory which may be applicable to non-linear problems and to the analysis of numerical procedures used to solve such boundary value problems.

The simplest linear boundary value problem is one in which the solution of a second order equation, say

$$(1a) \quad y'' - p(x)y' - q(x)y = 0,$$

is specified at two distinct points, say

$$(1b) \quad y(a) = \alpha, \quad y(b) = \beta.$$

The solution  $y(x)$ , is sought in the interval  $a \leq x \leq b$ . A formal approach to the exact solution of the boundary value problem is obtained by considering the related *initial value* problem,

$$(2a) \quad Y'' - p(x)Y' - q(x)Y = 0,$$

$$(2b) \quad Y(a) = \alpha, \quad Y'(a) = s.$$

The theory of solutions of such initial value problems is well known and if, for example, the functions  $p(x)$  and  $q(x)$  are continuous on  $[a, b]$ , the existence of a unique solution of (2) in  $[a, b]$  is assured. Let us denote this solution by

$$Y = Y(s; x),$$

and recall that *every* solution of (1a) or (2a) is a linear combination of two particular “independent” solutions of (1a),  $y^{(1)}(x)$  and  $y^{(2)}(x)$ , which satisfy, say,

$$(3a) \quad y^{(1)}(a) = 1, \quad y^{(1)'}(a) = 0;$$

$$(3b) \quad y^{(2)}(a) = 0, \quad y^{(2)'}(a) = 1.$$

Then the unique solution of (2a) which satisfies (2b) is

$$(4) \quad Y(s; x) = \alpha y^{(1)}(x) + s y^{(2)}(x).$$

Now if we take  $s$  such that

$$(5) \quad Y(s; b) \equiv \alpha y^{(1)}(b) + s y^{(2)}(b) = \beta,$$

then  $y(x) \equiv Y(s; x)$  is a solution of the boundary value problem (1). Clearly, there is at most one root of equation (5),

$$s = \frac{\beta - \alpha y^{(1)}(b)}{y^{(2)}(b)},$$

*provided* that  $y^{(2)}(b) \neq 0$ . If, on the other hand,  $y^{(2)}(b) = 0$  there *may not* be a solution of the boundary value problem (1). A solution would exist in this case only if  $\beta = \alpha y^{(1)}(b)$ , but it would not be unique since then  $Y(s; x)$  of (4) is a solution for arbitrary  $s$ .

Thus there are two mutually exclusive cases for the linear boundary value problem, the so-called *alternative principle*: *either a unique solution exists or else the homogeneous problem (i.e.,  $y(a) = y(b) = 0$ ) has a non-trivial solution* (which is  $s y^{(2)}(x)$  in this example).

These observations permit us to study the solution of the inhomogeneous equation

$$(6) \quad y'' - p(x)y' - q(x)y = r(x),$$

subject to the boundary conditions (1b). This problem can be reduced to

the previous case if a particular solution of (6), say  $y^{(p)}(x)$ , can be found. Then we define

$$(7) \quad w(x) \equiv y(x) - y^{(p)}(x),$$

and find that  $w(x)$  must satisfy the homogeneous equation (1a). The boundary conditions for  $w(x)$  become, from (7) and (1b)

$$w(a) = \alpha - y^{(p)}(a) \equiv \alpha',$$

$$w(b) = \beta - y^{(p)}(b) \equiv \beta'.$$

Thus, we can find the solution of (6) and (1b) by solving (1) with  $(\alpha, \beta)$  replaced by  $(\alpha', \beta')$ . A definite problem for the determination of  $y^{(p)}(x)$  is obtained by specifying particular initial conditions, say

$$(8) \quad y^{(p)}(a) = y^{(p)'}(a) = 0,$$

which provides a standard type of initial value problem for the equation (6). Again, the alternative principle holds; see Problem 1.

The formulation of boundary value problems for linear second order equations can be easily extended to more general  $n$ th order equations or equivalently to  $n$ th order systems of first order equations (not necessarily linear). For example, in the latter case we may consider the system

$$(9a) \quad \mathbf{y}' = \mathbf{f}(x; \mathbf{y}),$$

where we use the row vectors  $\mathbf{y} = (y_1, y_2, \dots, y_n)$ ,  $\mathbf{f} = (f_1, f_2, \dots, f_n)$  and the functions  $f_k \equiv f_k(x; \mathbf{y}) = f_k(x; y_1, \dots, y_n)$  are functions of  $n + 1$  variables. The  $n$  boundary conditions may be, say,

$$(9b) \quad \begin{aligned} y_1(a) &= \alpha_1, & y_2(a) &= \alpha_2, \dots, & y_{m_1}(a) &= \alpha_{m_1}, \\ y_{m_1+1}(b) &= \beta_1, & y_{m_1+2}(b) &= \beta_2, \dots, & y_n(b) &= \beta_{m_2}, \\ m_1 > 0, & m_1 + m_2 = n, & m_2 > 0. \end{aligned}$$

Thus, we specify  $m_1$  quantities at  $x = a$  and the remaining  $n - m_1 = m_2$  quantities at  $x = b$ .

In analogy with (2) we consider the related initial value problem:

$$(10a) \quad \mathbf{Y}' = \mathbf{f}(x, \mathbf{Y});$$

$$Y_i(a) = \alpha_i, \quad i = 1, 2, \dots, m_1$$

$$(10b) \quad Y_{m_1+j}(a) = s_j, \quad j = 1, 2, \dots, m_2.$$

We indicate the dependence on the  $m_2$  arbitrary parameters  $s_j$  by writing

$$Y_k = Y_k(s_1, s_2, \dots, s_{m_2}; x), \quad k = 1, 2, \dots, n.$$

These parameters are to be determined such that

$$(11) \quad Y_{m_1+}(s_1, s_2, \dots, s_{m_2}; b) = \beta_j, \quad j = 1, 2, \dots, m_2.$$

This represents a system of  $m_2$  equations in the  $m_2$  unknowns  $s_j$ . In the corresponding linear case (i.e., in which each  $f_k$  is linear in all the  $y_k$ ) the system (11) becomes a linear system and its solvability is thus reduced to a study of the non-singularity of a matrix of order  $m_2$ .

Note that the alternative principle is again valid. In the general case, however, the roots of a transcendental system (11) are required and the existence and uniqueness theory is more complicated (and in fact, is not as completely developed as it is for the linear case).

We shall examine two different types of numerical methods for approximating the solutions of boundary value problems, in Subsections 7.1 and 7.2.

### 7.1. Initial Value or “Shooting” Methods

The initial value or “shooting” methods attempt to carry out numerically the procedure indicated in equations (2) through (5). That is, roughly, the initial data are adjusted so that the solution of an initial value problem satisfies the required boundary condition at some distant (boundary) point.

We take, for definiteness, a uniform net

$$(12) \quad x_0 = a, \quad x_j = x_0 + jh, \quad j = 0, 1, \dots, N, \quad h = \frac{b - a}{N};$$

and shall try to approximate thereon the solution of the linear equation (6) subject to (1b). We first approximate the solutions  $y^{(1)}(x)$  and  $y^{(2)}(x)$  of the initial value problems (1a) and (3). This can be done, for example, by replacing (1a) by an equivalent first order system and then using a predictor-corrector or one-step method as indicated in Section 6. In the same manner, we can approximate the particular solution  $y^{(p)}(x)$  of (6) and (8). The respective numerical solutions are denoted at each point  $x_j$  of (12) by

$$(13a) \quad u_j^{(1)}, \quad u_j^{(2)}, \quad u_j^{(p)}, \quad j = 0, 1, \dots, N.$$

These solutions satisfy, at  $x_0 = a$ , the conditions

$$(13b) \quad u_0^{(1)} = 1, \quad u_0^{(2)} = 0, \quad u_0^{(p)} = 0.$$

Assume that the same numerical procedure has been used to compute each of these solutions and that we have

$$(14a) \quad e_j^{(1)} \equiv u_j^{(1)} - y^{(1)}(x_j) = \mathcal{O}(h^r),$$

$$(14b) \quad e_j^{(2)} \equiv u_j^{(2)} - y^{(2)}(x_j) = \mathcal{O}(h^r),$$

$$(14c) \quad e_j^{(p)} \equiv u_j^{(p)} - y^{(p)}(x_j) = \mathcal{O}(h^r).$$

[That is, the truncation error of the integration scheme is  $\mathcal{O}(h^r)$ , and the rounding errors are at most  $\mathcal{O}(h^{r+1})$  so that the estimates (14) apply.]

The exact solution of (6) and (1b) is, by the previous analysis, given by

$$(15a) \quad y(x) = y^{(p)}(x) + \alpha y^{(1)}(x) + s y^{(2)}(x), \quad a \leq x \leq b;$$

$$(15b) \quad s = \frac{\beta - y^{(p)}(b) - \alpha y^{(1)}(b)}{y^{(2)}(b)}.$$

Of course, we assume that  $y^{(2)}(b) \neq 0$ . Otherwise the homogeneous problem has a non-trivial solution and then, in general, the boundary value problem has no solution. With the use of (13), we take for the approximate solution the obvious combination

$$(16a) \quad U_j = u_j^{(p)} + \alpha u_j^{(1)} + s_h u_j^{(2)}, \quad j = 0, 1, \dots, N;$$

where

$$(16b) \quad s_h = \frac{\beta - u_N^{(p)} - \alpha u_N^{(1)}}{u_N^{(2)}}.$$

From (13b) and (16b) it clearly follows that, as required,

$$U_0 = \alpha, \quad U_N = \beta,$$

where we have neglected possible roundoff errors in forming  $U_j$  and  $s_h$ . Thus, in principle,  $U_j$  is an approximate solution of the boundary value problem (6) and (1b). In practice, we need only calculate the solution of two initial value problems to evaluate  $U_j$ . That is,  $y^{(p)}(x) + \alpha y^{(1)}(x)$  satisfies (6) and conditions (3a) so that  $u_j^{(p)} + \alpha u_j^{(1)}$  can be computed as the solution of a single initial value problem.

Upon recalling (14), we are led to the obvious, and in fact, correct conclusion that

$$e_j \equiv U_j - y(x_j) = \mathcal{O}(h^r).$$

However, as we now show, there may still be *practical* difficulties in obtaining an accurate approximation.

Upon subtracting (15a) with  $x = x_j$  from (16a) and using the definitions (14), we find

$$(17a) \quad e_j = (e_j^{(p)} + \alpha e_j^{(1)} + s e_j^{(2)}) + (s_h - s) u^{(2)}(x_j), \quad j = 0, 1, \dots, N.$$

Since  $b = x_N$ , (15b) and (16b) imply  $e_N = 0$  and

$$(17b) \quad (s_h - s) = -\frac{e_N^{(p)} + \alpha e_N^{(1)} + s e_N^{(2)}}{u_N^{(2)}}.$$

Use (17b) in (17a) to find

$$(18) \quad e_j = (e_j^{(p)} + \alpha e_j^{(1)} + s e_j^{(2)}) - (e_N^{(p)} + \alpha e_N^{(1)} + s e_N^{(2)}) \frac{u_j^{(2)}}{u_N^{(2)}}.$$

From this expression for the error we see that  $e_0 = e_N = 0$  and thus the error is, in general, small near the endpoints of the interval. However,

whenever  $|u_j^{(2)}/u_N^{(2)}|$  becomes large we may expect relatively large errors. This ratio can be computed and thus a practical assessment of the accuracy in the present method is possible. In particular, note that  $u_N^{(2)} = y^{(2)}(b) + e_N^{(2)}$ . Thus, whenever the fixed number  $y^{(2)}(b)$  is small and opposite in sign to the error  $e_N^{(2)}$ , which depends upon  $h$ , we may find a magnification of the intermediate errors,  $e_j$ .

The effect of roundoff errors in the present method can be very pronounced. While performing the calculations (16), significance is frequently lost when large almost equal quantities are subtracted from each other. This may be due to the occurrence of a small value of  $u_N^{(2)}$ , or to rapidly growing solutions  $y^{(1)}(x)$ ,  $y^{(2)}(x)$  and/or  $y^{(p)}(x)$ .

By using the estimates (14) in (18) we obtain the error bound

$$(19) \quad |U_j - y(x_j)| = |e_j| \leq Mh^r \left( 1 + \left| \frac{u_j^{(2)}}{u_N^{(2)}} \right| \right), \quad j = 1, 2, \dots, N-1.$$

Thus (for sufficiently small net spacing) the error behaves as theoretically expected. In practice, however, it may frequently be necessary to use many significant figures in the calculations to realize these error estimates.

The method (and its attendant difficulties) treated in this subsection is easily extended to more general *linear* boundary value problems.

We can alter the procedure slightly and, with considerably more computing effort, solve non-linear boundary value problems. For example if, in place of (1), the problem is

$$(20) \quad y'' = f(x, y, y'); \quad y(a) = \alpha, \quad y(b) = \beta;$$

we consider the initial value problem [in place of (2)]

$$(21) \quad Y'' = f(x, Y, Y'); \quad Y(a) = \alpha, \quad Y'(a) = s.$$

If  $Y(s; x)$  is the solution of (21) and  $s^*$  is such that

$$(22) \quad Y(s^*; b) = \beta,$$

then  $y(x) = Y(s^*; x)$  is a solution of (20). The equation (22) is, in general, transcendental, whereas in the linear case the corresponding equation, (5), is linear in  $s$ .

The problem of solving (20) is reduced to the determination of the root (or roots) of (22). The root  $s^*$  could be found by applying the iterative methods of Chapter 3. Of course, in each step of such iteration schemes at least one evaluation of  $Y(s; b)$  is required for some value of  $s$ . This may be found only approximately by integrating (21) numerically on some net (12). That is, the net function  $U_j(s)$ ,  $j = 0, 1, \dots, N$ , may be constructed by some method described in earlier sections. Then  $U_j(s)$  is an approximation to  $Y(s; x_j)$ . If the overall error of the integration scheme

is  $\mathcal{O}(h^r)$  and the function  $f(x, y, z)$  is sufficiently smooth, then for each  $s$  we will determine, in fact,

$$U_N(s) = Y(s; b) + \mathcal{O}(h^r).$$

If the solutions of (21) are such that (22) has a simple root,  $s^*$ , and

$$0 < \left| \frac{\partial Y(s; b)}{\partial s} \right| \leq K,$$

for  $|s - s^*| \leq \rho$ , then we can show that a functional iteration procedure will, if  $s_0$  is close enough to  $s^*$ , produce a sequence  $s_0, s_1, \dots$ , such that for some  $k$

$$U_N(s_k) - \beta = Y(s_k; b) - \beta + \mathcal{O}(h^r) = \mathcal{O}(h^r).$$

Hence

$$|s_k - s^*| = \mathcal{O}(h^r).$$

By using sufficiently many iterations, we can thus get within  $\mathcal{O}(h^r)$  of a root of (22) and hence compute a solution of (20) to within an error bounded by  $Mh^r$ . In Problems 4 and 5, we indicate some of the details of these results.

It is convenient for the application of Newton's iterative method in solving (22) to approximate  $\partial Y(s; b)/\partial s$ . By differentiating (21) with respect to  $s$ , we can formally find the differential equation, called the *variational equation*, i.e., satisfied by the function  $W(s; x) \equiv \partial Y(s; x)/\partial s$ :

$$W'' = \frac{\partial f(x, Y, Y')}{\partial y} W + \frac{\partial f(x, Y, Y')}{\partial z} W';$$

$$W(s; a) = 0, \quad W'(s; a) = 1.$$

A numerical approximation to the solution of the variational equation may be computed stepwise along with the evaluation of  $U_j(s)$ . Hence for  $j = N$  we would have an approximation for both  $Y(s; b)$  and  $\partial Y(s; b)/\partial s$ .

## 7.2. Finite Difference Methods

We consider here finite difference methods which are not based on solving the initial value problem. These are called *direct methods*. The truncation error of the particular difference method we use is  $\mathcal{O}(h^2)$  and the labor required for a given accuracy is comparable to that for the initial value method of some low order.

Let the boundary value problem be (6) and (1b) which we write as

$$(23a) \quad L\{y\} \equiv y'' - p(x)y' - q(x)y = r(x);$$

$$(23b) \quad y(a) = \alpha, \quad y(b) = \beta.$$

We impose here the restriction that

$$(24) \quad q(x) \geq Q_* > 0, \quad a \leq x \leq b.$$

The most simple existence and uniqueness proofs for solutions of boundary value problems of the form (23) require such a condition but with  $Q_* \geq 0$ . We assume a unique solution of (23) to exist with four continuous derivatives in  $a \leq x \leq b$ . A uniform net will be used with  $h = (b - a)/(N + 1)$ .

Now rather than seek high order accuracy in a difference approximation of (23a) we use the simple difference equations

$$(25a) \quad L_h\{u_j\} \equiv \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} - p(x_j) \frac{u_{j+1} - u_{j-1}}{2h} - q(x_j)u_j = r(x_j), \quad j = 1, 2, \dots, N.$$

The boundary conditions are replaced by

$$(25b) \quad u_0 = \alpha, \quad u_{N+1} = \beta.$$

Multiply (25a) by  $-h^2/2$  to obtain

$$-\frac{h^2}{2} L_h\{u_j\} = -b_j u_{j-1} + a_j u_j - c_j u_{j+1} = -\frac{h^2}{2} r(x_j), \quad j = 1, 2, \dots, N,$$

where

$$(26) \quad \begin{aligned} a_j &\equiv 1 + \frac{h^2}{2} q(x_j), & b_j &\equiv \frac{1}{2} \left[ 1 + \frac{h}{2} p(x_j) \right], \\ c_j &\equiv \frac{1}{2} \left[ 1 - \frac{h}{2} p(x_j) \right]. \end{aligned}$$

Using this notation the system of difference equations (25a) and boundary conditions (25b) can be written in the vector form

$$(27a) \quad A\mathbf{u} = \mathbf{r},$$

where

$$(27b) \quad \mathbf{u} \equiv \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_N \end{pmatrix}, \quad \mathbf{r} \equiv -\frac{h^2}{2} \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_N \end{pmatrix} + \begin{pmatrix} b_1\alpha \\ 0 \\ \vdots \\ 0 \\ c_{N+1}\beta \end{pmatrix},$$

$$A \equiv \begin{pmatrix} a_1 & -c_1 & & & & & \\ -b_2 & a_2 & -c_2 & & & & 0 \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -b_{N-1} & a_{N-1} & -c_{N-1} & \\ 0 & & & & -b_N & a_N & \end{pmatrix}.$$

Thus to solve the difference problem (25) we must, in fact, solve the  $N$ th order linear system (27a) with tridiagonal coefficient matrix,  $A$ , given in (27b).

Let us require that the net spacing  $h$  be so small that

$$(28) \quad \frac{h}{2} |p(x_j)| \leq 1, \quad j = 1, 2, \dots, N.$$

Then from (26) it follows that

$$|b_j| + |c_j| = b_j + c_j = 1,$$

while (24) implies  $a_j > 1$ . So we deduce that

$$\begin{aligned} |a_1| &> |c_1|; \\ |a_j| &> |b_j| + |c_j|, \quad 2 \leq j \leq N - 1; \\ |a_N| &> |b_N|; \end{aligned}$$

and hence Theorem 3.5 of Chapter 2 applies. The solution of (27a) can thus be computed by the simple direct factorization of  $A$  described in Section 3 of Chapter 2. Of course, this furnishes a proof of the existence of a unique solution of the difference equations (27) provided (28) is satisfied.

Let us now estimate the error in the numerical approximation defined above. The local truncation errors,  $\tau_j$ , are defined by:

$$(29) \quad L_h\{y(x_j)\} = r(x_j) + \tau_j; \quad j = 1, 2, \dots, N.$$

Since  $y(x)$  is a solution of (1a) we have, assuming  $y^{iv}(x)$  to be continuous,

$$\begin{aligned} (30) \quad \tau_j &= L_h\{y(x_j)\} - L\{y(x_j)\} \\ &= \left[ \frac{y(x_j - h) - 2y(x_j) + y(x_j + h)}{h^2} - y''(x_j) \right] \\ &\quad - p(x_j) \left[ \frac{y(x_j + h) - y(x_j - h)}{2h} - y'(x_j) \right] \\ &= \frac{h^2}{12} [y^{iv}(\xi_j) - 2p(x_j)y''(\eta_j)]; \quad j = 1, 2, \dots, N. \end{aligned}$$

Here  $\xi_j$  and  $\eta_j$  are in  $[x_{j-1}, x_{j+1}]$  and we have used Taylor's theorem.

The basic *error estimate* can now be stated as

**THEOREM 1.** *If the net spacing,  $h$ , satisfies (28) then*

$$(31a) \quad |u_j - y(x_j)| \leq h^2 \left( \frac{M_4 + 2P^* M_3}{12Q_*} \right), \quad j = 0, 1, \dots, N + 1;$$

where  $y(x)$  is the solution of (6) and (1b),  $\{u_j\}$  is the solution of (25) and

$$(31b) \quad P^* \equiv \max_{[a, b]} |p(x)|, \quad M_3 \equiv \max_{[a, b]} |y'''(x)|,$$

$$M_4 \equiv \max_{[a, b]} |y^{iv}(x)|.$$

*Proof.* Let us define

$$e_j = u_j - y(x_j), \quad j = 0, 1, \dots, N + 1.$$

Then subtracting (29) from (25a) yields, with the aid of (26),

$$(32) \quad a_j e_j = b_j e_{j-1} + c_j e_{j+1} + \frac{h^2}{2} \tau_j, \quad j = 1, 2, \dots, N.$$

Now with the norms

$$e \equiv \max_{0 \leq j \leq N+1} |e_j|, \quad \tau \equiv \max_{1 \leq j \leq N} |\tau_j|,$$

we obtain by taking absolute values in (32) and using the equation after (28)

$$|a_j e_j| \leq e + \frac{h^2}{2} \tau, \quad j = 1, 2, \dots, N.$$

However, by condition (24),  $|a_j| = a_j \geq 1 + (h^2/2)Q_*$  and so the above implies that

$$\left(1 + \frac{h^2}{2} Q_*\right) |e_j| \leq e + \frac{h^2}{2} \tau, \quad j = 1, 2, \dots, N.$$

From (23b) and (25b) we have  $e_0 = e_{N+1} = 0$  and so the above inequality is valid for all  $j$  in  $0 \leq j \leq N + 1$ . Thus we conclude that

$$e \leq \frac{1}{Q_*} \tau.$$

Finally, by using the quantities (31b) in (30) we find that

$$\tau \leq \frac{h^2}{12} (M_4 + 2P^*M_3)$$

and the theorem follows. ■

From Theorem 1 we see that the difference solution converges to the exact solution as  $h \rightarrow 0$  and, in fact, that the error is at most  $\mathcal{O}(h^2)$ . For equations in which  $p(x) \equiv 0$ , error bounds that are  $\mathcal{O}(h^4)$  are easily obtained by using a slight modification of (25a) (see Problem 6). Boundary conditions more general than those in (23b) can be treated with no essential change in these results (see Problem 7). The condition (24) can be relaxed to  $q(x) \geq 0$  and a somewhat more involved argument yields a result analogous to that of Theorem 1. These arguments are based on a so-called *maximum*

*principle* (see Problem 9). The use of this maximum principle is demonstrated in Problem 10.

The effects of roundoff in computing the solution of (25) can be estimated. In fact, let  $U_j$  be the computed quantities which, in place of (25), satisfy

$$(33a) \quad -\frac{h^2}{2} L_h\{U_j\} = -\frac{h^2}{2} r(x_j) + \rho_j, \quad j = 1, 2, \dots, N;$$

and the boundary conditions

$$(33b) \quad U_0 = \alpha + \rho_0, \quad U_{N+1} = \beta + \rho_{N+1}.$$

The quantities  $\rho_j$  represent the local roundoff errors committed in each of the indicated computations. Now we define

$$E_j \equiv U_j - y(x_j), \quad j = 0, 1, \dots, N+1$$

and exactly as in the proof of Theorem 1 we deduce that

$$\left(1 + \frac{h^2}{2} Q_*\right) |E_j| \leq E + \frac{h^2}{2} \tau + \rho, \quad j = 1, 2, \dots, N.$$

Here

$$\begin{aligned} \rho &\equiv \max_{0 \leq j \leq N+1} |\rho_j| \text{ and } |E_0| = |\rho_0|, \quad |E_{N+1}| = |\rho_{N+1}|, \\ E &\equiv \max_{0 < j \leq N+1} |E_j|. \end{aligned}$$

If, in addition to (28), we require that  $h^2 Q_*/2 \leq 1$ , then this inequality is also valid for  $j = 0$ , and  $j = N+1$  so we finally obtain

$$E \leq \frac{1}{Q_*} \left( \tau + 2 \frac{\rho}{h^2} \right).$$

Thus for sufficiently small net spacing,  $h$ , we have

$$(34) \quad |U_j - y(x_j)| \leq h^2 \left( \frac{M_4 + 2P^* M_3}{12Q_*} \right) + \frac{1}{h^2} \left( \frac{2\rho}{Q_*} \right),$$

$$j = 0, 1, \dots, N+1.$$

The roundoff affects this estimate somewhat differently than it did the corresponding estimates in Subsections 1.2 and 1.3, etc. Now to have an error bound which is  $\mathcal{O}(h^2)$  we must limit the roundoff by  $\rho = \mathcal{O}(h^4)$  as  $h \rightarrow 0$ . That is, two orders in  $h$  improvement over the local truncation error are required. Previously, only one additional order in  $h$  was required, since our difference equations were then approximations to first order differential equations (or systems of equations).

Difference methods can also be applied to fairly general non-linear second order boundary value problems. While such methods accurate to

$\mathcal{O}(h^2)$  can be determined, the difference equations are now no longer linear. Hence iterations are employed to solve these equations. It should be observed that the iterations are not employed in order to satisfy the correct boundary conditions, as would be the case in the initial value methods. The construction of iteration procedures for solving the difference equations is quite simple.

The non-linear boundary value problems we consider are of the form (20) where the function  $f(x, y, z)$  is assumed to satisfy the conditions

$$(35) \quad 0 < Q_* \leq \frac{\partial f(x, y, z)}{\partial y} \leq Q^*, \quad \left| \frac{\partial f(x, y, z)}{\partial z} \right| \leq P^*;$$

in some sufficiently large region. Furthermore, these partial derivatives and  $y^{iv}(x)$  are assumed to be continuous.

Again we use a uniform net. On this net the difference approximation of (20) is taken to be

$$(36a) \quad \frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = f\left(x_j, u_j, \frac{u_{j+1} - u_{j-1}}{2h}\right), \\ j = 1, 2, \dots, N;$$

$$(36b) \quad u_0 = \alpha, \quad u_{N+1} = \beta.$$

The local truncation error,  $\tau_j$ , of this method is defined in the usual manner by

$$(37) \quad \frac{y_{j-1} - 2y_j + y_{j+1}}{h^2} = f\left(x_j, y_j, \frac{y_{j+1} - y_{j-1}}{2h}\right) + \tau_j, \\ j = 1, 2, \dots, N.$$

From the assumed continuity properties of  $\partial f / \partial z$  and  $y^{iv}(x)$  it follows that

$$(38) \quad \tau_j = \frac{h^2}{12} \left[ y^{iv}(\xi_j) - 2 \frac{\partial f(x_j, y_j, y'(\zeta_j))}{\partial z} y'''(\eta_j) \right], \quad j = 1, 2, \dots, N.$$

Here,  $\xi_j$  and  $\eta_j$  in  $[x_{j-1}, x_{j+1}]$  are the appropriate mean values used in Taylor's theorem.

To examine the convergence of this procedure we introduce  $e_j \equiv u_j - y(x_j)$  and, for the further applications of Taylor's theorem,

$$(39) \quad p_j \equiv \frac{\partial f}{\partial z} \left( x_j, y_j + \theta_j e_j, \frac{y(x_{j+1}) - y(x_{j-1})}{2h} + \theta_j \frac{e_{j+1} - e_{j-1}}{2h} \right); \\ q_j \equiv \frac{\partial f}{\partial y} \left( x_j, y_j + \theta_j e_j, \frac{y(x_{j+1}) - y(x_{j-1})}{2h} + \theta_j \frac{e_{j+1} - e_{j-1}}{2h} \right); \\ 0 < \theta_j < 1.$$

Then subtracting (37) from (36a) we get, with the above notation and appropriate values for the  $\theta_j$ ,

$$(40) \quad \left(1 + \frac{h^2}{2} q_j\right) e_j = \frac{1}{2} \left(1 + \frac{h}{2} p_j\right) e_{j-1} + \frac{1}{2} \left(1 - \frac{h}{2} p_j\right) e_{j+1} + \frac{h^2}{2} \tau_j, \\ j = 1, 2, \dots, N.$$

This system of equations is formally identical to that in (32), with the same boundary conditions,  $e_0 = e_{N+1} = 0$ . So we may conclude, by using (35) in place of (24), exactly as in the proof of Theorem 1 that

$$(41) \quad |e_j| \leq h^2 \frac{M_4 + 2P^*M_3}{12Q_*}.$$

Here  $P^*$  is defined in (35) and  $M_3$  and  $M_4$  are the appropriate bounds on the derivatives of the solution of (20). Thus the order of convergence for the non-linear problem is the same as that for the linear case; the constants in (41) have only slightly different meanings from those in (31). The non-linear cases for which the difference method is applicable can be generalized as are the linear cases in Problems 7 and 8.

If  $f(x, y, z)$  is not a linear function of  $y$  and  $z$ , then the difference equations (36) constitute a non-linear system of equations. The general methods of Chapter 3 could be applied in order to solve such systems. In particular, Newton's method is frequently well suited for this purpose, and in special cases the convergence proof given in Subsection 3.2 of Chapter 3 can be applied. However, due to the special structure of this system some other iteration schemes are naturally suggested, and we shall consider one of them here. All of these methods proceed from an initial estimate of the solution, say

$$u_j^{(0)}, \quad j = 1, 2, \dots, N; \quad u_0^{(0)} = \alpha, \quad u_{N+1}^{(0)} = \beta.$$

A particularly simple iteration scheme for solving (36) is defined by:

$$(42a) \quad (1 + \omega) u_j^{(v+1)} = \frac{1}{2} (u_{j-1}^{(v)} + u_{j+1}^{(v)}) + \omega u_j^{(v)} \\ - \frac{h^2}{2} f\left(x_j, u_j^{(v)}, \frac{u_{j+1}^{(v)} - u_{j-1}^{(v)}}{2h}\right), \\ j = 1, 2, \dots, N;$$

$$(42b) \quad u_0^{(v+1)} = \alpha, \quad u_{N+1}^{(v+1)} = \beta.$$

Here  $\omega$  is a parameter to be determined so that the iterates converge. In fact, we can show, see Problem (11), that if  $\omega$  satisfies

$$(42c) \quad \omega \geq \frac{h^2}{2} Q^*$$

then the iterates satisfy

$$(43) \quad |u_j^{(v+1)} - u_j^{(v)}| \leq \left(1 - \frac{h^2 Q_*}{2(1 + \omega)}\right)^v \max_k |u_k^{(1)} - u_k^{(0)}|; \\ j = 1, 2, \dots, N.$$

From this result we see that the iterates form a Cauchy sequence. Thus not only do they converge but by the assumed continuity of  $f(x, y, z)$  we can show, exactly as in the proof of Theorem 1.1 in Chapter 3, that a *unique* solution of the difference equations (36) exists.

### 7.3. Eigenvalue Problems

We have shown previously that a linear boundary value problem may have non-unique solutions. In fact, this occurs if and only if the corresponding homogeneous boundary value problem has a non-trivial solution. If the coefficients of the homogeneous equation depend upon some parameter it is frequently of interest to determine the values of the parameter for which such non-trivial solutions exist. These special parameter values are called *eigenvalues* and the corresponding non-trivial solutions are called *eigenfunctions*. The simplest example is furnished by the homogeneous problem

$$y'' + \lambda y = 0; \quad y(a) = y(b) = 0.$$

For each of the parameter values

$$\lambda = \lambda_n \equiv \left[ \frac{n\pi}{b-a} \right]^2, \quad n = 1, 2, \dots;$$

there exists a non-trivial solution

$$y(x) = y_n(x) \equiv \sin \lambda_n^{1/2}(x - a), \quad n = 1, 2, \dots$$

A fairly general class of eigen-problems, which includes many of the cases that occur in applied mathematics, are the Sturm-Liouville problems,

$$(44a) \quad L\{y\} + \lambda r(x)y \equiv [p(x)y']' - q(x)y + \lambda r(x)y = 0,$$

$$(44b) \quad \alpha_0 y'(a) - \alpha_1 y(a) = 0, \quad \beta_0 y'(b) + \beta_1 y(b) = 0.$$

Here  $p(x) > 0$ ,  $r(x) > 0$ , and  $q(x) \geq 0$ ;  $p'(x)$ ,  $q(x)$ , and  $r(x)$  are continuous on  $[a, b]$ ; and the constants  $\alpha_v$  and  $\beta_v$  are non-negative and at least one of each pair does not vanish. It is known that for such problems there exists an infinite sequence of non-negative eigenvalues

$$(45) \quad 0 \leq \lambda_1 < \lambda_2 < \lambda_3 \dots$$

In addition, there exist corresponding eigenfunctions,  $y_n(x)$ , which satisfy the orthogonality relations

$$\int_a^b y_n(x) y_m(x) r(x) dx = \delta_{nm},$$

and the  $n$ th eigenfunction has  $n - 1$  distinct zeros in  $a < x < b$ .

We may again relate the solution of (44) to an initial value problem. For any fixed  $\lambda$  we consider

$$(46a) \quad L\{Y\} + \lambda r(x) Y = 0;$$

$$(46b) \quad \alpha_0 Y'(a) - \alpha_1 Y(a) = 0, \quad \gamma_0 Y'(a) - \gamma_1 Y(a) = 1.$$

Here  $\gamma_0$  and  $\gamma_1$  are any constants such that  $(\alpha_1 \gamma_0 - \alpha_0 \gamma_1) \neq 0$ . Then the two initial conditions in (46b) are linearly independent and a unique non-trivial solution of the initial value problem (46) exists. We denote this solution by  $Y(\lambda; x)$ . Now we consider the equation

$$(47) \quad \Phi(\lambda) \equiv \beta_0 Y'(\lambda; b) + \beta_1 Y(\lambda; b) = 0.$$

Clearly, each eigenvalue  $\lambda_n$  in (45) must satisfy this equation. Also every zero,  $\lambda^*$ , of  $\Phi(\lambda)$  is an eigenvalue of (44) and the corresponding solution  $Y(\lambda^*; x)$  of (46) is a corresponding eigenfunction of (44). Note that the present analysis differs from the corresponding discussion at the beginning of Section 7. Here, a parameter in the equation must be adjusted while the adjoined initial condition remains fixed, which reverses the previous situation. Of course, the present considerations apply to eigenvalue problems more general than those in (44); say for instance to problems in which the eigenvalue parameter  $\lambda$  enters into all of the coefficients of the equation and the boundary conditions. Extensions to homogeneous systems, of, say  $m$  second-order equations with  $m$  parameters are also clearly suggested. The initial value procedure can actually be used to prove the existence of the eigenvalues (45) and the oscillation properties of the eigenfunctions.

To approximate the eigenvalues and eigenfunctions for problems of the form (44), and various generalizations of these problems, we may apply numerical methods which are exactly analogous to those used in subsections 7.1 and 7.2. However, the proofs of convergence and estimates of the errors are now not always as easy to obtain as they were for those boundary value problems.

Some approximation methods for eigenvalue problems are based on *variational principles*. These have led to the construction of useful numerical methods. However, we do not treat them here, but refer the reader to the brief discussion of variational principles in Subsection 1.2 of Chapter 9.

A simple application of the basic error estimate for an eigenvalue of a symmetric matrix, Theorem 1.5 of Chapter 4, can be used to give an error estimate for the eigenvalue of a differential equation that is approximated by a difference method (e.g., the method in Subsection 7.2). Consider the eigenvalue problem

$$(48) \quad L\{y\} = \lambda y; \quad y(a) = y(b) = 0,$$

where  $L\{\cdot\}$  is defined in (44a).

Assume that  $\lambda$  is an eigenvalue and  $y(x)$  a corresponding eigenfunction, with a continuous fourth derivative. Let

$$(49) \quad L_h\{u\} = \Lambda u; \quad u(a) = u(b) = 0,$$

be a finite difference approximation to (48), on the net (12). Assume that the matrix form of (49), analogous to (27), is

$$(50) \quad A\mathbf{u} = -\frac{h^2}{2} \Lambda \mathbf{u},$$

where  $A$  is a *symmetric* matrix. Then the truncation error,  $\tau$ , of the eigen-solution is defined by

$$Ay + \frac{h^2}{2} \lambda y \equiv -\frac{h^2}{2} \tau.$$

If  $\|\tau\|_\infty \leq Mh^2$ , when  $\|y\|_\infty = 1$ , then  $\|\tau\|_2 \leq MN^{1/2}h^2$ . Furthermore, Theorem 1.5 of Chapter 4 implies

$$\min_{1 \leq j \leq N} \frac{h^2}{2} |\lambda - \Lambda_j| \leq \frac{h^4}{2} MN^{1/2},$$

whence we have shown,

### THEOREM 2.

$$\min_{1 \leq j \leq N} |\lambda - \Lambda_j| \leq h^2 MN^{1/2} = \mathcal{O}(h^{3/2}). \quad \blacksquare$$

Theorem 2 states that some eigenvalue,  $\Lambda_j$ , of the discrete problem (49) is a good approximation to a given eigenvalue  $\lambda$  of (48). But as  $h \rightarrow 0$ , the theorem fails to identify which eigenvalue  $\Lambda_j$  is the closest approximation. In Problem 14, we verify that, in a special case, the smallest eigenvalues  $\Lambda_j$  approximate respectively the lowest eigenvalues  $\lambda_j$ .

## PROBLEMS, SECTION 7

- Establish the *alternative principle*. Either the equations (6) and (1b) have a unique solution or else the homogeneous problem [i.e.,  $r(x) \equiv 0, \alpha = \beta = 0$ ] has a non-trivial solution.

**2. Solve by the initial value method**

$$y'' = -100y; \quad y(0) = 1, \quad y(2\pi + \epsilon) = 1.$$

Use

$$y^{(1)}(x) \equiv \cos 10x, \quad y^{(2)}(x) = \frac{\sin 10x}{10}.$$

For small  $\epsilon$ , show that  $s = 50\epsilon + \mathcal{O}(\epsilon^3)$ . Explain why the computational scheme corresponding to the initial value method would be difficult to apply for small  $\epsilon$ .

**3. Solve by the initial value method**

$$y'' = 100y; \quad y(0) = 1, \quad y(3) = e^{-30}.$$

Use

$$y^{(1)}(x) = \frac{e^{10x} + e^{-10x}}{2},$$

$$y^{(2)}(x) = \frac{e^{10x} - e^{-10x}}{20}.$$

Explain why the computational scheme of the initial value method would have to be applied with great care.

**4. The chord method for approximating the root  $s^*$  of (22) is based on the iteration scheme**

$$s_{k+1} = g(s_k),$$

where

$$g(s) \equiv s - m[Y(s; b) - \beta].$$

Show that if for some  $\rho > 0$ ,

$$0 < L \leq \left| \frac{\partial Y}{\partial s}(s; b) \right| \leq K, \quad \text{for } |s - s^*| \leq \rho,$$

then with

$$m = \frac{2}{L + K} \operatorname{sign} \left( \frac{\partial Y(s; b)}{\partial s} \right),$$

$$|g'(s)| \leq \frac{K - L}{K + L} < 1.$$

**5. Let the approximate solution of (21) be  $U_j(s)$ ,  $0 \leq j \leq N$ ; and assume that, in the notation of Problem 4,**

$$m|U_N(s) - Y(s; b)| \leq \delta = \mathcal{O}(h') \quad \text{for } |s - s^*| \leq \rho.$$

Define  $\lambda \equiv (K - L)/(K + L)$  and let  $h$  be small enough so that  $\delta \leq (1 - \lambda)\rho/2$ . Use Theorem 1.3 of Chapter 3 and Problem 4 to show that, with  $\sigma_{k+1} = \sigma_k - m[U_N(\sigma_k) - \beta]$ , then

$$|\sigma_k - s^*| \leq \frac{\delta}{1 - \lambda} + \lambda^k \left( \rho - \frac{2\delta}{1 - \lambda} \right),$$

if

$$|\sigma_0 - s^*| \leq \rho - \frac{\delta}{1 - \lambda}.$$

6. For the boundary value problem:  $y'' - q(x)y = r(x)$ ;  $y(a) = y(b) = 0$  use a difference scheme of the form:

$$\begin{aligned} [u_{j+1} - 2u_j + u_{j-1}] / h^2 - [\alpha_1 q(x_{j+1}) u_{j+1} + \alpha_0 q(x_j) u_j \\ + \alpha_{-1} q(x_{j-1}) u_{j-1}] = [\alpha_1 r(x_{j+1}) + \alpha_0 r(x_j) + \alpha_{-1} r(x_{j-1})], \end{aligned}$$

for  $j = 1, 2, \dots, N$ , with  $u_0 = u_{N+1} = 0$  (as usual  $h = (b - a)/(N + 1)$ ).

(a) Determine  $\alpha_0, \alpha_1, \alpha_{-1}$  such that the truncation error is  $\mathcal{O}(h^4)$ . We assume here that  $y^{iv}$ ,  $q^{iv}$ , and  $r^{iv}$  are continuous. Note that for the solution  $y(x)$  we have  $y^{iv} - [q(x)y]'' = r''(x)$ .

(b) If  $q(x) \geq Q_* > 0$ , then show that for sufficiently small  $h$ :

$$|u_j - y(x_j)| \leq \frac{h^4}{720} \frac{2M_6 + 5N_4 + 5R_4}{Q_*}$$

where  $M_6 \equiv \max_{[a, b]} |y^{vi}(x)|$ ,  $N_4 \equiv \max_{[a, b]} |[q(x)y(x)]^{iv}|$ ,  $R_4 \equiv \max_{[a, b]} |r^{iv}(x)|$ .

The proof is just as in Theorem 1.

7. Consider the boundary value problem

$$\begin{aligned} y'' - p(x)y' - q(x)y = r(x); \quad \alpha_0 y'(a) - \alpha_1 y(a) = \alpha, \\ \beta_0 y'(b) + \beta_1 y(b) = \beta \end{aligned}$$

where  $\alpha_0, \beta_0, \alpha_1$  and  $\beta_1$  are all positive. Use the difference equations

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} - p(x_j) \frac{u_{j+1} - u_{j-1}}{2h} - q(x_j)u_j = r(x_j) \quad \text{for } j = 0, 1, \dots, N+1$$

and the "boundary" conditions

$$\alpha_0 \left( \frac{u_1 - u_{-1}}{2h} \right) - \alpha_1 u_0 = \alpha, \quad \beta_0 \left( \frac{u_{N+2} - u_N}{2h} \right) - \beta_1 u_{N+1} = \beta.$$

[Note: Values at  $x_{-1} = a - h$  and  $x_{N+2} = b + h$  have been introduced and the difference approximations of the differential equation have been employed at  $x_0 = a$  and at  $x_{N+1} = b$ . Hence the values  $x_{-1}$  and  $x_{N+2}$  can be eliminated from the above difference equations.]

(a) Write these difference equations as a system of order  $N + 2$  in the form (27). If the tridiagonal coefficient matrix is

$$A \equiv \begin{pmatrix} & & & & \\ & \cdot & & & \\ & & \cdot & & \\ & & & \ddots & \\ & & & & \cdot \\ -B_j & & A_j & & -C_j \\ & \cdot & & \cdot & \\ & & \cdot & & \cdot \end{pmatrix}$$

with  $j = 0, 1, \dots, N + 1$ , show that from (26),

$$A_j = a_j, \quad B_j = b_j, \quad C_j = c_j \quad \text{for } j = 1, 2, \dots, N$$

and that

$$A_0 = \left( a_0 + 2h \frac{\alpha_0}{\alpha_1} b_0 \right), \quad C_0 = (c_0 + b_0).$$

Find similar expressions for  $A_{N+1}$  and  $B_{N+1}$ .

(b) If  $q(x) \geq Q_* > 0$  and the solution  $y(x)$  is sufficiently smooth in an open interval containing  $[a, b]$  show that for sufficiently small  $h$

$$|u_j - y(x_j)| \leq \mathcal{O}(h^2), \quad j = 0, 1, \dots, N+1.$$

8.\* Consider the boundary value problem:

$$[a(x)y']' - p(x)y' - q(x)y = r(x); \quad y(a) = y(b) = 0$$

and the corresponding difference problem:

$$\begin{aligned} & \left[ a\left(x_j + \frac{h}{2}\right)\left(\frac{u_{j+1} - u_j}{h^2}\right) - a\left(x_j - \frac{h}{2}\right)\left(\frac{u_j - u_{j-1}}{h^2}\right) \right] \\ & \quad - p(x_j)\left(\frac{u_{j+1} - u_{j-1}}{2h}\right) - q(x_j)u_j = r(x_j), \\ & \quad j = 1, 2, \dots, N; \quad u_0 = u_{N+1} = 0. \end{aligned}$$

(a) If  $y^{IV}$  and  $a''$  are continuous, show that the truncation error in this scheme is  $\mathcal{O}(h^2)$ .

(b) If  $q(x) \geq Q_* > 0$  and  $A^* \geq a(x) \geq A_* > 0$ , show that

$$|u_j - y(x_j)| \leq \frac{A^*}{A_* Q_*} \tau$$

provided  $A_* - (h/2)|p(x_j)| \geq 0$  for  $j = 1, 2, \dots, N$

[Hint: Proceed as in the proof of Theorem 1 but now divide by  $|b_j| + |c_j| = b_j + c_j \geq 2A_*$  before bounding the coefficients.]

9. We define the difference operator  $T$  by

$$Tu_j \equiv a_j u_j - b_j u_{j-1} - c_j u_{j+1}, \quad j = 1, 2, \dots, N,$$

where:

$$b_j > 0, \quad c_j > 0, \quad a_j \geq b_j + c_j.$$

Prove the

**MAXIMUM PRINCIPLE:** Let the net function  $\{V_j\}$  satisfy  $TV_j \leq 0, j = 1, 2, \dots, N$ . Then

$$\max_{0 \leq j \leq N+1} V_j = \max \{V_0, V_{N+1}\}.$$

Conversely if  $TV_j \geq 0, j = 1, 2, \dots, N$ ; then

$$\min_{0 \leq j \leq N+1} V_j = \min \{V_0, V_{N+1}\}.$$

[Hint: Use contradiction; assume  $\max V_j \equiv M$  is at  $V_k$  for some  $k$  in  $1 \leq k \leq N$  but that  $V_0 \neq M$  and  $V_{N+1} \neq M$ . Then conclude that  $V_j = M$  for all  $j$  which is a contradiction. The minimum result follows by changing sign.]

[Note: The conditions on the coefficients in  $T$  are satisfied by the quantities in (26) provided (28) is satisfied even if we allow  $q(x) = 0$  (i.e., if condition (24) is weakened)].

**10.** Let  $T$  be as in Problem 9 and  $\{e_j\}$  satisfy

$$Te_j = \sigma_j, \quad j = 1, 2, \dots, N.$$

Suppose  $\{g_j\}$  satisfies  $g_j > 0$  and

$$Tg_j \geq 1.$$

Then prove that

$$|e_j| \leq \max_{v=0, N+1} (|e_v| + \sigma g_v) + \sigma \cdot g,$$

where  $\sigma = \max |\sigma_j|$ ,  $g = \max |g_j|$ .

[Hint: Form  $\omega_j \equiv e_j \pm \sigma g_j$  and apply the maximum principle.]

**11.\*** Prove that (43) follows from (42).

[Hint: Subtract (42a) from the corresponding equation with  $v + 1$  replaced by  $v$ ; use Taylor's theorem and proceed as in the derivation of (41).]

**12.\*** Consider, in place of (36), the difference equations  $u_0 = u_{N+1} = 0$ ;

$$\frac{u_{j+1} - 2u_j + u_{j-1}}{h^2} = f\left(x_j, \frac{u_{j+1} + u_{j-1}}{2}, \frac{u_{j+1} - u_{j-1}}{2h}\right), \quad j = 1, 2, \dots, N.$$

(a) Show that  $|u_j - y(x_j)| = \mathcal{O}(h^2)$  where  $y(x)$  is the solution of (20); (28) and (35) hold; and  $y^{(v)}(x)$ ,  $\partial f/\partial y$  and  $\partial f/\partial y'$  are continuous.

(b) Under the above assumptions and  $h(P^* + hQ^*)/2 \leq 1$ , prove convergence of the iterations:

$$u_0^{(v+1)} = u_{N+1}^{(v+1)} = 0,$$

$$u_j^{(v+1)} = \frac{1}{2}[u_{j+1}^{(v)} + u_{j-1}^{(v)}] - \frac{h^2}{2} f\left(x_j, \frac{u_{j+1}^{(v)} + u_{j-1}^{(v)}}{2}, \frac{u_{j+1}^{(v)} - u_{j-1}^{(v)}}{2h}\right);$$

$$j = 1, 2, \dots, N.$$

Note that the parameter  $\omega$  is not required here, as it was in (42); i.e., we could employ the value  $\omega = 0$ .

**13.** Solve for the eigenvalues and eigenvectors of the problem  $y'' + \lambda y = 0$ ,  $y'(a) = y'(b) = 0$ , by using the initial value technique. For example, use the initial values  $y'(a) = 0$ ,  $y(a) = \text{constant} \neq 0$ .

**14.** Find the eigenvalues and eigenvectors of the scheme

$$\frac{u_{j-1} - 2u_j + u_{j+1}}{h^2} = \Lambda u_j, \quad 1 \leq j \leq N,$$

$$u_0 = u_{N+1} = 0, \quad h = \frac{\pi}{N+1}.$$

Compare them with the eigenvalues and eigenvectors of

$$y'' = \lambda y, \quad y(0) = y(\pi) = 0.$$

[Hint: Solve the difference equation in the form  $u_j = \alpha^j$ , for an appropriate  $\alpha$ . Show that the eigenfunctions are

$$u_j^{(k)} = A_k \sin j \left( \frac{\pi k}{N+1} \right), \quad 0 \leq j \leq N+1,$$

for  $k = 1, 2, \dots, N$ .]

**15.** (a) Use the results of Problems 6 and 14 to devise a difference scheme for  $y'' + \lambda y = 0$ ,  $y(0) = y(1) = 0$  which yields  $\mathcal{O}(h^4)$  approximations to the eigenvalues.

(b) Find the eigenvalues of the difference scheme and verify directly, with a comparison to  $\lambda_n = n^2\pi^2$ , that they are actually  $\mathcal{O}(h^4)$ . How accurate are the eigenvectors?

**16.\*** Derive a *variational differential equation* that is satisfied by  $\partial Y(\lambda; x)/\partial\lambda$ , where  $Y$  is a solution of (46). Describe how Newton's iterative method might be formulated to solve for an eigenvalue  $\lambda$  from (47).

# 9

## Difference Methods for Partial Differential Equations

### 0. INTRODUCTION

Although considerable study has been made of partial differential equation problems, the mathematical theory—existence, uniqueness, or well-posedness—is not nearly as complete as it is in the case of ordinary differential equations. Furthermore, except for some problems that are solved by explicit formulae, the analytical methods developed for the treatment of partial differential equations are, in general, not suited for the efficient numerical evaluation of solutions. Hence, as may be expected, the theory of numerical methods for partial differential equations is somewhat fragmented. Where the theory of the differential equations is well developed there has been a corresponding development of numerical methods. But the difference methods found thus far usually do not permit the construction of schemes of an arbitrarily high order of accuracy. For certain systems of partial differential equations convergent numerical methods of arbitrarily high order of accuracy have been devised (for instance, linear first order hyperbolic systems in two unknowns); while for others (say the simple case of the Laplace equation on a square) only relatively low order methods have been proved to converge. Furthermore, in contrast to the case of the numerical solution of ordinary differential equation problems, the facility with which one may use difference methods on modern electronic computers to solve problems involving partial differential equations is severely limited by (a) size of the high speed memory, (b) speed of the arithmetic unit, and (c) difficulty of programming a problem for and communicating with the computer.

In view of the limitations of the scope of this book and of the incompleteness of the theory of difference methods, we shall illuminate some of the highlights of this theory through the treatment of problems for the

$$(1a) \quad \frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} = 0 \quad \text{Laplace equation;}$$

$$(1b) \quad \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{Wave equation;}$$

$$(1c) \quad \frac{\partial u}{\partial t} - a^2 \frac{\partial^2 u}{\partial x^2} = 0 \quad \text{Diffusion or heat conduction equation.}$$

The applications of these equations are so varied and well known that we do not make specific mention of particular cases. Of course, in applied mathematics other partial differential equations occur; most of these are non-linear and not covered by a complete mathematical theory of existence, uniqueness, or well-posedness.

To each of the equations in (1) we must adjoin appropriate subsidiary relations, called *boundary* and/or *initial conditions*, which serve to complete the formulation of a “meaningful problem.” These conditions are related to the domain, say  $D$ , in which the equation (1) is to be solved. When the problem arises from a physical application it is usually clear (to anyone understanding the phenomenon) what these relations must be. Some familiar examples are, for the respective equations (1a, b, and c);

$$(2a, i) \quad u = f(x, y), \quad \text{for } (x, y) \text{ on the boundary of } D,$$

or with  $\partial/\partial n$  representing the normal derivative,

$$(2a, ii) \quad \alpha u + \beta \frac{\partial u}{\partial n} = f(x, y), \quad \text{for } (x, y) \text{ on the boundary of } D;$$

$$(2b, i) \quad u(0, x) = f(x), \quad \frac{\partial u(0, x)}{\partial t} = g(x), \quad -\infty < x < \infty,$$

where  $D \equiv \{(t, x) \mid t \geq 0, -\infty < x < \infty\}$ , i.e.,  $D \equiv$  half plane, or

$$(2b, ii) \quad \begin{cases} u(0, x) = f(x), \quad \frac{\partial u(0, x)}{\partial t} = g(x), \\ u(t, a) = \alpha(t), \quad u(t, b) = \beta(t), \quad t > 0, \end{cases}$$

where  $D \equiv \{(t, x) \mid t \geq 0, a \leq x \leq b\}$ , i.e.,  $D \equiv$  half strip;

$$(2c, i) \quad u(0, x) = f(x), \quad -\infty < x < \infty,$$

where  $D$  is the half plane, or

$$(2c, ii) \quad \begin{cases} u(0, x) = f(x), & a \leq x \leq b, \\ u(t, a) = \alpha(t), & u(t, b) = \beta(t), \end{cases}$$

where  $D$  is the half strip.

If the functions introduced in (2a, b, and c) satisfy appropriate smoothness conditions, then each set of relations (i) or (ii) adjoined to the corresponding equation in (1) yields a problem which has been termed *well-posed* or *properly-posed* by Hadamard. This implies that each such problem has a bounded solution, that the solution is unique, and that it depends continuously on the data (i.e., a “small” change in  $f$ ,  $g$ ,  $\alpha$ , or  $\beta$  produces a “correspondingly small” change in the solution). There are many other combinations of boundary and/or initial conditions which together with the equations in (1) (or more general equations) constitute properly posed problems. It is such problems for which there is a reasonably developed theory of difference approximations. We shall examine this theory briefly in Section 5, after first studying some special cases. However, as we shall see in Section 5, the theory serves mainly to determine whether a given method yields approximations of reasonable accuracy; but the theory does not directly suggest how to construct numerical schemes.

## 0.1. Conventions of Notation

For simplicity, let the domain  $D$  have boundary  $C$  and lie in the three dimensional space of variables  $(x, y, t)$ . Cover this space by a *net*, *grid*, *mesh*, or *lattice* of discrete points, with coordinates  $(x_i, y_j, t_k)$  given by

$$x_i = x_0 + i\delta x, \quad y_j = y_0 + j\delta y, \quad t_k = t_0 + k\delta t; \\ i, j, k = 0, \pm 1, \pm 2, \dots$$

Here, we have taken the *net spacings*  $\delta x$ ,  $\delta y$ , and  $\delta t$  to be uniform. The lattice points may be divided into three disjoint sets:  $D_\delta$ , the *interior net points*;  $C_\delta$ , the *boundary net points*; and the remainder which are *external points*. Here we assume, again for simplicity, that  $C$  is composed of sections of coordinate surfaces. The specific rules for assigning lattice points to a particular set will be clarified in the subsequent examples and discussion.

At the points of  $D_\delta + C_\delta$  the function  $u(x, y, t)$  is to be approximated by a net function,  $U(x_i, y_j, t_k)$ . It is convenient to denote the components of net functions by appropriate subscripts and/or superscripts. For instance, we may use

$$U(x_i, y_j) \equiv U_{i,j}; \quad U(x_i, y_j, t_k) \equiv U_{i,j}^k, \quad \text{etc.}$$

This notation is frequently cumbersome and at times difficult (if not unpleasant) to read. Thus, while we shall have occasion to use it, we

prefer another notation, more in keeping with the usual functional notation. If  $U$  has been defined to be a net function, then we may write  $U(x, y, t)$  and understand the argument point  $(x, y, t)$  to be some general point of the net on which  $U$  is defined. Furthermore, if we simply write  $U$  then the argument is understood to be a general point  $(x, y, t)$  of the appropriate net.

We shall make frequent use of various difference quotients of net functions (of course, in order to approximate partial derivatives). For this purpose we introduce a *subscript notation for difference quotients* of net functions

$$(3a) \quad U_x(x, y, t) \equiv \frac{U(x + \delta x, y, t) - U(x, y, t)}{\delta x}$$

$$(3b) \quad U_{\bar{x}}(x, y, t) \equiv \frac{U(x, y, t) - U(x - \delta x, y, t)}{\delta x}$$

$$(3c) \quad U_{\hat{x}}(x, y, t) \equiv \frac{1}{2}[U_x(x, y, t) + U_{\bar{x}}(x, y, t)].$$

Clearly, (3a, b, and c) are just the *forward*, *backward*, and *centered difference quotients* with respect to  $x$ . By our previous convention we might have written the left-hand sides of (3) as just  $U_x$ ,  $U_{\bar{x}}$ , and  $U_{\hat{x}}$ . This convenient notation was introduced by Courant, Friedrichs, and Lewy in a fundamental paper on difference methods for partial differential equations. The difference quotients with respect to other discrete variables are defined in analogy with (3), say  $U_y$ ,  $U_t$ , etc. It is a simple matter to verify that these difference operators commute; i.e.,

$$U_{xy} = U_{yx}, \quad U_{xt} = U_{tx}, \quad \text{etc.}$$

A particularly important case is the centered second difference quotient which can be written as

$$(4) \quad U_{yy} = U_{\bar{y}y} = \frac{1}{(\delta y)^2} [U(x, y + \delta y, t) - 2U(x, y, t) + U(x, y - \delta y, t)].$$

## 1. LAPLACE EQUATION IN A RECTANGLE

A standard type of problem which employs the Laplace operator or Laplacian,

$$\Delta \equiv \frac{\partial^2}{\partial x^2} + \frac{\partial^2}{\partial y^2},$$

is to determine a function,  $u(x, y)$ , such that

$$(1a) \quad -\Delta u(x, y) = f(x, y), \quad (x, y) \in D;$$

$$(1b) \quad u(x, y) = g(x, y), \quad (x, y) \in C.$$

Here  $D$  is some domain in the  $x, y$ -plane and  $C$  is its boundary. If the boundary  $C$  and inhomogeneous terms  $f(x, y)$  and  $g(x, y)$  satisfy mild regularity conditions, it is well known that the problem (1) is well-posed. If  $f \equiv 0$  this is called a *Dirichlet problem* for Laplace's equation while in its present form (1a) is called the *Poisson equation*. For simplicity of presentation, we take  $D$  to be a rectangle

$$(2a) \quad D \equiv \{(x, y) \mid 0 < x < a, 0 < y < b\};$$

whose boundary  $C$  is composed of four line segments

$$(2b) \quad C \equiv \{(x, y) \mid x = 0, a, 0 \leq y \leq b; y = 0, b, 0 \leq x \leq a\}.$$

To "solve" this problem numerically we introduce the net spacings  $\delta x = a/(J + 1)$ ,  $\delta y = b/(K + 1)$ , and the uniformly spaced net points

$$x_j = j\delta x, \quad y_k = k\delta y; \quad j, k = 0, \pm 1, \pm 2, \dots$$

Those net points interior to  $D$  we call  $D_\delta$ , i.e.,

$$(3a) \quad D_\delta \equiv \{(x_j, y_k) \mid 1 \leq j \leq J; 1 \leq k \leq K\}.$$

The net points on  $C$ , with the exception of the four corners of  $C$ , we call  $C_\delta$ , i.e.,

$$(3b) \quad C_\delta \equiv \{(x_j, y_k) \mid j = (0, J + 1), 1 \leq k \leq K; \\ k = (0, K + 1), 1 \leq j \leq J\}.$$

At the net points  $D_\delta + C_\delta$  we seek quantities  $U(x_j, y_k)$  which are to approximate the solution  $u(x_j, y_k)$  of (1). The net function will, of course, be defined as the solution of a system of difference equations that replaces the partial differential equation (1a) and boundary conditions (1b) on the net.

An obvious approximation to the Laplacian is obtained by replacing each second derivative by a centered second difference quotient. Thus at each point  $(x, y) \in D_\delta$  we define

$$(4a) \quad \Delta_\delta U(x, y) \equiv U_{xx}(x, y) + U_{yy}(x, y).$$

In the subscript notation, we could also write for each  $(x_j, y_k) \in D_\delta$

$$(4b) \quad \Delta_\delta U_{j,k} \equiv \frac{U_{j+1,k} - 2U_{j,k} + U_{j-1,k}}{(\delta x)^2} + \frac{U_{j,k+1} - 2U_{j,k} + U_{j,k-1}}{(\delta y)^2}.$$

It is frequently convenient with either of these notations to indicate the net points involved in the definition of  $\Delta_\delta U$  by means of a diagram as in Figure 1. The set of points marked with crosses is called the *star* or *stencil* associated with the difference operator  $\Delta_\delta$ .

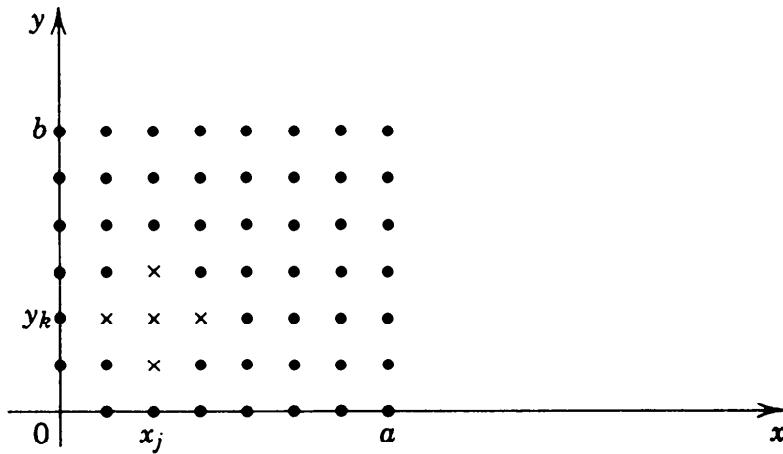


Figure 1. Net points for Laplace difference operator  $\Delta_\delta$ .

With the notation (4a), we write the difference problem as

$$(5a) \quad -\Delta_\delta U(x, y) = f(x, y), \quad (x, y) \in D_\delta;$$

$$(5b) \quad U(x, y) = g(x, y), \quad (x, y) \in C_\delta.$$

From (3), (4), and (5) we find that the values  $U$  at  $JK + 2(J + K)$  net points in  $D_\delta + C_\delta$  satisfy  $JK + 2(J + K)$  linear equations. Hence, we may hope to solve (5) for the unknowns  $U(x, y)$  in  $D_\delta + C_\delta$ . The  $2(J + K)$  values of  $U$  on  $C_\delta$  are specified in (5b) and so the  $JK$  equations of (5a) must determine the remaining  $JK$  unknowns. We shall first show that this system has a unique solution and then we will estimate the error in the approximation. Finally, we shall consider practical methods for solving the linear system (5).

To demonstrate that the difference equations have a unique solution we shall prove that the corresponding homogeneous system has only the trivial solution. For this purpose and for the error estimates to be obtained, we first prove a *maximum principle* for the operator  $\Delta_\delta$ .

### **THEOREM 1.**

(a) *If  $V(x, y)$  is a net function defined on  $D_\delta + C_\delta$  and satisfies*

$$\Delta_\delta V(x, y) \geq 0 \quad \text{for all } (x, y) \in D_\delta,$$

*then*

$$\max_{D_\delta} V(x, y) \leq \max_{C_\delta} V(x, y).$$

(b) *Alternatively, if  $V$  satisfies*

$$\Delta_\delta V(x, y) \leq 0 \quad \text{for all } (x, y) \in D_\delta,$$

*then*

$$\min_{D_\delta} V(x, y) \geq \min_{C_\delta} V(x, y).$$

*Proof.* We prove part (a) by contradiction. Assume that at some point  $P_0 \equiv (x^*, y^*)$  of  $D_\delta$ , we have  $V(P_0) = M$  where

$$M \geq V(P) \quad \text{for all } P \in D_\delta \quad \text{and} \quad M > V(P) \quad \text{for all } P \in C_\delta.$$

Let us introduce the notation  $P_1 \equiv (x^* + \delta x, y^*)$ ,  $P_2 \equiv (x^* - \delta x, y^*)$ ,  $P_3 \equiv (x^*, y^* + \delta y)$ ,  $P_4 \equiv (x^*, y^* - \delta y)$  and then use (4) to write

$$\Delta_\delta V(P_0) \equiv \theta_x[V(P_1) + V(P_2)] + \theta_y[V(P_3) + V(P_4)] - 2(\theta_x + \theta_y)V(P_0)$$

where  $\theta_x \equiv 1/(\delta x)^2$  and  $\theta_y \equiv 1/(\delta y)^2$ . However, by hypothesis  $\Delta_\delta V(P_0) \geq 0$ , so we have

$$M = V(P_0) \leq \frac{1}{\theta_x + \theta_y} \left[ \theta_x \frac{V(P_1) + V(P_2)}{2} + \theta_y \frac{V(P_3) + V(P_4)}{2} \right].$$

But  $M \geq V$  implies that  $V(P_\nu) = M$  for  $\nu = 1, 2, 3, 4$ . We now repeat this argument for each interior point  $P_\nu$  instead of the point  $P_0$ . By repetition, each point of  $D_\delta$  and  $C_\delta$  appears as one of the  $P_\nu$  for some corresponding  $P_0$ . Thus, we conclude that

$$V(P) = M \quad \text{for all } P \text{ in } D_\delta + C_\delta,$$

which contradicts the assumption that  $V < M$  on  $C_\delta$ . Part (a) of the theorem follows.†

To prove part (b), we could repeat an argument similar to the above. However, it is simpler to recall that

$$\max [-V(x, y)] = -\min V(x, y); \quad \Delta_\delta (-V) = -\Delta_\delta (V).$$

Hence, if  $V$  satisfies the hypothesis of part (b), then  $-V$  satisfies the hypothesis of part (a). But the conclusion of part (a) for  $-V$  is identical to the conclusion of part (b) for  $V$ .† ■

Let us now consider the homogeneous system corresponding to (5); i.e.,  $f \equiv g \equiv 0$ . From Theorem 1, it follows that the *max* and *min* of the solution of this homogeneous system vanish; hence, the only solution is the trivial one. Thus it follows by the alternative principle for linear systems that (5) has a unique solution for arbitrary  $f(x, y)$  and  $g(x, y)$ .

A bound for the solution of the difference equation (5) can also be obtained by an appropriate application of the *maximum principle*. The result, called an *a priori estimate*, may be stated as

**THEOREM 2.** *Let  $V(x, y)$  be any net function defined on the sets  $D_\delta$  and  $C_\delta$  defined by (3). Then*

$$(6) \quad \max_{D_\delta} |V| \leq \max_{C_\delta} |V| + \frac{\alpha^2}{2} \max_{D_\delta} |\Delta_\delta V|.$$

† We have in fact proved more; namely, that if the maximum, in case (a), or the minimum, in case (b), of  $V(x, y)$  occurs in  $D_\delta$ , then  $V(x, y)$  is constant on  $D_\delta + C_\delta$ .

*Proof.* We introduce the function

$$\phi(x, y) \equiv \frac{1}{2}x^2$$

and observe that for all  $(x, y) \in D_\delta + C_\delta$ ,

$$0 \leq \phi(x, y) \leq \frac{a^2}{2}; \quad \Delta_\delta \phi(x, y) = 1.$$

Now define the two net functions  $V_+(x, y)$  and  $V_-(x, y)$  by

$$V_\pm(x, y) \equiv \pm V(x, y) + N\phi(x, y),$$

where

$$N \equiv \max_{D_\delta} |\Delta_\delta V|.$$

Clearly for all  $(x, y) \in D_\delta$ , it follows that

$$\Delta_\delta V_\pm(x, y) = \pm \Delta_\delta V(x, y) + N \geq 0.$$

Thus we may apply the maximum principle, part (a) of Theorem 1, to each of  $V_\pm(x, y)$  to obtain for all  $(x, y) \in D_\delta$ ,

$$\begin{aligned} V_\pm(x, y) &\leq \max_{C_\delta} V_\pm(x, y) \\ &= \max_{C_\delta} [\pm V(x, y) + N\phi] \leq \max_{C_\delta} [\pm V(x, y)] + N \frac{a^2}{2}. \end{aligned}$$

But from the definition of  $V_\pm$  and the fact that  $\phi \geq 0$ ,

$$\pm V(x, y) \leq V_\pm(x, y).$$

Hence,

$$\begin{aligned} \pm V(x, y) &\leq \max_{C_\delta} [\pm V(x, y)] + N \frac{a^2}{2}, \\ &\leq \max_{C_\delta} |V| + \frac{a^2}{2} N. \end{aligned}$$

Since the right-hand side in the final inequality is independent of  $(x, y)$  in  $D_\delta$  the theorem follows. ■

Note that we could readily replace  $a^2/2$  in (6) by  $b^2/2$  since the function  $\psi(x, y) = y^2/2$  can be used in place of  $\phi(x, y)$  in the proof of the theorem.

It is now a simple matter to estimate the error  $U - u$ . We introduce the *local truncation error*,  $\tau\{\Phi\}$ , for the difference operator  $\Delta_\delta$  on  $D_\delta$  by

$$(7) \quad \tau\{\Phi(x, y)\} \equiv \Delta_\delta \Phi(x, y) - \Delta \Phi(x, y), \quad (x, y) \text{ in } D_\delta,$$

where  $\Phi(x, y)$  is any sufficiently smooth function defined on  $D$ . Now if  $u(x, y)$  is the solution of the boundary value problem (1) we have from (1a) at the points of  $D_\delta$

$$-\Delta_\delta u(x, y) = f(x, y) - \tau\{u(x, y)\}.$$

Subtracting this from (5a) at each point of  $D_\delta$  yields

$$(8a) \quad -\Delta_\delta[U(x, y) - u(x, y)] = \tau\{u(x, y)\}, \quad (x, y) \text{ in } D_\delta.$$

Also from (1b) and (5b) we obtain

$$(8b) \quad U(x, y) - u(x, y) = 0, \quad (x, y) \text{ in } C_\delta.$$

Now apply Theorem 2 to the net function  $U(x, y) - u(x, y)$  and we get by (8)

$$\max_{D_\delta} |U(x, y) - u(x, y)| \leq \frac{a^2}{2} \max_{D_\delta} |\tau\{u\}|.$$

Upon introducing the *maximum norm* defined for any net function  $W(x, y)$  by  $\|W\| = \max_{D_\delta} |W|$ , we have the

**COROLLARY.** *With  $u$ ,  $U$ , and  $\tau$  defined respectively by (1), (5), and (7), we have*

$$(9) \quad \|U(x, y) - u(x, y)\| \leq \frac{a^2}{2} \|\tau\{u\}\|. \quad \blacksquare$$

Note that the error bound is proportional to the truncation error!

It is easy to estimate  $\|\tau\|$ . If the solution  $u(x, y)$  of (1) has *continuous and bounded fourth order partial derivatives* in  $D$ , then

$$(10) \quad \begin{aligned} u(x \pm \delta x, y) &= u(x, y) \pm \delta x \frac{\partial u(x, y)}{\partial x} + \frac{(\delta x)^2}{2!} \frac{\partial^2 u(x, y)}{\partial x^2} \\ &\quad \pm \frac{(\delta x)^3}{3!} \frac{\partial^3 u(x, y)}{\partial x^3} + \frac{(\delta x)^4}{4!} \frac{\partial^4 u(x + \theta_\pm \delta x, y)}{\partial x^4}, \\ &\quad |\theta_\pm| < 1. \end{aligned}$$

Thus we find, as in Chapter 6, that

$$u_{xx}(x, y) - \frac{\partial^2 u(x, y)}{\partial x^2} = \frac{(\delta x)^2}{12} \frac{\partial^4 u(x + \theta \delta x, y)}{\partial x^4}, \quad |\theta| < 1,$$

with a similar result for the  $y$  derivatives. Hence,

$$\begin{aligned} \tau\{u(x, y)\} &= \Delta_\delta u(x, y) - \Delta u(x, y) \\ &= \frac{1}{12} \left[ (\delta x)^2 \frac{\partial^4 u(x + \theta \delta x, y)}{\partial x^4} + (\delta y)^2 \frac{\partial^4 u(x, y + \psi \delta x)}{\partial y^4} \right]. \end{aligned}$$

If we denote the bounds of the respective fourth order derivatives by  $M_x^{(4)}$  and  $M_y^{(4)}$ , then

$$(11a) \quad \|\tau\{u\}\| \leq \frac{1}{12} [(\delta x)^2 M_x^{(4)} + (\delta y)^2 M_y^{(4)}].$$

If  $u(x, y)$  has only *continuous third order derivatives*, we terminate the expansions in (10) one term earlier and get

$$u_{x\bar{x}}(x, y) - \frac{\partial^2 u(x, y)}{\partial x^2} = \frac{\delta x}{3!} \left[ \frac{\partial^3 u(x + \theta_+ \delta x, y)}{\partial x^3} - \frac{\partial^3 u(x + \theta_- \delta x, y)}{\partial x^3} \right].$$

If the moduli of continuity in  $D$  of the third derivatives  $\partial^3 u / \partial x^3$  and  $\partial^3 u / \partial y^3$  are denoted by  $\omega_x^{(3)}(\delta)$  and  $\omega_y^{(3)}(\delta)$ , respectively, we have

$$(11b) \quad \|\tau\{u\}\| \leq \frac{1}{6} [\delta x \omega_x^{(3)}(2\delta x) + \delta y \omega_y^{(3)}(2\delta y)].$$

Clearly by these procedures, we find that if  $u(x, y)$  has only continuous second derivatives with moduli of continuity  $\omega_x^{(2)}(\delta)$  and  $\omega_y^{(2)}(\delta)$ , then

$$(11c) \quad \|\tau\{u\}\| \leq \omega_x^{(2)}(\delta x) + \omega_y^{(2)}(\delta y).$$

With the aid of any of the estimates (11a, b, or c) that may be appropriate, the corollary establishes convergence of the approximate solution to the exact solution *as*  $\delta x \rightarrow 0$  and  $\delta y \rightarrow 0$  *in any manner*. We see that the convergence rate is generally faster for “smoother” solutions  $u(x, y)$ . For solutions which have more than four continuous derivatives, we cannot deduce better truncation error estimates than that given by (11a). It is possible to construct more accurate difference approximations to the Laplacian, which then have solutions  $U$  of greater accuracy than  $\mathcal{O}[(\delta x)^2 + (\delta y)^2]$ . But there is no general way of constructing convergence proofs for similar schemes of *arbitrarily* high order truncation error. In fact, it is unlikely that such schemes, which are of maximum order of accuracy, converge in general.

The effects of roundoff can also be estimated by means of Theorem 2. Let the numbers actually computed be denoted by  $\bar{U}(x, y)$ . Then we can write

$$(12a) \quad -\Delta_\delta \bar{U}(x, y) = f(x, y) + \frac{1}{\delta x \delta y} \rho(x, y), \quad (x, y) \in D_\delta;$$

$$(12b) \quad \bar{U}(x, y) = g(x, y) + \rho'(x, y), \quad (x, y) \in C_\delta.$$

Here  $\rho'(x, y)$  is the roundoff error in approximating the boundary data. After noting that the coefficients in  $\Delta_\delta$  are proportional to  $1/(\delta x)^2$  and  $1/(\delta y)^2$ , we have defined the roundoff errors,  $\rho(x, y)$ , in the computations (12a) to be proportional to a similar factor. This corresponds to the fact that the actual computations are done with the form of (4) which results after multiplication by the factor  $\delta x \delta y$ . We now obtain from (1), (7) and (12)

$$-\Delta_\delta [\bar{U}(x, y) - u(x, y)] = \tau\{u(x, y)\} + \frac{\rho(x, y)}{\delta x \delta y}, \quad (x, y) \in D_\delta;$$

$$\bar{U}(x, y) - u(x, y) = \rho'(x, y), \quad (x, y) \in C_\delta.$$

Thus for the net function  $\bar{U}(x, y) - u(x, y)$ , Theorem 2 implies

**THEOREM 3.** *With  $u$ ,  $\bar{U}$ , and  $\tau$  defined by (1), (12), and (7) respectively, we have*

$$(13) \quad \|\bar{U}(x, y) - u(x, y)\| \leq \|\rho'\| + \frac{a^2}{2} \left[ \|\tau\{u\}\| + \frac{1}{\delta x \delta y} \|\rho\| \right].$$

Here

$$\|\rho'\| = \max_{C_\delta} |\rho'(x, y)| \quad \text{and} \quad \|\rho\| = \max_{D_\delta} |\rho(x, y)|. \quad \blacksquare$$

Thus we find that the boundary roundoff error and the interior roundoff error have quite different effects on the accuracy as  $\delta x$  and  $\delta y \rightarrow 0$ . In fact, to be consistent with the truncation error, the interior roundoff error,  $\rho$ , should be of the same order as  $\delta x \delta y \tau\{u\}$  and the boundary roundoff error,  $\rho'$ , should be of the same order as  $\tau$  when  $\delta x$  and  $\delta y \rightarrow 0$ . This result for  $\rho$  is analogous to that in (7.34), of Chapter 8 where simple difference approximations of an ordinary boundary value problem were considered.

The maximum principle and its applications given here can be generalized in various ways (see Problems 1–4). Extensions to rectangular domains in higher dimensions are straightforward, and non-rectangular domains may also be treated (with suitable modifications of the difference equations near the boundary surface).

### 1.1. Matrix Formulation

The system of linear equations (5) can be written in matrix-vector notation in various ways. For this purpose, we use the subscript notation for any net function  $V(x, y)$  defined on  $D_\delta + C_\delta$

$$V(x_j, y_k) \equiv V_{jk}; \quad 0 \leq j \leq J + 1, 0 \leq k \leq K + 1.$$

From the values of such a net function, construct the  $J$ -dimensional vector

$$(14a) \quad \mathbf{V}_k \equiv \begin{pmatrix} V_{1k} \\ V_{2k} \\ \vdots \\ V_{Jk} \end{pmatrix}, \quad k = 0, 1, 2, \dots, K + 1.$$

Each vector  $\mathbf{V}_k$  consists of the elements of the net function  $V(x_j, y_k)$  on the coordinate line segment  $y = y_k$ ,  $x_1 \leq x \leq x_J$ . (We note that the elements on the line segments  $y_1 \leq y \leq y_K$ ,  $x = x_0$ , and  $x = x_{J+1}$  are not included.) We also introduce the  $J$ th order square matrices

$$I_J \equiv \begin{bmatrix} 1 & & & \\ & \ddots & & \\ & & \ddots & \\ & & & 1 \end{bmatrix} \equiv (\delta_{ij}),$$

$$(14b) \quad L_J \equiv \begin{bmatrix} 0 & & & & \\ 1 & 0 & & & \\ & \ddots & \ddots & & \\ & & \ddots & \ddots & \\ & & & \ddots & \\ & & & & 1 & 0 \end{bmatrix} \equiv (a_{ij}), \quad a_{ij} = \delta_{i-1,j},$$

and the quantities

$$(14c) \quad \delta^2 \equiv \frac{(\delta x)^2(\delta y)^2}{2[(\delta x)^2 + (\delta y)^2]}, \quad \theta_x \equiv \frac{(\delta y)^2}{2[(\delta x)^2 + (\delta y)^2]},$$

$$\theta_y \equiv \frac{(\delta x)^2}{2[(\delta x)^2 + (\delta y)^2]}.$$

Upon multiplying (5a) by  $\delta^2$ , we can write the result for  $(x, y) = (x_j, y_k)$ , in subscript notation as

$$(15) \quad U_{jk} - \theta_x(U_{j-1,k} + U_{j+1,k}) - \theta_y(U_{j,k-1} + U_{j,k+1}) = \delta^2 f_{jk},$$

$$1 \leq j \leq J, \quad 1 \leq k \leq K.$$

Or with the vector and matrix notation of (14) this system becomes

$$(16a) \quad [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_1 - \theta_y\mathbf{U}_2 = \delta^2 \mathbf{F}_1,$$

$$-\theta_y\mathbf{U}_{k-1} + [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_k - \theta_y\mathbf{U}_{k+1} = \delta^2 \mathbf{F}_k,$$

$$2 \leq k \leq K-1;$$

$$-\theta_y\mathbf{U}_{K-1} + [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_K = \delta^2 \mathbf{F}_K.$$

Here we have introduced

$$(16b) \quad \mathbf{F}_1 = \mathbf{f}_1 + \frac{1}{(\delta x)^2} \mathbf{w}_1 + \frac{1}{(\delta y)^2} \mathbf{U}_0,$$

$$\mathbf{F}_k = \mathbf{f}_k + \frac{1}{(\delta x)^2} \mathbf{w}_k, \quad 2 \leq k \leq K-1,$$

$$\mathbf{F}_K = \mathbf{f}_K + \frac{1}{(\delta x)^2} \mathbf{w}_K + \frac{1}{(\delta y)^2} \mathbf{U}_{K+1},$$

where

$$\mathbf{w}_k \equiv (w_{ik}) \equiv \begin{pmatrix} U_{0k} \\ 0 \\ \vdots \\ 0 \\ U_{J+1,k} \end{pmatrix},$$

i.e.,  $w_{ik} = 0$  for  $2 \leq i \leq J - 1$ ;  $w_{1k} = U_{0k}$ ;  $w_{Jk} = U_{J+1,k}$ . Of course, all of the  $U_{jk}$  which enter into (16b) are known quantities given in (5b).

Further simplification is obtained by introducing  $JK$ -dimensional vectors or  $K$ -dimensional *compound vectors* (i.e., vectors whose components are  $J$ -dimensional vectors)

$$(17a) \quad \mathbf{U} \equiv \begin{pmatrix} \mathbf{U}_1 \\ \mathbf{U}_2 \\ \vdots \\ \mathbf{U}_K \end{pmatrix} \equiv \begin{pmatrix} U_{11} \\ U_{21} \\ \vdots \\ U_{JK} \end{pmatrix}, \quad \mathbf{F} \equiv \begin{pmatrix} \mathbf{F}_1 \\ \mathbf{F}_2 \\ \vdots \\ \mathbf{F}_K \end{pmatrix};$$

and the square matrices of order  $JK$

$$I \equiv (\delta_{ij}),$$

$$(17b) \quad L \equiv \begin{pmatrix} L_J & & & \\ & L_J & & \\ & & \ddots & \\ & & & L_J \end{pmatrix}, \quad B \equiv \begin{pmatrix} 0 & & & \\ I_J & 0 & & \\ & \ddots & \ddots & \\ & & \ddots & \ddots \\ & & & I_J & 0 \end{pmatrix},$$

$$H \equiv \theta_x(L + L^T), \quad V \equiv \theta_y(B + B^T), \quad A \equiv I - H - V.$$

Now the system (16), or equivalently (5), can be written as

$$(18) \quad A\mathbf{U} = \delta^2\mathbf{F}.$$

The vectors in (17a) associate a component with each net point  $(x, y)$  of  $D_\delta$ . In the indicated vectors of dimension  $N = JK$ , the  $r$ th component is the value of the net function at the point  $(x_r, y_k)$  such that  $r = j + (k - 1)J$ . If the assignment of integers,  $r$ , to net points of  $D_\delta$  is done in some other order, then the vectors and matrices are changed by some permutation. (Another ordering of interest would be to list the elements on lines  $x = \text{constant}$  of  $D_\delta$ ). The previous proof that the system (5) has a

unique solution now implies that  $A$  is non-singular. We shall prove this fact directly by showing that the eigenvalues of  $A$  are positive.

Let us consider again the problem of obtaining error estimates for the approximate solution. Multiply (8a) by  $\delta^2$  and employ the present notation to obtain in place of the system (8)

$$(19a) \quad A(\mathbf{U} - \mathbf{u}) = \delta^2 \boldsymbol{\tau}.$$

Here  $\mathbf{U}$  is as before,  $\mathbf{u}$  is the vector of the exact solution on  $D_\delta$ , and  $\boldsymbol{\tau}$  is the vector of local truncation errors  $\tau\{u(x, y)\}$  on  $D_\delta$  with no adjustments now required as in (16b) since  $U - u = 0$  on  $C_\delta$ . Then as  $A$  is non-singular, we have

$$(19b) \quad \mathbf{U} - \mathbf{u} = \delta^2 A^{-1} \boldsymbol{\tau},$$

which is an exact representation for the error. By using any vector norm and the corresponding natural matrix norm, we have from (19b),

$$(20) \quad \|\mathbf{U} - \mathbf{u}\| \leq \delta^2 \|A^{-1}\| \cdot \|\boldsymbol{\tau}\|.$$

We note from (17b), that  $A = A^T$  and thus  $A^{-1}$  is symmetric. If we use the Euclidean norm in (20), i.e., for any vector  $\mathbf{v}$ ,

$$\|\mathbf{v}\|_2 = \sqrt{\sum_{v=1}^{JK} v_v^2},$$

then

$$\|A^{-1}\|_2 = \max_{1 \leq v \leq JK} (1/|\Lambda_v|) = 1/\min_{1 \leq v \leq JK} |\Lambda_v|$$

where the  $\Lambda_v$  are the eigenvalues of  $A$ . The eigenvalues of  $A$  satisfy

$$(21) \quad A\mathbf{W} = \Lambda\mathbf{W}.$$

However, we see that this is equivalent to the *finite difference eigenvalue problem*

$$(22a) \quad -\Delta_\delta W(x, y) = \frac{\Lambda}{\delta^2} W(x, y), \quad (x, y) \text{ in } D_\delta$$

$$(22b) \quad W(x, y) = 0, \quad (x, y) \text{ in } C_\delta,$$

since multiplication of (22a) by  $\delta^2$  yields (21).

We determine the eigenvalues of problem (21) by using the technique called *separation of variables* for (22). Let us try to find a solution of the form  $W(x, y) = \phi(x)\psi(y)$  of (22a), i.e.,

$$-\Delta_\delta \phi(x)\psi(y) = -\phi_{xx}\psi(y) - \phi(x)\psi_{yy} = \frac{\Lambda}{\delta^2} \phi(x)\psi(y).$$

Now divide by  $W(x, y)$  to get

$$-\frac{\phi_{xx}}{\phi(x)} - \frac{\psi_{yy}}{\psi(y)} = \frac{\Lambda}{\delta^2}, \quad (x, y) \text{ in } D_\delta.$$

But the only way that the sum of a function of  $x$  and a function of  $y$  can be constant is for each function to be a constant. Hence we may write  $\Lambda = \xi + \eta$  and have the two sets of equations

$$(23a) \quad -\phi_{xx}(x) = \frac{\xi}{\delta^2} \phi(x) \quad (x, y) \text{ in } D_\delta$$

$$(23b) \quad -\psi_{yy}(y) = \frac{\eta}{\delta^2} \psi(y)$$

If  $\xi$  and  $\eta$  are known, (23) would be ordinary difference equations of second order with constant coefficients. We solve them as we did the difference equations in Section 4 of Chapter 8. Thus, let us use the form  $\phi(x) = \alpha^x$  in (23a) to get by using (14c),

$$\frac{\alpha^x}{(\delta x)^2} \left[ -\alpha^{-\delta x} + \left( 2 - \frac{\xi}{\theta_x} \right) - \alpha^{\delta x} \right] = 0, \quad \delta x \leq x \leq a - \delta x.$$

If we set  $\omega = \alpha^{\delta x}$ , then these equations are satisfied provided

$$\frac{\xi}{\theta_x} = 2 - \omega - \omega^{-1}.$$

Furthermore, it is clear that  $\phi(x) = \alpha^{-x}$  yields the same condition, and hence the general solution of (23a) is of the form

$$\phi(x_j) = c\alpha^{x_j} + d\alpha^{-x_j} = c\omega^j + d\omega^{-j}.$$

To satisfy the boundary conditions (22b), we have  $\phi(x)\psi(y) = 0$  for  $(x, y)$  in  $C_\delta$ . This implies that

$$(24) \quad \phi(0) = \phi(x_0) = 0; \quad \phi(a) = \phi(x_{J+1}) = 0.$$

From the condition (24) at  $j = 0$ , we have  $c = -d$ ; hence at  $j = J + 1$ ,

$$\omega^{2(J+1)} = 1.$$

The  $2(J + 1)$  roots of this equation are the roots of unity

$$\omega_p = e^{i[p\pi/(J+1)]}, \quad p = 1, 2, \dots, 2(J+1).$$

However, if we replace  $\omega$  by  $\omega^{-1}$ , the solution  $\phi(x)$  of the difference equation becomes  $-\phi(x)$ . Hence, we need consider only the first  $J + 1$  such roots. But the  $(J + 1)$ st root is  $\omega = -1$  which leads to the trivial solution,  $\phi \equiv 0$ . Thus we have found  $J$  non-trivial solutions of (23a) which satisfy (24) and they are

$$(25a) \quad \phi_p(x_j) = c(\omega_p^j - \omega_p^{-j}) = \sin \left( j \frac{p\pi}{J+1} \right) \quad p = 1, 2, \dots, J.$$

$$(25b) \quad \xi_p = 2\theta_x \left( 1 - \cos \frac{p\pi}{J+1} \right) = 4\theta_x \sin^2 \left( \frac{\pi}{2} \frac{p}{J+1} \right)$$

Here we have chosen the arbitrary normalization constant of  $\phi(x)$  to be  $c = -i/2$ . In an exactly analogous manner, we find  $K$  non-trivial solutions of (23b) which satisfy  $\psi(0) = \psi(b) = 0$ ;

$$(26a) \quad \psi_q(y_k) = \sin\left(k \frac{q\pi}{K+1}\right) \quad q = 1, 2, \dots, K.$$

$$(26b) \quad \eta_q = 4\theta_y \sin^2\left(\frac{\pi}{2} \frac{q}{K+1}\right)$$

By combining these results, we find the solutions of the eigenvalue problem (22)

$$(27) \quad \begin{aligned} W_{p,q}(x, y) &= \phi_p(x)\psi_q(y) \\ \Lambda_{p,q} &= \xi_p + \eta_q \end{aligned} \quad 1 \leq p \leq J, \quad 1 \leq q \leq K.$$

We have thus found  $JK$  different eigenfunctions  $W_{p,q}(x, y)$ , with corresponding eigenvalues  $\Lambda_{p,q}$  (which may not all be distinct). In the vector representation of the net functions  $W_{p,q}(x, y)$ , we have  $JK$  distinct eigenvectors,  $\mathbf{W}_{p,q}$ , of the matrix  $A$  in (21). [In fact, it can be shown that the  $JK$  eigenvectors in (27) are orthogonal.] Hence, all of the eigenvalues of  $A$  are in the set  $\Lambda_{p,q}$ . We observe that all eigenvalues of  $A$  are positive and  $A$  is not only non-singular, but is also *positive definite*.

The norm of  $A^{-1}$  is now found to be

$$\begin{aligned} \|A^{-1}\|_2 &= [\min_{p,q} (\xi_p + \eta_q)]^{-1} = \frac{1}{\xi_1 + \eta_1} \\ &= \left[ 4\theta_x \sin^2\left(\frac{\pi}{2a} \delta x\right) + 4\theta_y \sin^2\left(\frac{\pi}{2b} \delta y\right) \right]^{-1} \\ &= \frac{1}{\delta^2 \pi^2} \left( \frac{a^2 b^2}{a^2 + b^2} \right) \{1 + \mathcal{O}[(\delta x)^2 + (\delta y)^2]\}. \end{aligned}$$

Thus the error estimate in (20) becomes in this norm,

$$(28) \quad \|\mathbf{U} - \mathbf{u}\|_2 \leq \frac{a^2 b^2}{\pi^2(a^2 + b^2)} \|\boldsymbol{\tau}\|_2 \cdot \{1 + \mathcal{O}[(\delta x)^2 + (\delta y)^2]\}.$$

This bound is similar to that in (9) but it must be recalled that the norms are different. We have presented here a convergence proof which is independent of the maximum principle. There are still other proofs that could have been given. In particular, if we were to consider the problem (1) with  $f(x, y) \equiv 0$  on  $D$ , then the solution could easily be written in terms of Fourier series [assuming  $g(x, y)$  to have piecewise continuous derivatives on  $C$ ]. The solution of the corresponding difference problem

(5), with  $f(x, y) \equiv 0$ , can also be given in terms of (finite) Fourier series. A comparison of these explicit solutions would then show convergence as  $\delta x$  and  $\delta y$  vanish, and the rate of convergence would depend upon the smoothness properties of the boundary data,  $g(x, y)$ . Of course, the determination of the explicit solutions used in the calculation of  $\|A^{-1}\|_2$  cannot be made in most of the applications of the present difference method. In particular, if the domain is not composed of coordinate lines and/or if the equation is replaced by one with variable coefficients, then these special methods must be modified to give analogous results. However, the maximum principle is readily extended to include many such applications. Often it may be possible to obtain a *bound* on  $\|A^{-1}\|$  (in some norm) without having to determine the eigenvalues of  $A$ .

## 1.2. An Eigenvalue Problem for the Laplacian Operator

In view of the development of the previous subsection, we can readily find approximations to the *eigenfunctions*,  $u(x, y)$ , and *eigenvalues*,  $\lambda$ , of the Laplacian operator for a rectangular region. The eigenfunction is not identically zero, i.e.,  $u \not\equiv 0$ , and for some constant,  $\lambda$ , (the eigenvalue) satisfies

$$(29a) \quad -\Delta u = \lambda u, \quad (x, y) \text{ in } D,$$

$$(29b) \quad u = 0, \quad (x, y) \text{ in } C.$$

We can solve this continuous problem by the *separation of variables* technique. Thus we set

$$u = f(x)g(y),$$

whence from (29a)

$$-\frac{f''}{f} - \frac{g''}{g} = \lambda,$$

while from (29b)

$$f(0) = f(a) = g(0) = g(b) = 0.$$

But since  $\lambda$  is a constant, we find that

$$-\frac{f''}{f} = \text{constant}, \quad 0 \leq x \leq a,$$

$$-\frac{g''}{g} = \text{constant}, \quad 0 \leq y \leq b.$$

The only possible *non-trivial solutions* of these differential equations and boundary conditions are proportional to

$$(30a) \quad f_m(x) = \sin \left( m\pi \frac{x}{a} \right)$$

$$(30b) \quad g_n(y) = \sin \left( n\pi \frac{y}{b} \right).$$

Hence the eigenfunctions and eigenvalues of (29) are

$$(31a) \quad u_{m,n}(x, y) = \sin\left(m\pi \frac{x}{a}\right) \sin\left(n\pi \frac{y}{b}\right),$$

$$(31b) \quad \lambda_{m,n} = \pi^2 \left( \frac{m^2}{a^2} + \frac{n^2}{b^2} \right); \quad m, n = 1, 2, \dots$$

[It can be shown that these are all of the independent eigenfunctions and eigenvalues of (29).] We now exhibit the eigenfunctions  $U$  and eigenvalues  $\mu$  of the approximating difference equations defined by  $U \not\equiv 0$ ,

$$(32a) \quad -\Delta_\delta U = \mu U, \quad (x, y) \text{ in } D_\delta$$

$$(32b) \quad U = 0, \quad (x, y) \text{ in } C_\delta.$$

That is, from (27), (26), and (25), with  $\mu = \Lambda/\delta^2$ ,

$$(33a) \quad U_{p,q}(x_j, y_k) = \sin\left(j \frac{p\pi}{J+1}\right) \sin\left(k \frac{q\pi}{K+1}\right),$$

$$(33b) \quad \mu_{p,q} = 4 \left[ \frac{(J+1)^2 \sin^2\left(\frac{\pi}{2} \frac{p}{J+1}\right)}{a^2} + \frac{(K+1)^2 \sin^2\left(\frac{\pi}{2} \frac{q}{K+1}\right)}{b^2} \right],$$

$$1 \leq p \leq J, \quad 1 \leq q \leq K.$$

From (31a) and (33a), we note that

$$(34a) \quad u_{p,q}(x_j, y_k) = U_{p,q}(x_j, y_k).$$

This is an exceptional coincidence! On the other hand, if we use (31b) and expand (33b) for fixed  $(p, q)$  and large  $(J, K)$ , we find

$$(34b) \quad \mu_{p,q} - \lambda_{p,q} = \mathcal{O}[p^4(\delta x)^2 + q^4(\delta y)^2].$$

Equation (34b) expresses the fact, also valid in more general problems, that the lowest eigenvalues of the difference operator approximate the respective lowest eigenvalues of the differential operator with an error proportional to the square of the mesh width. Frequently the error in the approximation of the corresponding eigenfunctions is also proportional to the square of the mesh width.

In most cases, where the eigenvalues of the differential operator obey a *variational principle*, the practical problem of determining the eigenvalues of the difference operator is made simpler by characterizing them as the *stationary values* of some functional.

For example, in the case of (29) the eigenvalues are the stationary values,  $\lambda = G[u]$ , of

$$(35) \quad G[u] \equiv \frac{\int_D \int \left[ \left( \frac{\partial u}{\partial x} \right)^2 + \left( \frac{\partial u}{\partial y} \right)^2 \right] dx dy}{\int_D \int u^2 dx dy},$$

where  $u(x, y)$  ranges over the class  $\mathcal{G}$  of non-trivial functions with continuous first derivatives and such that  $u \equiv 0$  on  $C$ . We say  $G[u]$  is *stationary* at  $u$ , if

$$\frac{d}{d\epsilon} G[u + \epsilon v] = 0 \quad \text{at } \epsilon = 0,$$

for all functions  $v$  in  $\mathcal{G}$ . It can be shown that if  $\lambda = G[u]$  is stationary at  $u$ , then  $u$  has continuous second derivatives and satisfies (29). On the other hand, the corresponding functional that characterizes the eigenvalues of the difference operator in (32) is

$$(36) \quad H[U] \equiv \frac{Q[U]}{L[U]} \equiv \frac{\frac{1}{2} \sum [(U_x)^2 + (U_{\bar{x}})^2 + (U_y)^2 + (U_{\bar{y}})^2]}{\sum U^2}.$$

The sums in (36) are taken over all net points of the infinite lattice that covers the plane and  $U$  is in the class  $\mathcal{F}$  of non-trivial net functions which satisfy

$$U(x, y) = 0 \quad \text{for } (x, y) \text{ not in } D_\delta.$$

**THEOREM 4.**  $\mu = H[U]$  is stationary at  $U$  iff  $\mu$  and  $U$  are an eigenvalue and eigenfunction that satisfy (32).

*Proof.* Let

$$(37) \quad \begin{aligned} V(x_0, y_0) &= 1, \\ V(x, y) &= 0 \quad \text{if } (x, y) \neq (x_0, y_0). \end{aligned}$$

It is easy to calculate

$$\frac{d}{d\epsilon} H[U + \epsilon V] \Big|_{\epsilon=0}$$

by expanding the numerator and denominator of  $H$  to first order in  $\epsilon$ . That is,

$$\begin{aligned} Q[U + \epsilon V] &\cong Q[U] + \epsilon \sum U_x V_x + U_{\bar{x}} V_{\bar{x}} + U_y V_y + U_{\bar{y}} V_{\bar{y}} \\ &= Q[U] - 2\epsilon \Delta_\delta U(x_0, y_0); \\ L[U + \epsilon V] &\cong L[U] + 2\epsilon \sum UV \\ &= L[U] + 2\epsilon U(x_0, y_0). \end{aligned}$$

Hence,

$$(38) \quad H[U + \epsilon V] \cong H[U] - \frac{2\epsilon}{L[U]} [\Delta_\delta U(x_0, y_0) + \mu U(x_0, y_0)].$$

Therefore, if  $H[U]$  is stationary, (32) holds, since we may pick  $(x_0, y_0)$  to be any point in  $D_\delta$ . On the other hand, in Problem 5 it is shown that (32) implies  $H[U]$  is stationary. ■

We remark that the variational principles can be used as a basis for constructing methods to determine the eigenfunctions and eigenvalues as in the Rayleigh-Ritz methods, which we do not treat. Another application of the related functionals (e.g.,  $H[u]$  and  $G[u]$ ) is to determine estimates for  $\mu_{p,q} - \lambda_{p,q}$  in more general cases. But we cannot pursue this topic further.

The eigenvalue problem (32) corresponds to the matrix eigenvalue problem, similar to (21),

$$A\mathbf{U} = \mu \delta^2 \mathbf{U},$$

where  $A$  is symmetric. Hence we may use the argument of Theorem 7.2 of Chapter 8 to prove a result analogous to that cited therein.

## PROBLEMS, SECTION 1

In Problems 1, 2, and 3 we indicate how to generalize Theorems 1, 2, and 3 for a non-rectangular region. For example, let

$$D \equiv \{(x, y) \mid x^2 + y^2 < a^2\}, \quad C \equiv \{(x, y) \mid x^2 + y^2 = a^2\};$$

in the notation of Theorem 1, let  $P$  be any lattice point and define

$$D_\delta \equiv \{P \mid P, P_1, P_2, P_3, P_4 \in D\}.$$

Now, if  $P \in D$  but  $P \notin D_\delta$ , we set  $P \in C_\delta$  and note that at least one pair of its opposite neighbors is separated by  $C$ , say

$$P_1 \notin D, \quad P_2 \in D.$$

Let  $P_C \in C$  be on the line segment  $PP_1$ ; let  $\theta \equiv \text{distance } PP_C$ , therefore,  $0 < \theta \leq \delta x$ . Define  $U(P_C) \equiv u(P_C)$  for any point  $P_C$  on  $C$ .

**1. Maximum Principle:** In the above notation, for  $P \in D$  but  $P \notin D_\delta$ , define the *linear interpolation operator*

$$B_\delta U(P) \equiv \frac{\theta U(P_2) + \delta x U(P_C)}{\delta x + \theta} - U(P).$$

Show that

(a) If

$$\Delta_\delta U(P) \geq 0, \quad \text{for } P \in D_\delta,$$

and

$$B_\delta U(P) \geq 0, \quad \text{for } P \in C_\delta,$$

then

$$\max_{P \in D} U(P) \leq \max_{P_C \in C} U(P_C).$$

(b) If

$$\Delta_\delta U(P) \leq 0, \quad \text{for } P \in D_\delta;$$

$$B_\delta U(P) \leq 0, \quad \text{for } P \in C_\delta,$$

then

$$\min_{P \in D} U(P) \geq \min_{P_C \in C} U(P_C).$$

(c) The equations

$$-\Delta_\delta U(P) = f(P), \quad P \in D_\delta,$$

$$B_\delta U(P) = g(P), \quad P \in C_\delta,$$

have a unique solution.

2. With the *linear interpolation operator*, (i.e.,  $B_\delta U(P)$ ), prove the *a priori estimate*, for lattice points in  $D$ , and any lattice function  $U(P)$ ,

$$\max_{P \in D_\delta} |U(P)| \leq \max_{P_C \in C} |U(P_C)| + \frac{a^2}{2} K,$$

where

$$K = \max \left[ \max_{P \in D_\delta} |\Delta_\delta U(P)|, \quad \max_{P \in C_\delta} |B_\delta U(P)| \right].$$

3. Derive a bound for the error,  $E = U - u$ , when  $U$  is found with rounding errors  $\rho, \rho_1$  that satisfy

$$\begin{aligned} -\Delta_\delta U(P) &= f(P) + \frac{\rho}{\delta x \delta y}, & P \in D_\delta \\ B_\delta U(P) &= \rho_1, & P \in C_\delta. \end{aligned}$$

If  $u$  has continuous derivatives of fourth order and  $\delta x = \mathcal{O}(h)$ ,  $\delta y = \mathcal{O}(h)$ , show that

$$\max_{P \in D} |E(P)| = \mathcal{O}(h^2),$$

for sufficiently small  $\rho, \rho_1$ .

[Hint: Define the *truncation error* as in (7) for  $P \in D_\delta$ . Otherwise, set

$$\tau\{u(P)\} \equiv B_\delta u(P), \quad P \in C_\delta.$$

Apply the a priori estimate to  $E(P)$ , for  $P \in D$ .]

4. Show how the statements of the maximum principle, the a priori estimate, and the error bound must be modified for a more general bounded domain  $D$ .

5.\* With  $U$  and  $V$  in the class  $\mathcal{F}$  of Theorem 4, show that

$$\sum U_x V_x + U_{\bar{x}} V_{\bar{x}} + U_y V_y + U_{\bar{y}} V_{\bar{y}} = -2 \sum V \Delta_\delta U.$$

Hence

$$H[U + \epsilon V] \cong H[U] - \frac{2\epsilon}{L[U]} \sum V (\Delta_\delta U + \mu U).$$

Therefore if  $U$  satisfies (32),  $H[U]$  is stationary.

[Hint: Use summation by parts to remove the difference quotients of  $V$ . For example, in  $\sum U_x V_x$ , the value  $V(P)$  for a fixed  $P \in D_\delta$  occurs only in  $V_x(P_2)$  and  $V_x(P)$ . Thus in this sum the coefficient of  $V(P)$  is found to be:

$$-[U(P_2) - 2U(P) + U(P_1)]/\delta x^2.]$$

## 2. SOLUTION OF LAPLACE DIFFERENCE EQUATIONS

The linear algebraic system of equations determined by the difference scheme (1.5) is, for the rectangular region, of order  $JK$ . For small net spacings  $\delta x$  and  $\delta y$ , this may be extremely large since  $JK = \text{constant}/(\delta x \delta y)$ . (In practice,  $JK > 2500$  is not at all unusual.) Thus the standard elimination procedures for the equivalent system (1.18) of order  $JK$  require on the order of  $(JK)^3$  operations for solution and are too inefficient. Now from the definition (1.17b) of  $A$ , we see that many of its elements are zero and in fact, that it is block tridiagonal. The Gaussian elimination procedures which take account of large blocks of zero elements (in particular, the methods of Subsection 3.3 in Chapter 2) are then naturally suggested. This block elimination method requires at most on the order of  $J^3K$  operations (for rectangular regions) and is efficiently carried out on modern digital computers. (The storage requirements are for  $2K - 1$  matrices of order  $J$  and one vector of order  $JK$ . But this data is used only in dealing with systems of order  $J$  and hence is not all required at the same time.) In fact, since only tridiagonal systems need to be solved, efficient organization requires only  $\mathcal{O}(J^2K)$  operations!

Nevertheless, *iterative methods* seem to be the ones most often employed to solve the Laplace difference equations. Again, the large number of zero elements in the coefficient matrix greatly reduces the computational effort required in each iteration. However, some care must be taken to insure that sufficient accuracy will be obtained in a “reasonable” number of iterations. We consider such methods for the rectangular region.

The simplest iteration method begins with an initial estimate of the solution, say  $U^{(0)}$ , and then defines the sequence of net functions  $U^{(v)}$  by

$$(1a) \quad U^{(v+1)}(x, y) = U^{(v)}(x, y) + \delta^2 \Delta_\delta U^{(v)}(x, y) + \delta^2 f(x, y), \quad (x, y) \text{ in } D_\delta,$$

$$(1b) \quad U^{(v+1)}(x, y) = g(x, y), \quad (x, y) \text{ in } C_\delta, \quad v = 0, 1, \dots$$

Here  $\delta^2$  is defined in (1.14c) and the boundary condition (1.1b) is to be satisfied by  $U^{(0)}$ . From the other definitions in (1.14c) and (1.4), we find that (1a) is, in subscript notation,

$$(2) \quad U_{j,k}^{(v+1)} = \theta_x(U_{j-1,k}^{(v)} + U_{j+1,k}^{(v)}) + \theta_y(U_{j,k-1}^{(v)} + U_{j,k+1}^{(v)}) + \delta^2 f_{j,k}, \quad 1 \leq j \leq J, 1 \leq k \leq K.$$

[Note the relation between (2) and (1.15).] The calculations required in (1) or equivalently (2) can be carried out in any order on the net.

This iteration scheme is easily written in matrix form by using the notation of the previous subsection. We get that

$$(3a) \quad \mathbf{U}^{(v+1)} = (H + V)\mathbf{U}^{(v)} + \delta^2 \mathbf{F}, \quad v = 0, 1, \dots$$

Here the  $JK$ -order matrices are defined by (1.14b) and (1.17b),  $\mathbf{F}$  is defined in (1.16b) and (1.17a), and  $\mathbf{U}^{(v)}$  is the vector with components  $U_{j,k}^{(v)}$  ordered as in (1.17a). From the definition of  $A$  in (1.17b), we see that (3a) can be written as

$$(3b) \quad \mathbf{U}^{(v+1)} = (\mathbf{I} - A)\mathbf{U}^{(v)} + \delta^2\mathbf{F},$$

and thus, this scheme applied to (1.18) is a special case of the general iterative methods studied in Section 4 of Chapter 2. In fact, this is just the *Jacobi or simultaneous iteration scheme* of Subsection 4.1 in Chapter 2 applied to the system (1.18).

From the general theory of iterative methods, we know that the necessary and sufficient condition for the convergence of the sequence  $\{\mathbf{U}^{(v)}\}$  to the solution  $\mathbf{U}$  for an arbitrary initial guess  $\mathbf{U}^{(0)}$  is that all of the eigenvalues of  $(H + V)$  are in magnitude less than unity (see Theorem 4.1 of Chapter 2). The eigenvalues of this matrix are the roots of the characteristic polynomial

$$(4) \quad \Psi(\eta) \equiv \det |\eta\mathbf{I} - (H + V)| = \det |\eta\mathbf{I} - (I - A)|.$$

However, we have determined the eigenvalues of  $A$  in the previous subsection; they are given in (1.27). Thus the eigenvalues of  $(I - A)$ , and hence the roots of  $\Psi(\eta) = 0$ , are

$$(5a) \quad \begin{aligned} \eta &= \eta_{p,q} \equiv 1 - \Lambda_{p,q} \\ &= 1 - 4\theta_x \sin^2 \left( \frac{\pi}{2} \frac{p}{J+1} \right) - 4\theta_y \sin^2 \left( \frac{\pi}{2} \frac{q}{K+1} \right), \\ &\quad 1 \leq p \leq J, 1 \leq q \leq K. \end{aligned}$$

Now we easily find that  $-1 < \eta < 1$  and for small  $\delta x$  and  $\delta y$

$$(5b) \quad \begin{aligned} \rho(H + V) &\equiv \max_{p,q} |\eta_{p,q}| = \eta_{1,1} = 1 - \lambda_{1,1} \\ &= 1 - \delta^2 \pi^2 \left( \frac{1}{a^2} + \frac{1}{b^2} \right) + \mathcal{O}(\delta^4). \end{aligned}$$

Since  $0 < \lambda_{1,1} < 1$ , the method clearly converges and the rate of convergence is by (4.11) of Chapter 2

$$(6) \quad \begin{aligned} R_J &= -\log(1 - \lambda_{1,1}) \\ &= \delta^2 \pi^2 \left( \frac{1}{a^2} + \frac{1}{b^2} \right) + \mathcal{O}(\delta^4). \end{aligned}$$

We find that the rate of convergence decreases with

$$\delta^2 = \frac{\delta x^2 \delta y^2}{2[(\delta x)^2 + (\delta y)^2]},$$

and thus for difference equations with a small net spacing we may expect very slow convergence.

The *Gauss-Seidel or successive iteration* method for the Laplace difference equations can be written as

$$(7a) \quad U_{j,k}^{(v+1)} = \theta_x(U_{j-1,k}^{(v+1)} + U_{j+1,k}^{(v)}) + \theta_y(U_{j,k-1}^{(v+1)} + U_{j,k+1}^{(v)}) + \delta^2 f_{j,k}, \\ 1 \leq j \leq J, 1 \leq k \leq K.$$

In matrix form this becomes, with the use of (1.17b),

$$(7b) \quad [I - (\theta_x L + \theta_y B)] \mathbf{U}^{(v+1)} = (\theta_x L + \theta_y B)^T \mathbf{U}^{(v)} + \delta^2 \mathbf{F}.$$

In the present application this iteration scheme is frequently called the *Liebmann method*. The new iterates cannot be evaluated in a completely arbitrary order in this method. We first compute  $U_{1,1}^{(v+1)}$  and then, in order, the other elements on the coordinate lines with  $j = 1$  and  $k = 1$ . Next  $U_{2,2}^{(v+1)}$  is determined, etc. By slight changes in the scheme we could start the calculations at either of the other three “corners” in  $D_\delta$ . However, as we shall see, all of these methods have the same rate of convergence. This successive scheme is easier to employ on a digital computer than the simultaneous scheme since now each new component can immediately replace the previous value in storage. In addition, we shall find that the Gauss-Seidel method converges exactly twice as fast as the Jacobi method (when they are used on the same problem) and thus one should *never use the Jacobi method* on such difference problems.

The convergence of the iteration method (7) is determined by the magnitude of the eigenvalues of the matrix

$$(8) \quad S_1 \equiv [I - (\theta_x L + \theta_y B)]^{-1}(\theta_x L + \theta_y B)^T.$$

The indicated inverse exists since  $\theta_x L + \theta_y B$  is a strictly lower triangular matrix. Thus the eigenvalues,  $\xi$ , of the matrix  $S_1$  are the roots of the characteristic polynomial

$$(9) \quad \Phi_1(\xi) \equiv \det |\xi[I - (\theta_x L + \theta_y B)] - (\theta_x L + \theta_y B)^T| \\ = \det |\xi I - \theta_x(\xi L + L^T) - \theta_y(\xi B + B^T)|.$$

To examine the roots of this polynomial we shall use the following

**THEOREM 1.** *Let the matrices  $L$ ,  $B$ , and  $A$  be defined as in (1.14b) and (1.17b). Then for any non-zero scalars  $\alpha$  and  $\beta$*

$$(10) \quad \det |A| = \det |I - \theta_x(\alpha L + \alpha^{-1}L^T) - \theta_y(\beta B + \beta^{-1}B^T)|.$$

*Proof.* Let the elements of  $A$  be  $a_{r,s}$  where  $r, s = 1, 2, \dots, N = JK$ . Then each term in the formal expansion of  $\det |A|$  is given by a product of the form

$$\pm a_{1,\pi(1)} a_{2,\pi(2)} \cdots a_{r,\pi(r)} \cdots a_{N,\pi(N)}.$$

Here  $\pi$  is one of the  $N!$  permutations of the first  $N$  integers. Let each point  $(x_r, y_k)$  of the net  $D_\delta$  be identified with a unique integer (see Figure 1)

$$(x_r, y_k) \leftrightarrow r \equiv j + (k - 1)J.$$

Then any given permutation  $\pi$  can be represented by  $N$  vectors on  $D_\delta$  by drawing lines from  $r$  to  $\pi(r)$  for  $1 \leq r \leq N$  [i.e., from the point corresponding to  $r$  to the point corresponding to  $\pi(r)$ ]. Now by the definition of the matrix  $A$  it follows that  $a_{r,\pi(r)} \neq 0$  only if  $r = \pi(r)$  or the point corresponding to  $\pi(r)$  is one of the four neighboring net points,  $(x \pm \delta x, y)$  or  $(x, y \pm \delta y)$ , in the star about the point  $(x, y)$  corresponding to  $r$ . Thus, the only terms in the expansion of  $\det |A|$  which may not vanish correspond to permutations whose geometric representation is composed entirely of unit vectors in the  $(\pm x)$  and  $(\pm y)$  directions and null vectors.

Now every permutation is a product of disjoint cycles and in the above representation a cycle is a closed path of vectors on  $D_\delta$  (see Figure 1). Thus for any cycle corresponding to a non-vanishing product of elements, there must be the same number of unit vectors in the  $(+x)$  direction as in the  $(-x)$  direction and similarly for the  $(\pm y)$  directions. Now we recall that  $a_{r,\pi(r)}$  is an element of  $L$  if  $\pi(r) = r - 1$  and is an element of  $L^T$  if  $\pi(r) = r + 1$ . Thus there are as many factors from  $L$  as from  $L^T$  in any non-

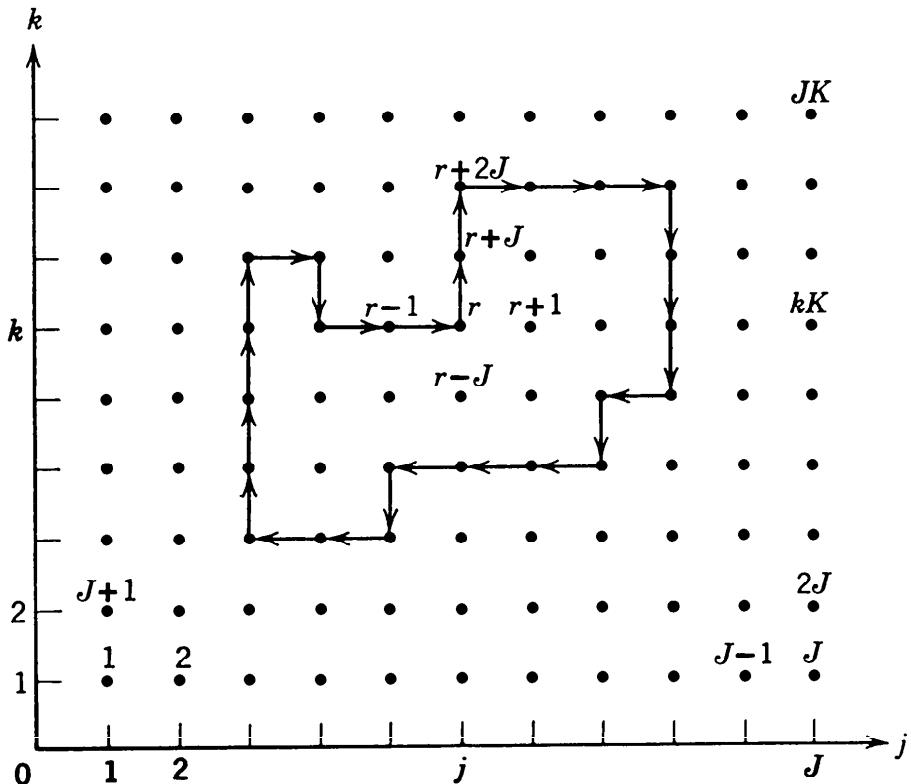


Figure 1. Geometric representation of a non-vanishing cycle. In the permutation  $n \rightarrow \pi(n)$ , of which this cycle is a factor,  $r = j + (k - 1)J$  and the cycle is given by:

$$\pi(r) = r + J, \pi(r + J) = r + 2J, \dots, \pi(r - 1) = r.$$

vanishing term in the expansion. Similarly,  $a_{r,\pi(r)}$  is an element of  $B$  if  $\pi(r) = r - J$  and of  $B^T$  if  $\pi(r) = r + J$ . Thus factors from  $B$  and  $B^T$  also enter pairwise in any non-vanishing term.

These results hold for the expansion of the right-hand determinant in (10) if elements from  $L$ ,  $L^T$ ,  $B$ , and  $B^T$  are replaced by those of  $\alpha L$ ,  $\alpha^{-1}L^T$ ,  $\beta B$ , and  $\beta^{-1}B^T$  respectively. Thus in any non-vanishing term the scalars  $\alpha$  and  $\beta$  do not appear and the proof is concluded. ■

We note that the proof of Theorem 1 depends *only upon the location of the zero elements in the matrix A*. Hence if the non-zero elements of  $A$  are changed in any manner we have the

**COROLLARY.** *If  $\mu$ ,  $\xi$ ,  $\eta$ ,  $\zeta$ , and  $\nu$  are matrices with zero components wherever  $I$ ,  $L$ ,  $L^T$ ,  $B$ , and  $B^T$  respectively have zero components, then*

$$\det |\mu - (\alpha\xi + \alpha^{-1}\eta) - (\beta\zeta + \beta^{-1}\nu)| = \det |\mu - (\xi + \eta + \zeta + \nu)|. \quad \blacksquare$$

In particular, we now consider the determinant  $\Phi(\xi)$  in (9). The identity  $I$  and the matrices  $L$  and  $B$  have been multiplied by a scalar  $\xi$ , so no zero elements of  $A$  have been altered. Thus we may apply the corollary to get

$$\Phi_1(\xi) = \det |\xi I - \theta_x(\alpha\xi L + \alpha^{-1}L^T) - \theta_y(\beta\xi B + \beta^{-1}B^T)|.$$

Take  $\alpha = \beta = \xi^{-1/2}$  and recall (4) to find that

$$\begin{aligned} \Phi_1(\xi) &= \det |\xi^{1/2}I| \cdot \det |\xi^{1/2}I - (H + V)| \\ &= \xi^{JK/2}\Psi(\xi^{1/2}). \end{aligned}$$

Thus every non-zero root  $\xi$  of  $\Phi_1(\xi) = 0$  satisfies  $\Psi(\xi^{1/2}) = 0$  and every root  $\eta$  of  $\Psi(\eta) = 0$  satisfies  $\Phi_1(\eta^2) = 0$ . So all non-zero eigenvalues of the matrix  $S_1$  in (8) are given by

$$\xi = \eta^2 = (1 - \lambda_{p,q})^2, \quad 1 \leq p \leq J, 1 \leq q \leq K,$$

and from (5) we find that the maximum eigenvalue of  $S_1$  is

$$\begin{aligned} \rho(S_1) &= [\rho(H + V)]^2 = (1 - \lambda_{1,1})^2 \\ &= 1 - 2\delta^2\pi^2\left(\frac{1}{a^2} + \frac{1}{b^2}\right) + \mathcal{O}(\delta^4). \end{aligned}$$

The rate of convergence for the Gauss-Seidel scheme is thus

$$(12) \quad R_{GS} = 2\delta^2\pi^2\left(\frac{1}{a^2} + \frac{1}{b^2}\right) + \mathcal{O}(\delta^4),$$

or twice that for the Jacobi scheme.

The convergence rate of the Gauss-Seidel method (7) may be improved

by introducing an appropriate *acceleration parameter*, as discussed in Section 5 of Chapter 2. That is, set

$$(13a) \quad V_{j,k}^{(v+1)} = \theta_x(U_{j-1,k}^{(v+1)} + U_{j+1,k}^{(v)}) + \theta_y(U_{j,k-1}^{(v+1)} + U_{j,k+1}^{(v)}) + \delta^2 f_{j,k}$$

and then, at the point  $(x_j, y_k)$ , take

$$(13b) \quad \begin{aligned} U_{j,k}^{(v+1)} &= \omega V_{j,k}^{(v+1)} + (1 - \omega)U_{j,k}^{(v)} \\ &= U_{j,k}^{(v)} + \omega(V_{j,k}^{(v+1)} - U_{j,k}^{(v)}). \end{aligned}$$

Here  $\omega$  is the acceleration parameter to be determined. We note that for  $\omega = 1$  this scheme reduces to that in (7a), i.e., to the ordinary Gauss-Seidel method. The order in which the components of the new iterates are to be computed is just as in the previous successive scheme.

To examine the convergence of the accelerated Gauss-Seidel method we first write it in matrix form. Obviously (13a) implies

$$(14a) \quad \mathbf{V}^{(v+1)} = (\theta_x L + \theta_y B)\mathbf{U}^{(v+1)} + (\theta_x L + \theta_y B)^T \mathbf{U}^{(v)} + \delta^2 \mathbf{F},$$

and (13b) implies

$$(14b) \quad \mathbf{U}^{(v+1)} = \omega \mathbf{V}^{(v+1)} + (1 - \omega) \mathbf{U}^{(v)}.$$

Upon eliminating  $\mathbf{V}^{(v+1)}$ , we obtain

$$(15) \quad [I - \omega(\theta_x L + \theta_y B)]\mathbf{U}^{(v+1)} = [(1 - \omega)I + \omega(\theta_x L + \theta_y B)^T]\mathbf{U}^{(v)} + \omega\delta^2 \mathbf{F}.$$

The convergence of these iterations is thus determined by the magnitude of the eigenvalues of the matrix

$$(16) \quad S_\omega \equiv [I - \omega(\theta_x L + \theta_y B)]^{-1}[(1 - \omega)I + \omega(\theta_x L + \theta_y B)^T].$$

Note that for  $\omega = 1$  the above matrix reduces to the  $S_1$  defined in (8) for the ordinary unaccelerated successive iterations. The eigenvalues of  $S_\omega$  are the roots  $\zeta$  of the characteristic polynomial

$$(17) \quad \begin{aligned} \Phi_\omega(\zeta) &\equiv \det |[I - \omega(\theta_x L + \theta_y B)]\zeta - [(1 - \omega)I + \omega(\theta_x L + \theta_y B)^T]| \\ &= \det |(\zeta + \omega - 1)I - \omega\zeta(\theta_x L + \theta_y B) - \omega(\theta_x L + \theta_y B)^T|. \end{aligned}$$

The matrix in (17) has zero elements wherever the matrix  $A$  has them and so the corollary to Theorem 1 is applicable. If we use the scalars  $\alpha = \beta = \zeta^{-1/2}$  then we obtain from (17) and (4)

$$\begin{aligned} \Phi_\omega(\zeta) &= \det |(\zeta + \omega - 1)I - \omega\zeta^{1/2}(\theta_x L + \theta_y B) - \omega\zeta^{1/2}(\theta_x L + \theta_y B)^T| \\ &= \det |\omega\zeta^{1/2}I| \cdot \det \left| \frac{(\zeta + \omega - 1)}{\omega\zeta^{1/2}} I - (H + V) \right| \\ &= (\omega\zeta^{1/2})^{JK} \Psi \left( \frac{\zeta + \omega - 1}{\omega\zeta^{1/2}} \right). \end{aligned}$$

From this result we conclude, for each  $\omega \neq 0$ , that any non-zero root  $\zeta$  of  $\Phi_\omega(\zeta) = 0$  satisfies  $\Psi(\eta) = 0$ , and that every root  $\eta$  of  $\Psi(\eta) = 0$  satisfies  $\Phi_\omega(\zeta) = 0$  provided  $(\zeta + \omega - 1)/(\omega\zeta^{1/2}) = \eta$ . Thus the non-zero eigenvalues of the matrix  $S_\omega$  are the roots  $\zeta$  of

$$(18a) \quad \zeta + \omega - 1 = \omega\zeta^{1/2}\eta$$

where  $\eta$  ranges over the roots of  $\Psi(\eta) = 0$  [i.e., the eigenvalues of  $I - A$  given in (5a)]. Since (18a) is quadratic in  $\zeta^{1/2}$ , we find that all  $\zeta$  which satisfy this equation are given by

$$(18b) \quad \zeta = \zeta_{\pm} = \left[ \left( \frac{\omega\eta}{2} \right) \pm \sqrt{\left( \frac{\omega\eta}{2} \right)^2 + (1 - \omega)} \right]^2.$$

We may now determine  $\omega$  such that the iteration scheme (15) converges. First observe that since  $\eta$  is real it follows from (18b) that  $|\zeta_-| \geq 1$  for  $\omega \leq 0$ . Thus an eigenvalue of  $S_\omega$  will have magnitude larger than unity and we conclude that the accelerated Gauss-Seidel method is not convergent for any non-positive  $\omega$ . For fixed  $\omega > 0$  we see that some eigenvalues may be complex (only if  $\omega > 1$ ) but then their magnitude is

$$(19a) \quad |\zeta| = \omega - 1.$$

For the real eigenvalues it follows from (18b) with  $\omega > 0$  and  $\eta > 0$  that  $\zeta_+$  is an increasing function of  $\eta$  and that  $|\zeta_+| > |\zeta_-|$ . Thus the largest real eigenvalue of  $S_\omega$  is, since  $\eta \leq \eta_{1,1}$ ,

$$(19b) \quad \zeta = \zeta_1(\omega) \equiv \left[ \frac{\omega\eta_{1,1}}{2} + \sqrt{\left( \frac{\omega\eta_{1,1}}{2} \right)^2 + (1 - \omega)} \right]^2.$$

From (19) we obtain for  $\omega > 0$

$$\rho(S_\omega) = \max [\omega - 1, \zeta_1(\omega)].$$

As  $0 < \eta_{1,1} < 1$  it follows that  $\rho(S_\omega) < 1$  if  $0 < \omega < 2$  since in this interval, when  $\zeta_1$  is real,

$$\begin{aligned} \zeta_1(\omega) &= \left[ \left( \frac{\omega\eta_{1,1}}{2} \right) + \sqrt{\left( 1 - \frac{\omega}{2} \right)^2 - \left( \frac{\omega}{2} \right)^2 (1 - \eta_{1,1}^2)} \right]^2, \\ &< \left[ \left( \frac{\omega}{2} \right) + \sqrt{\left( 1 - \frac{\omega}{2} \right)^2} \right]^2 = 1. \end{aligned}$$

On the other hand, if  $\omega \geq 2$  then  $\zeta_1(\omega)$  is complex, and by (19a) some eigenvalue has modulus not less than unity. Thus we have

**THEOREM 2.** *The accelerated Gauss-Seidel iterations converge iff the acceleration parameter  $\omega$  lies in the interval  $0 < \omega < 2$ .* ■

The optimal value for the acceleration parameter is that value  $\omega = \omega^*$ , in  $0 < \omega < 2$ , for which

$$\rho(S_{\omega^*}) = \min_{0 < \omega < 2} \rho(S_\omega) = \min_{0 < \omega < 2} \{\max [\omega - 1, \zeta_1(\omega)]\}.$$

[We know that the indicated minimum exists since  $\rho(S_\omega)$  is continuous in  $0 \leq \omega \leq 2$  and satisfies  $\rho(S_0) = \rho(S_2) = 1$ ,  $\rho(S_\omega) < 1$  in  $0 < \omega < 2$ .] It is clear that the expression in the radical of (19b) is a decreasing function of  $\omega$  for  $0 < \omega < 2$ . Thus  $\zeta_1(\omega)$  becomes complex when this expression vanishes, i.e., for

$$\omega = \omega_b = \frac{2}{1 + \sqrt{1 - \eta_{1,1}^2}}.$$

For  $\omega \geq \omega_b$  we have now  $\rho(S_\omega) = \omega - 1$  and

$$\min_{\omega_b \leq \omega < 2} \rho(S_\omega) = \omega_b - 1.$$

For  $0 < \omega \leq \omega_b$ , since  $\zeta_1'(\omega) < 0$ ,  $\zeta_1(\omega)$  is a decreasing function of  $\omega$ . Hence  $\rho(S_{\omega^*})$  occurs for  $\omega^* = \omega_b$  at which  $\omega_b - 1 = \zeta_1(\omega_b) = \rho(S_{\omega^*})$ . Thus, in summary, we have for the *optimal application of the accelerated Gauss-Seidel method*

$$(20a) \quad \omega^* = \frac{2}{1 + \sqrt{1 - \eta_{1,1}^2}},$$

$$(20b) \quad \rho(S_{\omega^*}) = \omega^* - 1 = \frac{1 - \sqrt{1 - \eta_{1,1}^2}}{1 + \sqrt{1 - \eta_{1,1}^2}}.$$

From (5), we have

$$\rho(S_{\omega^*}) = 1 - 2\delta\pi \sqrt{2\left(\frac{1}{a^2} + \frac{1}{b^2}\right)} + \mathcal{O}(\delta^2)$$

and so the rate of convergence is now

$$(21) \quad R_{\text{AGS}} = 2\delta\pi \sqrt{2\left(\frac{1}{a^2} + \frac{1}{b^2}\right)} + \mathcal{O}(\delta^2).$$

By comparing (21) with (6) and (12), we see that the power of  $\delta$  in the rate of convergence for the optimal accelerated Gauss-Seidel method is lower than the power of  $\delta$  appearing in the ordinary Gauss-Seidel or Jacobi methods. The same result is obtained if the iterations were to proceed in one of the other orders indicated after (7b). This is suggested by the form of our results in which the coordinate directions and related dimensions enter symmetrically (see, however, the discussion at the end of the next subsection).

## 2.1. Line or Block Iterations

Since the linear system (1.18), which we are solving iteratively, has the simple block structure indicated in (1.16a) it is rather natural to consider corresponding block iterations (i.e., Subsection 4.3 of Chapter 2). In the present application, these are more properly called “line” methods since the net function is altered by changing the data on a complete coordinate line of net points in  $D_\delta$  simultaneously. A particularly simple line iteration for the system in (1.16) is

$$(22a) \quad \begin{aligned} [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_1^{(v+1)} - \theta_y\mathbf{U}_2^{(v)} &= \delta^2\mathbf{F}_1, \\ -\theta_y\mathbf{U}_{k-1}^{(v)} + [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_k^{(v+1)} - \theta_y\mathbf{U}_{k+1}^{(v)} &= \delta^2\mathbf{F}_k, \\ 2 \leq k \leq K-1, \\ -\theta_y\mathbf{U}_{K-1}^{(v)} + [I_J - \theta_x(L_J + L_J^T)]\mathbf{U}_K^{(v+1)} &= \delta^2\mathbf{F}_K. \end{aligned}$$

The  $K$  systems for the  $\mathbf{U}_k^{(v+1)}$  can be solved in any order. At each of the  $K$  steps in one of these iterations, a linear system of order  $J$  must be solved with the coefficient matrix  $I_J - \theta_x(L_J + L_J^T)$ . However, this matrix is tridiagonal and can easily be factored by the method of Subsection 3.2 in Chapter 2. This is done only once and then each linear system in the succeeding iterations is solved by evaluating two simple recursions of the forms (3.12) and (3.13) of Chapter 2. The present scheme is frequently called a *line Jacobi method*.

By using the matrices and vectors in (1.17) we can write the iterative scheme (22a) as [compare with (3a)]

$$(22b) \quad (I - H)\mathbf{U}^{(v+1)} = V\mathbf{U}^{(v)} + \delta^2\mathbf{F}.$$

The convergence is thus determined by the matrix

$$(23a) \quad (I - H)^{-1}V$$

whose eigenvalues,  $\rho$ , are the roots of the characteristic polynomial

$$(23b) \quad P(\rho) \equiv \det |\rho I - \rho H - V|.$$

It is not difficult to show that the matrices  $H$  and  $V$  have common eigenvectors (since they are symmetric and commute). In fact, the eigenvalues and eigenvectors of these matrices are easily computed. Just as the eigenvalue problems (1.21) and (1.22) are equivalent, it follows that the following pairs of eigenvalue problems are also equivalent

$$(24a) \quad (2\theta_x I - H)\mathbf{W} = \xi\mathbf{W}, \quad \begin{cases} -W_{x\bar{x}} = (\xi/\delta^2)W \text{ on } D_\delta, \\ W = 0 \quad \text{on } C_\delta; \end{cases}$$

$$(24b) \quad (2\theta_y I - V)\mathbf{W} = \eta\mathbf{W}, \quad \begin{cases} -W_{y\bar{y}} = (\eta/\delta^2)W \text{ on } D_\delta, \\ W = 0 \quad \text{on } C_\delta. \end{cases}$$

[In fact, if we set  $\Lambda = \xi + \eta$  and add corresponding equations we obtain (1.21) and (1.22), by using  $\theta_x + \theta_y = \frac{1}{2}$ .] The problems in (24) may be solved by separating variables and recalling (1.23)–(1.27). We find that these problems have common eigenvectors  $\mathbf{W}_{p,q}$  with the components

$$(25) \quad W_{p,q}(x_j, y_k) = \sin\left(j \frac{p\pi}{J+1}\right) \sin\left(k \frac{q\pi}{K+1}\right),$$

$$1 \leq p \leq J, \quad 1 \leq q \leq K,$$

and the eigenvalues

$$(26a) \quad \xi = \xi_p = 4\theta_x \sin^2\left(\frac{\pi}{2} \frac{p}{J+1}\right) = 2\theta_x \left[1 - \cos\left(\pi \frac{p}{J+1}\right)\right],$$

$$1 \leq p \leq J;$$

$$(26b) \quad \eta = \eta_q = 4\theta_y \sin^2\left(\frac{\pi}{2} \frac{q}{K+1}\right) = 2\theta_y \left[1 - \cos\left(\pi \frac{q}{K+1}\right)\right],$$

$$1 \leq q \leq K.$$

Each eigenvalue  $\xi_p$  of the problem (24a) has multiplicity  $K$  and each eigenvalue  $\eta_q$  of (24b) has multiplicity  $J$ . The eigenvalues of  $H$  and  $V$  are easily obtained from the above and are

$$2\theta_x \cos\left(\pi \frac{p}{J+1}\right) \quad \text{and} \quad 2\theta_y \cos\left(\pi \frac{q}{K+1}\right),$$

respectively.

The vectors  $\mathbf{W}_{p,q}$  are also eigenvectors of  $(I - H)^{-1}V$  and multiplication by this matrix yields the eigenvalues, which are also the roots of  $P(\rho) = 0$

$$(27) \quad \rho_{p,q} = \frac{2\theta_y \cos \frac{q\pi}{K+1}}{1 - 2\theta_x \cos \frac{p\pi}{J+1}}, \quad 1 \leq p \leq J, \quad 1 \leq q \leq K.$$

The maximum magnitude of the eigenvalues is found by the usual expansions and some simplification to be

$$(28a) \quad \max_{p,q} |\rho_{p,q}| = \rho_{1,1} = 1 - \frac{\delta y^2}{2} \pi^2 \left( \frac{1}{a^2} + \frac{1}{b^2} \right) + \mathcal{O}(\delta^4).$$

Hence the rate of convergence for this *line Jacobi scheme* is

$$(28b) \quad R_{LJ} = \frac{\delta y^2}{2} \pi^2 \left( \frac{1}{a^2} + \frac{1}{b^2} \right) + \mathcal{O}(\delta^4).$$

Note the similarity between this result and that in (6) for the (point) Jacobi iterations and that in (12) for the successive iterations. If  $\delta x = \delta y$ ,

then  $\delta y^2 = 4\delta^2$  and the above rate is essentially that of the Gauss-Seidel method given in (12).

Of course, an analog of the method of successive iterations is also possible for the line methods. We need only use the latest improved data as soon as it is obtained. Thus in (22) we replace  $\mathbf{U}_{k-1}^{(v)}$  by  $\mathbf{U}_{k-1}^{(v+1)}$  for  $k = 2, 3, \dots, K$ , to obtain a *line Gauss-Seidel scheme*. In matrix form this successive-line method is written as

$$(29) \quad (I - H - \theta_y B)\mathbf{U}^{(v+1)} = \theta_y B^T \mathbf{U}^{(v)} + \delta^2 \mathbf{F}.$$

However, an accelerated version of these iterations is of interest, and we directly consider this more general procedure. As before, an intermediate iterate  $\mathbf{V}^{(v+1)}$  is defined by

$$(30a) \quad (I - H)\mathbf{V}^{(v+1)} = \theta_y B\mathbf{U}^{(v+1)} + \theta_y B^T \mathbf{U}^{(v)} + \delta^2 \mathbf{F}.$$

Then with an arbitrary parameter  $\omega$  we set

$$(30b) \quad \mathbf{U}^{(v+1)} = \omega \mathbf{V}^{(v+1)} + (1 - \omega) \mathbf{U}^{(v)}.$$

The calculations are performed a line at a time, as in the line Jacobi method, to determine the  $\mathbf{V}_k^{(v+1)}$  and then the  $\mathbf{U}_k^{(v+1)}$  before going on to  $k + 1$ . However, now they must be done in a fixed order (say increasing or decreasing  $k$ ). For  $\omega = 1$  this scheme reduces to that in (29).

The *accelerated successive-line method* becomes upon the elimination of  $\mathbf{V}^{(v+1)}$  in (30)

$$\begin{aligned} (I - H - \omega \theta_y B)\mathbf{U}^{(v+1)} \\ = [(1 - \omega)I - (1 - \omega)H + \omega \theta_y B^T] \mathbf{U}^{(v)} + \omega \delta^2 \mathbf{F}. \end{aligned}$$

To examine the convergence of the scheme, we must determine the eigenvalues of the matrix

$$(31a) \quad T_\omega \equiv (I - H - \omega \theta_y B)^{-1}[(1 - \omega)I - (1 - \omega)H + \omega \theta_y B^T],$$

which is determined from the roots  $\tau$  of the characteristic polynomial

$$(31b) \quad Q_\omega(\tau) \equiv \det |(\tau + \omega - 1)(I - H) - \tau \omega \theta_y B - \omega \theta_y B^T|.$$

We see that the matrix in (31b) has zero elements wherever the matrix  $A$  has them and so just as in (17), we can apply the corollary to Theorem 1. With the scalars  $\alpha = 1$  and  $\beta = \tau^{-1/2}$  we then get from (31b) and (23b)

$$\begin{aligned} Q_\omega(\tau) &= \det |(\tau + \omega - 1)(I - H) - \omega \tau^{1/2} V| \\ &= \det |\omega \tau^{1/2} I| \cdot \det \left| \frac{(\tau + \omega - 1)}{\omega \tau^{1/2}} (I - H) - V \right| \\ &= (\omega \tau^{1/2})^{JK} P \left( \frac{\tau + \omega - 1}{\omega \tau^{1/2}} \right). \end{aligned}$$

It follows that  $\tau$  and the roots  $\rho$  of  $P(\rho) = 0$  are related by

$$(32) \quad \tau + \omega - 1 = \omega\tau^{1/2}\rho,$$

by the reasoning that led to (18a). For  $\omega = 1$ , the iterations reduce to the ordinary successive-line iterations and the non-zero roots  $\tau$  are given by  $\tau = \rho^2$ . Thus this method converges twice as fast as the line Jacobi method. Finally, since the eigenvalues  $\rho$  lie in  $0 < \rho < 1$ , the arguments used in (18)–(21) can be applied to the roots  $\tau(\omega)$  and the acceleration parameter  $\omega$ , which satisfy (32). Now the optimal parameter value  $\omega_*$  and minimum value  $\rho(T_{\omega_*})$  of  $\rho(T_\omega)$  become, where  $\rho(T)$  denotes the spectral radius of  $T$ ,

$$(33a) \quad \omega_* = \frac{2}{1 + \sqrt{1 - \rho_{1,1}^2}},$$

$$(33b) \quad \rho(T_{\omega_*}) = \omega_* - 1 = \frac{1 - \sqrt{1 - \rho_{1,1}^2}}{1 + \sqrt{1 - \rho_{1,1}^2}}.$$

By using (28a), we find

$$\rho(T_{\omega_*}) = 1 - 2\delta y \pi \sqrt{\left(\frac{1}{a^2} + \frac{1}{b^2}\right)} + \mathcal{O}(\delta^2),$$

and hence the rate of convergence of the optimum *accelerated line Gauss-Seidel* method is

$$(34) \quad R_{\text{ALGS}} = 2\delta y \pi \sqrt{\left(\frac{1}{a^2} + \frac{1}{b^2}\right)} + \mathcal{O}(\delta^2).$$

To compare rates of convergence we note, using (21), that

$$(35) \quad \begin{aligned} \frac{R_{\text{ALGS}}}{R_{\text{AGS}}} &= \frac{\delta y}{\sqrt{2} \delta} + \mathcal{O}(\delta^2) \\ &= \left[1 + \left(\frac{\delta y}{\delta x}\right)^2\right]^{1/2} + \mathcal{O}(\delta^2). \end{aligned}$$

Thus it follows that *for any mesh ratio*,  $\delta y / \delta x$ , the optimum accelerated successive-line method has a larger rate of convergence than the corresponding optimum accelerated successive (point) iterations. For equal net spacing in the  $x$ - and  $y$ -directions the factor of improvement is, asymptotically,  $\sqrt{2}$ . However, if  $\delta y > \delta x$ , even greater improvement results. We observe here that the net lines along which the new data are obtained at each step should be in the direction of the smallest mesh width; i.e., the “closest” neighbors are grouped together on a line and improved as a group. All of the above could be repeated with  $H$  and  $V$

interchanged which corresponds to taking lines in the  $y$ -direction. The only change in (35) that would result is the interchange of  $\delta x$  and  $\delta y$ . A decision as to whether the ALGS scheme is more efficient than the AGS scheme must depend upon the size of

$$\frac{\# \text{ ops}_{\text{ALGS}}}{\# \text{ ops}_{\text{AGS}}} \equiv \alpha.$$

$\alpha$  measures the ratio of the amounts of work involved in one iteration step for each of the two methods. If

$$\alpha \frac{R_{\text{ALGS}}}{R_{\text{AGS}}} < 1,$$

then the ALGS scheme is more efficient; otherwise, the AGS scheme is more efficient.

## 2.2. Alternating Direction Iterations

One of the most effective iteration schemes for solving the system (1.16) or (1.18) employs a combination of horizontal and vertical line iterations. In terms of an acceleration parameter  $\omega$ , and recalling that  $2\theta_x + 2\theta_y = 1$ , such a scheme due to Peaceman and Rachford can be defined as follows:

$$(36a) \quad [(\omega + 2\theta_x)I - H]\mathbf{U}^{v+\frac{1}{2}} = [(\omega - 2\theta_y)I + V]\mathbf{U}^v + \delta^2\mathbf{F},$$

$$(36b) \quad [(\omega + 2\theta_y)I - V]\mathbf{U}^{v+1} = [(\omega - 2\theta_x)I + H]\mathbf{U}^{v+\frac{1}{2}} + \delta^2\mathbf{F}.$$

The vector  $\mathbf{U}^{v+\frac{1}{2}}$  is an intermediate quantity used to define the scheme and of course it is actually computed in carrying out the procedure. The first step, (36a), is just a horizontal line scheme, similar to line-Jacobi. (In fact, with  $\omega = 2\theta_y$  in (36a), we obtain (22b) with  $\mathbf{U}^{v+1}$  replaced by  $\mathbf{U}^{v+\frac{1}{2}}$ .) Clearly then (36b) is essentially a vertical line-Jacobi iteration. The vector to be found in each of the stages (36a and b) is easily evaluated by solving a tridiagonal system.

To study the convergence of this scheme, we eliminate  $\mathbf{U}^{v+\frac{1}{2}}$  in (36) and obtain, assuming for the moment that the required inverses exist,

$$\mathbf{U}^{v+1} = Q_\omega \mathbf{U}^v + \mathbf{f}_\omega,$$

where

$$(37) \quad Q_\omega \equiv [(\omega + 2\theta_y)I - V]^{-1}[(\omega - 2\theta_x)I + H] \\ \times [(\omega + 2\theta_x)I - H]^{-1}[(\omega - 2\theta_y)I + V],$$

and

$$\mathbf{f}_\omega \equiv [(\omega + 2\theta_y)I - V]^{-1} \\ \times \{[(\omega - 2\theta_x)I + H][(\omega + 2\theta_x)I - H]^{-1} + I\}\delta^2\mathbf{F}.$$

The eigenvalues of  $Q_\omega$  are easily obtained since the matrices  $(2\theta_x I - H)$

and  $(2\theta_y I - V)$  have common eigenvectors given in (25). We obtain using (24) and the eigenvalues given in (26)

$$\mathcal{Q}_\omega \mathbf{W}_{p,q} = \frac{(\omega - \xi_p)(\omega - \eta_q)}{(\omega + \xi_p)(\omega + \eta_q)} \mathbf{W}_{p,q}.$$

Thus the eigenvalues, say  $\lambda(\omega)$ , of  $\mathcal{Q}_\omega$  are

$$(38) \quad \lambda_{p,q}(\omega) \equiv \frac{(\omega - \xi_p)(\omega - \eta_q)}{(\omega + \xi_p)(\omega + \eta_q)}, \quad \begin{cases} p = 1, 2, \dots, J, \\ q = 1, 2, \dots, K. \end{cases}$$

Since  $\xi_p > 0$  and  $\eta_q > 0$  for all  $p$  and  $q$ , it follows that the *alternating direction scheme* (36) converges for any choice of  $\omega > 0$ . We also note that all relevant inverses exist for positive  $\omega$ .

The trick in the proper use of the alternating direction type schemes is *not* to use a single acceleration parameter  $\omega$  as above but rather to use a sequence of them, say  $\omega_1, \omega_2, \dots, \omega_m$  applied periodically (or cyclically). That is, the calculations in (36) are to be carried out  $m$  times (using each  $\omega_i$  for a complete double sweep of the net) in order to compute  $\mathbf{U}^{v+1}$  from  $\mathbf{U}^v$ . To actually write this scheme out we should introduce  $2m - 1$  intermediate quantities  $\mathbf{U}^{v+1/(2m)}, \mathbf{U}^{v+2/(2m)}, \dots, \mathbf{U}^{v+1-1/(2m)}$  and successively use (36a and b) for the pairs  $\mathbf{U}^{v+(2j-1)/(2m)}, \mathbf{U}^{v+2j/(2m)}$ . As before, we find that the eigenvalues which determine convergence are now

$$(39) \quad \lambda_{p,q}(\omega_1, \omega_2, \dots, \omega_m) \equiv \prod_{i=1}^m \frac{(\omega_i - \xi_p)(\omega_i - \eta_q)}{(\omega_i + \xi_p)(\omega_i + \eta_q)}, \quad \begin{cases} p = 1, 2, \dots, J, \\ q = 1, 2, \dots, K. \end{cases}$$

If we take  $m = J$  and choose  $\omega_j = \xi_j$  for  $j = 1, 2, \dots, J$ , then it clearly follows from (39) that

$$\lambda_{p,q}(\omega_1, \omega_2, \dots, \omega_J) = 0$$

for all  $p$  and  $q$ . In this case the exact solution is obtained in a finite number of steps. Of course we could also employ  $\omega_i = \eta_i$  with  $m = K$  to get similar results. However, both  $J$  and  $K$  are extremely large in general and we desired to obtain an accurate approximation in only  $v$  iterations where  $mv \ll J$  and  $mv \ll K$ . Thus we consider the problem, with fixed small  $m$ , to find  $\omega_i$  such that

$$\max_{p,q} |\lambda_{p,q}(\omega_1, \omega_2, \dots, \omega_m)|$$

is minimized with respect to all possible choices of the acceleration parameters  $\omega_i$ .

This problem is related to the subject of best approximations, Section 4 of Chapter 5. Specifically let us define the function

$$(40) \quad F(z) \equiv \prod_{i=1}^m \frac{(\omega_i - z)}{(\omega_i + z)}.$$

Then from (39) we have, recalling (26),

$$\max_{p,q} |\lambda_{p,q}(\omega_1, \omega_2, \dots, \omega_m)| \leq \max_{\substack{\xi_1 \leq x \leq \xi_J \\ \eta_1 \leq y \leq \eta_K}} |F(x)F(y)|.$$

Thus we seek  $\omega_i$  such that the rational function  $F(x)F(y)$  is the best (uniform) approximation to zero on the rectangle  $\xi_1 \leq x \leq \xi_J$ ,  $\eta_1 \leq y \leq \eta_K$ . The optimization problem is further simplified by noting that for all  $x, y$  on this rectangle

$$|F(x)F(y)| \leq \max_{\alpha \leq z \leq \beta} F^2(z) \equiv \|F(z)\|_\infty^2$$

where  $\alpha \equiv \min(\xi_1, \eta_1)$  and  $\beta \equiv \max(\xi_J, \eta_K)$ . Thus our problem is reduced to *finding the best approximation to zero of the form (40) on an interval*  $0 < \alpha \leq z \leq \beta$ . The existence and uniqueness of such a best rational approximation can be proved in a manner analogous to the treatment in Section 4 of Chapter 5 of best polynomial approximations. We shall not present the analysis here of how to determine the optimum parameters  $\omega_i$ . Rather, we show how to find a set of parameters  $\omega_i$ , for which we can estimate  $\|F\|_\infty$  in order to compare the rate of convergence of the cyclic alternating direction method with the previously studied iterative methods.

In Problem 1, we verify that for  $m = 1$  the choice  $\omega_1 = \sqrt{\alpha\beta}$  minimizes  $\|F\|_\infty$ , and

$$\|F\|_\infty = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}}.$$

Hence we divide the interval  $[\alpha, \beta]$  by points  $0 < \alpha_0 = \alpha < \alpha_1 < \dots < \alpha_m = \beta$ , such that

$$\frac{\alpha_0}{\alpha_1} = \frac{\alpha_1}{\alpha_2} = \dots = \frac{\alpha_{m-1}}{\alpha_m}.$$

The values  $\alpha$ , which have this property are

$$(41a) \quad \alpha_j = \alpha \left( \frac{\beta}{\alpha} \right)^{j/m}, \quad j = 0, 1, \dots, m.$$

We now set

$$(41b) \quad \omega_j = \sqrt{\alpha_{j-1}\alpha_j}$$

and find that, since the magnitude of each factor of  $F(z)$  is bounded by unity,

$$\max_{\alpha_{i-1} \leq z \leq \alpha_i} |F(z)| \leq \max_{\alpha_{i-1} \leq z \leq \alpha_i} \left| \frac{\omega_i - z}{\omega_i + z} \right|, \quad i = 1, 2, \dots, m.$$

On the other hand, with the choice (41), the result in Problem 1 implies that

$$\max_{\alpha_{i-1} \leq z \leq \alpha_i} \left| \frac{\omega_i - z}{\omega_i + z} \right| = \frac{1 - \sqrt{\frac{\alpha_{i-1}}{\alpha_i}}}{1 + \sqrt{\frac{\alpha_{i-1}}{\alpha_i}}} = \frac{1 - \left(\frac{\alpha}{\beta}\right)^{1/2m}}{1 + \left(\frac{\alpha}{\beta}\right)^{1/2m}}$$

for all  $i$ . Hence

$$\|F\|_\infty \leq \frac{1 - \left(\frac{\alpha}{\beta}\right)^{1/(2m)}}{1 + \left(\frac{\alpha}{\beta}\right)^{1/(2m)}}.$$

From the definitions of  $\alpha$  and  $\beta$ , and the results (1.25) and (1.26), we note that  $\alpha = \mathcal{O}(\delta^2)$ ,  $\beta \simeq 1$ . Therefore,

$$\|F\|_\infty \leq 1 - \mathcal{O}(\delta^{1/m}).$$

Hence we have shown that the rate of convergence of the cyclic alternating direction method is less than  $\mathcal{O}(\delta^{1/m})$ , for a complete cycle. But the amount of work required to compute  $m$  sweeps as in (36a and b) is equivalent to the work required for about  $2m$  applications of the line accelerated schemes. Now, the convergence rate of  $2m$  applications of a line accelerated scheme is  $\mathcal{O}(2m\delta)$ . This is much smaller than  $\mathcal{O}(\delta^{1/m})$ , the convergence rate for one cycle of the alternating direction method for small  $m$ . We have thus shown that the alternating direction method is more efficient than any of the other iterative schemes, even when parameters that are not necessarily optimal are employed. For detailed comparisons we refer to the book of Varga. In practice it is wise to start each cycle with the largest parameter value,  $\omega_m$ , and then successively to use the smaller values.

## PROBLEM, SECTION 2

- Given  $0 < \alpha < \beta$ , show that

$$\min_{0 \leq \omega} \left\{ \max_{\alpha \leq z \leq \beta} \left| \frac{\omega - z}{\omega + z} \right| \right\} = \frac{1 - \sqrt{\alpha/\beta}}{1 + \sqrt{\alpha/\beta}},$$

and that the minimum value is attained for  $\omega = \omega^* \equiv \sqrt{\alpha\beta}$ .

[Hint: The function  $(\omega - z)/(\omega + z)$  is a monotonic function of  $z$  for any fixed  $\omega$ . Hence it attains its extreme values at  $z = \alpha$  and  $z = \beta$ . Equal extreme values are attained for  $\omega = \omega^*$ .]

### 3. WAVE EQUATION AND AN EQUIVALENT SYSTEM

We consider the *initial value* or *Cauchy problem* for the *wave equation*: Find a function  $u(x, t)$  continuous in the half plane

$$D \equiv \{x, t \mid t \geq 0, -\infty < x < \infty\}$$

which satisfies, for  $t > 0$ ,

$$(1) \quad \frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0;$$

and for  $t = 0$ ,

$$(2a) \quad u(x, 0) = f(x),$$

$$(2b) \quad \frac{\partial u(x, 0)}{\partial t} = g(x).$$

This problem may be solved explicitly in terms of quadratures. That is, by using the change of variables

$$\xi = x + ct, \quad \eta = x - ct, \quad \phi(\xi, \eta) \equiv u(x, t),$$

we find

$$\frac{\partial}{\partial x} \equiv \frac{\partial}{\partial \xi} + \frac{\partial}{\partial \eta} \quad \text{and} \quad \frac{\partial}{\partial t} \equiv c \left( \frac{\partial}{\partial \xi} - \frac{\partial}{\partial \eta} \right),$$

whence equation (1) reduces to

$$4c^2 \frac{\partial^2 \phi}{\partial \xi \partial \eta} = 0.$$

The general solution of this equation is found, by two integrations, to be of the form

$$\phi(\xi, \eta) = P(\xi) + Q(\eta).$$

Thus the general solution of (1) is

$$(3) \quad u(x, t) = P(x + ct) + Q(x - ct),$$

where  $P$  and  $Q$  are arbitrary (twice differentiable) functions. Since  $P(x + ct)$  is constant along lines  $x + ct = \text{constant}$ , this part of the solution can be considered as a *signal* or *wave* which propagates to the left with speed  $c > 0$  as time increases. Similarly,  $Q(x - ct)$  represents a wave moving to the right with speed  $c$ . The lines in the  $x, t$ -plane along which the signals travel,

$$x \pm ct = \text{constant},$$

are called the *characteristics* of equation (1).

The initial conditions (2), when applied to (3), yield

$$P(x) + Q(x) = f(x),$$

$$P'(x) - Q'(x) = \frac{1}{c} g(x).$$

Thus by integrating the second relation over  $[0, x]$  and using the first, we may solve the pair of equations for  $P(x)$  and  $Q(x)$ . That is, set  $K = \frac{1}{2}[P(0) - Q(0)]$  and find

$$P(x) = \frac{1}{2}f(x) + \frac{1}{2c} \int_0^x g(\zeta) d\zeta + K,$$

$$Q(x) = \frac{1}{2}f(x) - \frac{1}{2c} \int_0^x g(\zeta) d\zeta - K.$$

If we replace  $x$  in  $P(x)$  by  $x + ct$  and in  $Q(x)$  by  $x - ct$ , we get from (3) the solution of the initial value problem

$$(4) \quad u(x, t) = \frac{1}{2}[f(x + ct) + f(x - ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\zeta) d\zeta.$$

Clearly, the solution at any point  $(x^*, t^*)$  depends upon the initial data only in the interval  $[x^* - ct^*, x^* + ct^*]$  on the initial line,  $t = 0$ . This interval is cut out by the two characteristics passing through  $(x^*, t^*)$  shown in Figure 1. The shaded triangle in this figure is called the *domain of dependence* of the point  $(x^*, t^*)$  and its base is the *interval of dependence*.

The Cauchy problem, (1) and (2), can also be formulated as an initial value problem for a first order system of partial differential equations. In particular, introduce the function  $v(x, t)$  and consider

$$(5) \quad \begin{aligned} \frac{\partial u}{\partial t} &= c \frac{\partial v}{\partial x}, \\ \frac{\partial v}{\partial t} &= c \frac{\partial u}{\partial x}, \end{aligned}$$

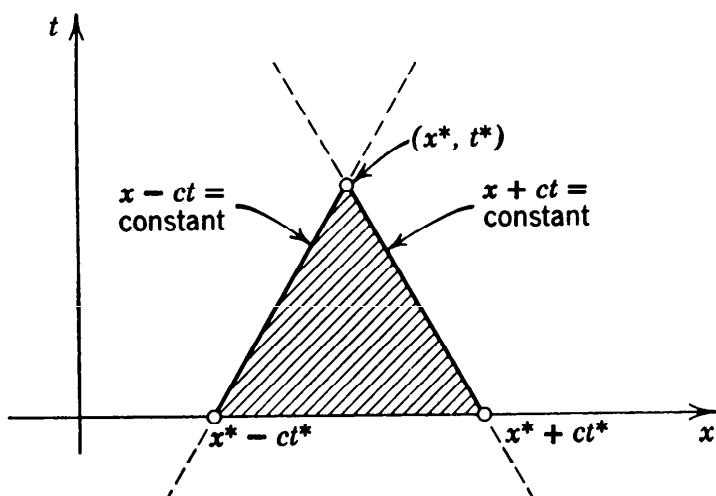


Figure 1. Characteristics and domain of dependence for the wave equation.

subject to the initial values

$$(6a) \quad u(x, 0) = f(x),$$

$$(6b) \quad v(x, 0) = G(x).$$

Equation (1) results after the elimination of  $v(x, t)$  from equations (5). With the same change of variables used before, we find that the general solution of (5) has the form

$$u(x, t) = p(x + ct) + q(x - ct),$$

$$v(x, t) = p(x + ct) - q(x - ct),$$

where  $p$  and  $q$  are again arbitrary functions but only required to possess one derivative. To satisfy the initial conditions (6), we must have

$$p(x) = \frac{1}{2}f(x) + \frac{1}{2}G(x),$$

$$q(x) = \frac{1}{2}f(x) - \frac{1}{2}G(x),$$

and hence the solution of the initial value problem (5) and (6) is

$$(7) \quad \begin{aligned} u(x, t) &= \frac{1}{2}[f(x + ct) + f(x - ct)] + \frac{1}{2}[G(x + ct) - G(x - ct)], \\ v(x, t) &= \frac{1}{2}[f(x + ct) - f(x - ct)] + \frac{1}{2}[G(x + ct) + G(x - ct)]. \end{aligned}$$

A comparison of the solutions  $u$  given in (7) and (4) shows that the two Cauchy problems are equivalent if in (6b) we take

$$G(x) = \frac{1}{c} \int_0^x g(\zeta) d\zeta + \text{constant}.$$

This relation could have been derived directly by satisfying the first equation of (5) at  $t = 0$  and using (2b).

For the system (5), the lines  $x \pm ct = \text{constant}$  are again the characteristics, and the domain of dependence is still as in Figure 1. Of course, as is clear from (7), the solution at any point  $(x^*, t^*)$  is now determined by the values of the initial data (6) at the points where the characteristics through  $(x^*, t^*)$  intersect the initial line,  $t = 0$ . These properties of the system (5) become particularly transparent if we first add and then subtract the equations in this system to get

$$\left( \frac{\partial}{\partial t} - c \frac{\partial}{\partial x} \right)(u + v) = 0,$$

$$\left( \frac{\partial}{\partial t} + c \frac{\partial}{\partial x} \right)(u - v) = 0.$$

This is called the *characteristic form* of the system (5) and the combinations  $u \pm v$  are the *characteristic (dependent) variables*. In the  $(x, t)$ -plane

the operators  $\left(\frac{\partial}{\partial t} \pm c \frac{\partial}{\partial x}\right)$  represent differentiation in two specific directions, the *characteristic directions*, and each characteristic variable is differentiated in an appropriate one of these directions. The notions of characteristic direction, characteristic variable and domain of dependence are useful in the treatment of more general equations of *hyperbolic† type* in two independent variables  $(x, t)$ . For example, consider the simplest case of a linear system of  $n$  first order differential equations for the  $n$  functions  $\{u_i\}$  that are the components of  $\mathbf{u}$ ,

$$\frac{\partial}{\partial t} \mathbf{u} + A \frac{\partial}{\partial x} \mathbf{u} = \mathbf{b}.$$

Suppose that the square matrix  $A = A(x, t)$  has  $n$  real eigenvalues  $(\alpha_i)$  and a complete set of eigenvectors. Let  $P$  be the matrix whose columns are the eigenvectors of  $A$ . Then define  $\mathbf{v}$  by  $\mathbf{u} = P\mathbf{v}$  and insert in the above system. This yields

$$\frac{\partial}{\partial t}(P\mathbf{v}) + A \frac{\partial}{\partial x}(P\mathbf{v}) = \mathbf{b},$$

or, by differentiation

$$P \frac{\partial}{\partial t}(\mathbf{v}) + AP \frac{\partial}{\partial x}(\mathbf{v}) = \mathbf{b} - \left(\frac{\partial}{\partial t}P\right)\mathbf{v} - A\left(\frac{\partial}{\partial x}P\right)\mathbf{v}.$$

If we multiply both sides on the left by  $P^{-1}$ , we find

$$\left(I \frac{\partial}{\partial t} + P^{-1}AP \frac{\partial}{\partial x}\right)\mathbf{v} = P^{-1}\left[\mathbf{b} - \left(\frac{\partial}{\partial t}P\right)\mathbf{v} - A\left(\frac{\partial}{\partial x}P\right)\mathbf{v}\right].$$

This system is in the simple *characteristic form*. That is, differentiation in only a single (*characteristic*) *direction*,

$$\frac{dx}{dt} = \alpha_j,$$

occurs in each equation, since  $P^{-1}AP$  is a diagonal matrix with the  $(\alpha_j)$  on the diagonal. The components of  $\mathbf{v} = P^{-1}\mathbf{u}$  are the *characteristic variables*.

We refrain from giving the definition of characteristic surface, which plays a vital role in the theory of partial differential equations in more dimensions. It is sufficient to say that the notion of domain of dependence

† A system of partial differential equations is said to be of hyperbolic type if the Cauchy initial value problem is well posed for this system. For a linear system of equations, simple algebraic properties of the coefficients have been shown to imply the hyperbolicity of the system, e.g., the conditions on the matrix  $A$  above.

is important for *hyperbolic* equations in higher dimensions but the notion of characteristic form of the system does not generalize.

The *homogeneous characteristic equation* has the form

$$(8a) \quad \frac{\partial w}{\partial t} + \alpha \frac{\partial w}{\partial x} = 0.$$

Now on any curve  $x = x(t)$  in the  $(x, t)$ -plane,  $w$  is a function of  $t$  given by  $w(x(t), t)$  and has the total derivative

$$\frac{dw}{dt} = \frac{\partial w}{\partial t} + \frac{\partial w}{\partial x} \frac{dx}{dt}.$$

Thus if the curve is chosen such that

$$(8b) \quad \frac{dx}{dt} = \alpha,$$

then any solution  $w$  of (8) satisfies  $dw/dt = 0$  and hence, is constant on such a curve. The curves (8b) are the characteristics and if  $\alpha$  is not a constant, they are not straight lines.

The Cauchy problems previously formulated and solved could have been solved by the *method of separation of variables*. Instead, we shall now apply this method to a special *mixed initial-boundary value problem* for the wave equation. The problem of interest is to solve the wave equation

$$\frac{\partial^2 u}{\partial t^2} - c^2 \frac{\partial^2 u}{\partial x^2} = 0$$

subject to the *initial conditions*

$$(9a) \quad u(x, 0) = f(x),$$

$$(9b) \quad \frac{\partial u}{\partial t}(x, 0) = 0, \quad 0 < x < L,$$

and the *boundary conditions*

$$(10a) \quad u(0, t) = 0,$$

$$(10b) \quad u(L, t) = 0, \quad t > 0.$$

The solution is to be determined in the strip  $R \equiv \{x, t \mid 0 \leq x \leq L, t \geq 0\}$ . For convenience we have used a homogeneous initial condition in (9b).

Let us seek solutions of the wave equation in the form

$$u(x, t) = \phi(x)\psi(t)$$

which satisfy the boundary conditions (10). Then we must have

$$\frac{\phi''(x)}{\phi(x)} = \frac{1}{c^2} \frac{\ddot{\psi}(t)}{\psi(t)} = k^2 \equiv \text{constant}.$$

and

$$\phi(0) = \phi(L) = 0,$$

where two primes indicate  $d^2/dx^2$  and two dots indicate  $d^2/dt^2$ . The general solutions of the differential equations which result from this separation are

$$\begin{aligned}\phi(x) &= \alpha e^{kx} + \beta e^{-kx}, \\ \psi(t) &= ae^{ckt} + be^{-ckt}.\end{aligned}$$

From the boundary condition  $\phi(0) = 0$ , we get  $\alpha = -\beta$ ; while  $\phi(L) = 0$  implies that

$$e^{kL} = e^{-kL} \quad \text{or} \quad e^{2kL} = 1.$$

Thus the boundary conditions can only be satisfied if  $k$  has pure imaginary values such that  $2kL = 2n\pi i$ , or

$$k = i \frac{n\pi}{L}, \quad n = 1, 2, \dots$$

We omit  $n = 0$  since it leads to the trivial result  $\phi(x) \equiv 0$ . With  $\alpha = 1$  and any coefficients  $a$  and  $b$ , we have shown that  $\phi(x)\psi(t)$  is a solution of (1) satisfying (10), if

$$\phi(x) = \phi_n(x) = \sin \frac{n\pi x}{L}$$

$$\psi(t) = \psi_n(t) = a_n \sin c \frac{n\pi t}{L} + b_n \cos c \frac{n\pi t}{L}, \quad n = 1, 2, \dots$$

Thus, formally, a solution of the wave equation which satisfies (10) is given by

$$u(x, t) = \sum_{n=1}^{\infty} \phi_n(x)\psi_n(t).$$

To satisfy the initial conditions (9) with this solution, we require that

$$(11a) \quad \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} = f(x), \quad \sum_{n=1}^{\infty} n a_n \sin \frac{n\pi x}{L} = 0.$$

Multiply each of these relations by  $\sin(m\pi x/L)$  and integrate over  $[0, L]$ , if the series converge uniformly, to find, since

$$\int_0^\pi \sin n\theta \sin m\theta d\theta = \frac{\pi}{2} \delta_{mn},$$

$$(11b) \quad b_n = \frac{2}{L} \int_0^L f(x) \sin \frac{n\pi x}{L} dx, \quad a_n = 0, n = 1, 2, \dots$$

The coefficients  $b_n$  are just the Fourier coefficients for the expansion of  $f(x)$  in a sine series. If  $f'(x)$  is piecewise continuous, this series converges

uniformly to  $f(x)$ . The solution of the mixed problem (1), (9), and (10) is given by

$$(11c) \quad \begin{aligned} u(x, t) &= \sum_{n=1}^{\infty} b_n \sin \frac{n\pi x}{L} \cos c \frac{n\pi t}{L} \\ &= \sum_{n=1}^{\infty} \frac{b_n}{2} \left[ \sin \frac{n\pi}{L} (x + ct) + \sin \frac{n\pi}{L} (x - ct) \right]. \end{aligned}$$

If the function  $f(x)$  has a piecewise continuous third derivative, the series in (11) defines a function  $u(x, t)$  with continuous second derivatives which can be evaluated by differentiating (11c) termwise. Hence,  $u(x, t)$  defined by (11) is a solution of the mixed problem. Equation (11c) again shows that  $u(x, t)$  is the sum of functions of the two variables  $x \pm ct$ . In fact, in this special case,

$$u(x, t) = P(x + ct) + P(x - ct),$$

where

$$P(x) \equiv \sum_{n=1}^{\infty} \frac{b_n}{2} \sin \frac{n\pi x}{L} = \frac{f(x)}{2}.$$

### 3.1. Difference Approximations and Domains of Dependence

On the half space  $t \geq 0$ ,  $|x| \leq \infty$ , we introduce the uniformly spaced net points

$$x_j = j\Delta x, \quad t_n = n\Delta t; \quad |j|, n = 0, 1, 2, \dots$$

The set of net points  $D_\Delta$  is defined by

$$D_\Delta \equiv \{x_j, t_n \mid j = 0, \pm 1, \pm 2, \dots; n = 1, 2, \dots\}.$$

A direct approximation of the wave equation (1) is obtained by using centered difference quotients, as in Section 1, to replace derivatives. Thus if  $U(x, t)$  is a net function, we consider the difference equations

$$(12a) \quad U_{tt}(x, t) - c^2 U_{xx}(x, t) = 0, \quad (x, t) \in D_\Delta.$$

If we take the point  $(x, t) = (x_j, t_n)$  and use the subscript notation  $U(x_j, t_n) = U_{j,n}$ , then (12a) can be multiplied by  $\Delta t^2$  and the result re-written as

$$(12b) \quad \begin{aligned} U_{j,n+1} &= 2 \left[ 1 - \left( c \frac{\Delta t}{\Delta x} \right)^2 \right] U_{j,n} \\ &\quad + \left( c \frac{\Delta t}{\Delta x} \right)^2 (U_{j+1,n} + U_{j-1,n}) - U_{j,n-1}, \\ &\quad n \geq 1, |j| = 0, 1, \dots \end{aligned}$$

The star of net points entering into (12) is the same as that in Figure 1 of Section 1 (with  $y$  replaced by  $t$ ). To calculate  $U$  on any time line  $t = t_{n+1}$ , say, the values of  $U$  must be known on the two preceding time lines. Thus in order to start the computations indicated in (12), we require data on the two initial time lines  $t = 0$  and  $t = \Delta t$ . This is consistent with the form of initial data given for the wave equation in (2). A simple adaptation of these conditions is

$$(13a) \quad U(x, 0) = f(x),$$

$$(13b) \quad U_t(x, 0) \equiv U_t(x, \Delta t) = g(x).$$

From (13b), we have

$$(13c) \quad \begin{aligned} U(x, \Delta t) &\equiv U(x, 0) + \Delta t U_t(x, 0) \\ &= f(x) + \Delta t g(x). \end{aligned}$$

More accurate approximations to  $u(x, \Delta t)$  can be obtained if we assume that  $f$  and  $g$  are sufficiently differentiable and that the wave equation (1) is satisfied on the initial line,  $t = 0$ . That is, by Taylor's theorem

$$u(x, \Delta t) = u(x, 0) + \Delta t \frac{\partial u(x, 0)}{\partial t} + \frac{\Delta t^2}{2!} \frac{\partial^2 u(x, 0)}{\partial t^2} + \mathcal{O}(\Delta t^3).$$

But since  $u(x, t)$  satisfies (1) and (2),

$$\frac{\partial^2 u(x, 0)}{\partial t^2} = c^2 \frac{\partial^2 u(x, 0)}{\partial x^2} = c^2 f''(x),$$

hence

$$u(x, \Delta t) = f(x) + \Delta t g(x) + \frac{\Delta t^2}{2} c^2 f''(x) + \mathcal{O}(\Delta t^3).$$

This suggests replacing (13c) by the formula

$$(14a) \quad U(x, \Delta t) = f(x) + \Delta t g(x) + \frac{\Delta t^2}{2} c^2 f_{xx}(x),$$

or equivalently the replacement of (13b) by

$$(14b) \quad U_t(x, 0) \equiv U_t(x, \Delta t) = g(x) + \frac{\Delta t}{2} c^2 f_{xx}(x, 0).$$

Even more accurate approximations than (14a) can be derived by continuing this procedure. For instance, the next term would involve

$$\frac{\partial^3 u(x, 0)}{\partial t^3} = c^2 \frac{\partial^3 u(x, 0)}{\partial x^2 \partial t} = c^2 g''(x).$$

The difference problem posed by (12b) and (13a and c) [or (12b), (13a), and (14a)] is called *explicit* since it is in a form in which the solution is obtained recursively by evaluating the given formulae. (This was not the case for the elliptic difference equations of Section 1 and 2, where a major part of the task was to solve the difference equations efficiently.) A glance at (12), (14), and the star in Figure 1 on p. 447 indicates that the solution at any fixed net point,  $(x^*, t^*)$ , depends only on the values of  $U$  at the net points in the triangle formed by the initial line and the two lines with slopes  $\pm \Delta t / \Delta x$ , say  $x \pm \frac{\Delta x}{\Delta t} t = \text{constant}$ , which pass through  $(x^*, t^*)$ .

This region is shown in Figure 2 and it may be called the *numerical domain of dependence* for the difference equations (12).

Clearly, the numerical domain of dependence will be greater than or equal to the domain of dependence of the wave equation, for the same point  $(x^*, t^*)$ , iff

$$\frac{\Delta t}{\Delta x} \leq \frac{1}{c}.$$

We refer to  $1/c$  as the *characteristic slope* and to  $\Delta t / \Delta x$  as the *net slope*. Therefore, if the characteristic slope is greater than or equal to the net slope, then the numerical domain of dependence includes the domain of dependence of the wave equation. We introduce the ratio of these slopes as

$$(15) \quad \lambda \equiv \frac{\text{net slope}}{\text{characteristic slope}} \equiv \frac{c\Delta t}{\Delta x}$$

and then the above condition becomes  $\lambda \leq 1$ . Note that since  $c$  is the speed of propagation of a signal or wave for the wave equation,  $\lambda$  is the

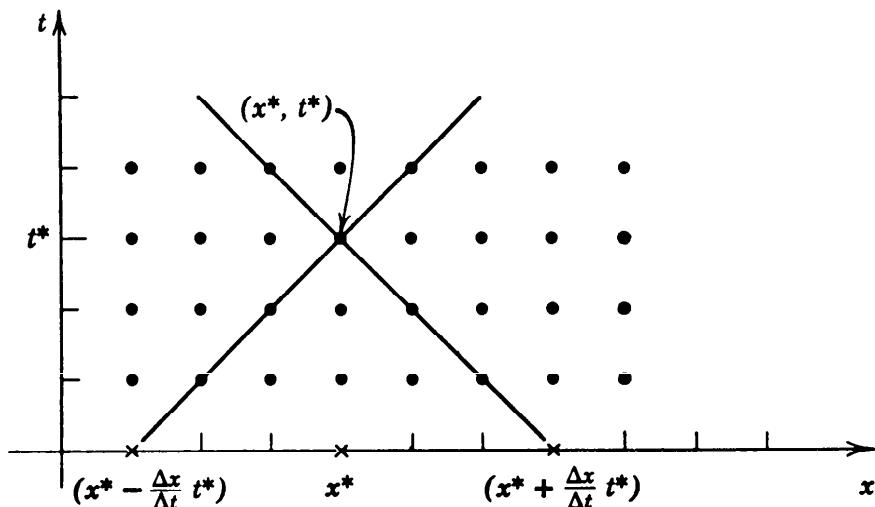


Figure 2. Net points and numerical domain of dependence for difference scheme (12).

ratio of the distance such a signal travels in one time step to the length of a spacial step of the net. Thus if such signals cannot move more than the distance  $\Delta x$  in the time  $\Delta t$ , then the numerical domain contains the analytical domain of dependence.

To understand the significance, for difference schemes, of these domains of dependence, we consider two Cauchy problems for the wave equation. The first is that posed by (1) and (2) with the solution  $u(x, t)$  given in (4). In the second problem we retain (2a) and replace  $g(x)$  in (2b) by

$$g^*(x) = g(x) + \begin{cases} 0 & x \leq a, \\ 4cm(x - a) & x \geq a; \end{cases}$$

where  $x = a$  is an arbitrary fixed point. By using (4) with the new initial data, the solution,  $u^*(x, t)$ , of the altered problem is found to be

$$(16a) \quad u^*(x, t) = u(x, t) + \begin{cases} 0 & x + ct \leq a, \\ \frac{1}{2c} \int_{\max(a, x-ct)}^{x+ct} 4cm(\zeta - a) d\zeta & x + ct \geq a, \end{cases}$$

$$= u(x, t) + \begin{cases} 0 & x + ct \leq a, \\ m(x + ct - a)^2 & x + ct \geq a \geq x - ct, \\ 4cmt(x - a) & x - ct \geq a. \end{cases}$$

Now for each of these problems let us consider the corresponding difference problem (12) and (13) for a net, chosen such that  $x = a$  is a net point on the initial line. If the difference solutions are denoted by  $U$  and  $U^*$  respectively, then, since they have identical initial data on  $x \leq a$ , it follows from a consideration of the numerical domains of dependence that

$$(16b) \quad U^*(x, t) = U(x, t), \quad \text{if } x + \frac{\Delta x}{\Delta t} t \leq a.$$

If the net spacing is such that  $\lambda > 1$ , then there are net points  $(x, t)$  which satisfy

$$x + \frac{\Delta x}{\Delta t} t = a \quad \text{and} \quad x + ct > a \geq x - ct.$$

At such points we have from (16)

$$u^*(x, t) - u(x, t) = m(x + ct - a)^2$$

$$a - ct < x = a - \frac{\Delta x}{\Delta t} t.$$

$$U^*(x, t) - U(x, t) = 0$$

If we let  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$  while  $\lambda = \text{constant} > 1$  and  $x = a$  remains a net point on  $t = 0$ , then clearly  $U(x, t)$  and  $U^*(x, t)$  cannot both converge to the corresponding solutions  $u(x, t)$  and  $u^*(x, t)$ . Thus we deduce

**THEOREM 1.** *In general, the difference solution of (12) and (13) cannot converge to the exact solution of (1) and (2) as  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$  for constant  $\lambda = c\Delta t/\Delta x > 1$ .* ■

The requirement  $\lambda \leq 1$ , which by the above observations is seen to be a necessary condition for convergence in general (i.e., for “all” initial value problems) is called the *Courant-Friedrichs-Lowy condition* (or sometimes just the *Courant condition* for brevity). In other words, *the numerical domain of dependence of a difference scheme should include the domain of dependence of the differential equation or else convergence is not always possible*. We also call this the *domain of dependence condition*.

The relationship between the notion of a domain of dependence and the convergence of a difference method is easily studied for the initial value problem of the single characteristic equation (8a), in which  $\alpha > 0$  is a constant. Then the characteristic curves, determined by (8b), are the lines

$$x = \alpha t + \text{constant}.$$

The solution of (8a) which satisfies the initial condition

$$w(x, 0) = f(x),$$

is thus

$$w(x, t) = f(x - \alpha t).$$

The domain of dependence of the point  $(x^*, t^*)$  is the set of points  $(x, t)$  on the characteristic,  $x = \alpha t + x^* - \alpha t^*$ .

With the uniform net of spacing  $\Delta x$  and  $\Delta t$  we consider the difference equations, for a net function  $W(x, t)$ ,

$$(17a) \quad W_t(x, t) + \alpha W_x(x, t) = 0.$$

In subscript notation, with  $(x, t) = (x_j, t_n)$  and  $\lambda \equiv \alpha\Delta t/\Delta x$ , this becomes

$$W_{j,n+1} = (1 + \lambda)W_{j,n} - \lambda W_{j+1,n}.$$

For this scheme,  $W(x, t + \Delta t)$  is determined by data to the right of  $x$  or directly below it. However, since  $\alpha > 0$ , the exact solution depends upon data to the left of  $x$ . Thus for any  $\lambda > 0$ , the numerical domain of dependence cannot contain that of the differential equation. Clearly then, this procedure is not convergent, in general, as  $\Delta x, \Delta t \rightarrow 0$ .

Now, let us replace the forward  $x$ -difference by a backward difference to get

$$(17b) \quad W_t(x, t) + \alpha W_{\bar{x}}(x, t) = 0,$$

or equivalently

$$W_{j,n+1} = (1 - \lambda)W_{j,n} + \lambda W_{j-1,n}.$$

Clearly, the actual domain of dependence is contained within the numerical domain if  $\lambda \leq 1$ . If the exact solution is sufficiently smooth (i.e., say  $f''(x)$  is continuous), then we find that the *local truncation error*,  $\tau$ , is

$$\tau(x, t) \equiv w_t(x, t) + \alpha w_{\bar{x}}(x, t) = \mathcal{O}(\Delta t + \Delta x).$$

With the definition

$$e(x, t) = W(x, t) - w(x, t)$$

we obtain

$$e_{j,n+1} = (1 - \lambda)e_{j,n} + \lambda e_{j-1,n} - \Delta t \tau_{j,n}.$$

Now let  $E_n \equiv \underset{j}{\text{l.u.b.}} |e_{j,n}|$  and take the absolute value of both sides to get, since  $\lambda \leq 1$ ,

$$\begin{aligned} |e_{j,n+1}| &\leq (1 - \lambda)|e_{j,n}| + \lambda|e_{j-1,n}| + \Delta t |\tau_{j,n}| \\ &\leq (1 - \lambda)E_n + \lambda E_n + \Delta t \mathcal{O}(\Delta t + \Delta x) \\ &\leq E_n + \Delta t \mathcal{O}(\Delta t + \Delta x). \end{aligned}$$

Thus

$$E_{n+1} \leq E_n + \Delta t \mathcal{O}(\Delta t + \Delta x)$$

and a simple recursion yields,

$$E_{n+1} \leq E_0 + t_{n+1} \mathcal{O}(\Delta t + \Delta x),$$

or

$$|W(x, t) - w(x, t)| \leq \|W(x, 0) - f(x)\|_\infty + t \mathcal{O}(\Delta t + \Delta x).$$

Convergence now follows as  $\Delta t$  and  $\Delta x$  vanish while  $\lambda \leq 1$ , provided the initial data  $W(x, 0)$  approaches  $f(x)$ .

The second scheme converges for special choices of the *mesh ratio*,  $\Delta t/\Delta x$ , and is therefore said to be *conditionally convergent*. Of course, the first scheme never satisfies the domain of dependence condition while the second converges when it does satisfy this condition. *However, there are schemes which satisfy the domain of dependence condition, are reasonable approximations to the differential equation but still do not converge for any value of the mesh ratio.* Consider, for example, the scheme

$$(17c) \quad W_t(x, t) + \alpha W_{\hat{x}}(x, t) = 0,$$

which uses the centered  $x$ -difference quotient. The truncation error is now  $\tau = \mathcal{O}(\Delta t + \Delta x^2)$ , and is at least as good as in the previous case. If  $\lambda \leq 1$ , the domain of dependence condition is satisfied but this scheme does not converge, in general, for any mesh ratio (see Problem 1). Thus to determine convergent schemes it is not sufficient to examine domains of dependence and truncation error alone. Now the difference scheme (17c) can be modified in a simple way to yield a convergent scheme. We use

$$\frac{1}{\Delta t} \{W(x, t + \Delta t) - \frac{1}{2}[W(x + \Delta x, t) + W(x - \Delta x, t)]\} + \alpha W_{\hat{x}}(x, t) = 0,$$

in which  $W(x, t)$  in the forward difference  $W_t(x, t)$  has been replaced by an average of two adjacent values. In subscript notation this becomes

$$W_{j,n+1} = \frac{1}{2}(1 - \lambda)W_{j+1,n} + \frac{1}{2}(1 + \lambda)W_{j-1,n},$$

and the truncation error is again  $\mathcal{O}(\Delta t + \Delta x^2)$ . If  $\lambda \leq 1$ , the numerical domain of dependence includes that of the differential equation (8a); the coefficients  $1 \pm \lambda$  are non-negative with sum unity; and convergence can be proved as above [see also the convergence proof in (31)–(35)]. It should be noted that this difference scheme can be written as

$$W_t(x, t) + \alpha W_{\hat{x}}(x, t) - \Delta x \frac{\alpha}{2\lambda} W_{x\bar{x}}(x, t) = 0.$$

### 3.2. Convergence of Difference Solutions

The difference solution determined by (12) and (13) converges to the solution of the initial value problem (1) and (2), provided  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$  while  $\lambda \leq 1$ . The proof of this fact is somewhat complicated for  $\lambda < 1$  but is much simpler if the special mesh ratio condition  $\lambda = 1$  holds. Hence we first consider this case in which the characteristic slope and net slope are equal. It follows that, with the definition

$$D_{i,j} \equiv U_{i,j} - U_{i-1,j-1},$$

the difference equations (12) can be written, when  $\lambda = 1$ , as

$$D_{j,n+1} = D_{j+1,n}.$$

Note that the value  $U_{j,n}$  does not enter into the above difference equation. In fact, the net points may be divided into two groups, corresponding to the red and black squares on a checkerboard, and the difference equations do not couple net points of different groups. Thus we need consider only

one such group of net points. Application of the above form of the difference equation recursively yields

$$\begin{aligned} D_{j,n+1} &= D_{j+1,n} \\ &= D_{j+2,n-1} \\ &\vdots \\ &= D_{j+n,1} \end{aligned}$$

These relations are equivalent to the fact that the  $D_{i,m}$  are constant on the diagonal  $x + ct = \text{constant}$  through  $(x_j, t_{n+1})$ . We also observe by summing the  $D_{i,m}$  along the diagonal  $x - ct = \text{constant}$  through  $(x_j, t_{n+1})$  that

$$U_{j,n+1} - U_{j-n-1,0} = \sum_{v=0}^n D_{j-v,n-v+1}.$$

By combining the last two results and recalling the initial conditions (13a and c), we get

$$\begin{aligned} (18) \quad U_{j,n+1} &= U_{j-n-1,0} + \sum_{v=0}^n D_{j+n-2v,1} \\ &= f_{j-n-1} + \sum_{v=0}^n \Delta t g_{j+n-2v} + \sum_{v=0}^n (f_{j+n-2v} - f_{j+n-2v-1}). \end{aligned}$$

This is an explicit representation of the solution of the difference problem (12)–(13). To examine convergence we shall let  $\Delta x = c\Delta t \rightarrow 0$ . Since  $t_n = n\Delta t$  and  $x_n = n\Delta x = ct_n$ , it follows that  $F_{j+n} \equiv F(x_{j+n}) = F(x_j + ct_n)$  for any function  $F(x)$ . Then if  $f(x)$  has a continuous first derivative

$$\begin{aligned} f_{j+n-2v} - f_{j+n-2v-1} &= f(x_{j-2v} + ct_n) - f(x_{j-2v} + ct_n - \Delta x) \\ &= \Delta x f'(x_{j-2v} + ct_n + \theta_v \Delta x), \quad 0 > \theta_v > -1. \end{aligned}$$

Now take the limit as  $\Delta x \rightarrow 0$  and  $n, j \rightarrow \infty$  in (18), while  $t_{n+1} = t$  and  $x_j = x$ , for any *fixed*  $(x, t)$ , to get

$$\begin{aligned} U(x, t) &= f(x - ct) + \lim_{\Delta x \rightarrow 0} \frac{1}{2c} \sum_{v=0}^n g(x + ct - [2v+1]\Delta x) 2\Delta x \\ &\quad + \lim_{\Delta x \rightarrow 0} \frac{1}{2} \sum_{v=0}^n f'(x + ct + \theta_v \Delta x - [2v+1]\Delta x) 2\Delta x \\ &= f(x - ct) + \frac{1}{2c} \int_0^{2ct} g(x + ct - \xi) d\xi + \frac{1}{2} \int_0^{2ct} f'(x + ct - \xi) d\xi \\ &= \frac{1}{2} [f(x + ct) + f(x - ct)] + \frac{1}{2c} \int_{x-ct}^{x+ct} g(\eta) d\eta \\ &= u(x, t). \end{aligned}$$

Thus the proof of convergence, when  $\lambda = 1$ , has been completed by using the representation of  $u(x, t)$  given in (4).

We study the case of  $\lambda < 1$ , first for the mixed initial-boundary value problem formulated in (1), (9), and (10) whose solution is given in (11). The difference problem can be formulated as the difference equations (12) where now  $D_\Delta$  are the net points in  $0 < x < L, t > 0$ , and the mesh is such that, say,

$$(J + 1)\Delta x = L.$$

The initial conditions (13) or (14) are to be satisfied for  $0 < x < L$ , where  $g(x) \equiv 0$ , and the boundary conditions for the difference problem are

$$(19) \quad U(0, t_n) \equiv U_{0,n} = 0, \quad U(L, t_n) \equiv U_{J+1,n} = 0, \quad n = 1, 2, \dots$$

We now seek solutions of the difference problem by the method of separation of variables. In fact, from the experience gained in Subsection 1.1, we try the forms

$$U(x, t) = \Psi^{(p)}(t) \sin\left(p \frac{\pi x}{L}\right), \quad p = 1, 2, \dots$$

From (12)

$$\begin{aligned} \Psi_{tt}^{(p)} \sin\left(p \frac{\pi x}{L}\right) &= c^2 \Psi^{(p)}(t) \sin_{xx} \left(p \frac{\pi x}{L}\right) \\ &= -\frac{c^2 \zeta_p^2}{\Delta x^2} \Psi^{(p)} \sin\left(p \frac{\pi x}{L}\right) \end{aligned}$$

where  $\zeta_p = 2 \sin\left(\frac{\pi}{2} \frac{p}{J+1}\right)$  and we have employed the trigonometric identities

$$\begin{aligned} \sin\left[p \frac{\pi(x + \Delta x)}{L}\right] + \sin\left[p \frac{\pi(x - \Delta x)}{L}\right] &= 2 \cos\left(p \frac{\pi \Delta x}{L}\right) \sin\left(p \frac{\pi x}{L}\right), \\ 1 - \cos\left(p \frac{\pi \Delta x}{L}\right) &= 2 \sin^2\left(p \frac{\pi \Delta x}{2L}\right). \end{aligned}$$

It follows that  $\Psi^{(p)}(t)$  must satisfy

$$\Delta t^2 \Psi_{tt}^{(p)}(t) = -\lambda^2 \zeta_p^2 \Psi^{(p)}(t),$$

and we try the quantities

$$\Psi^{(p)}(t) = \sin \mu_p t.$$

By using the same trigonometric identities, we get

$$4 \sin^2 \frac{\mu_p \Delta t}{2} \sin \mu_p t = \lambda^2 \zeta_p^2 \sin \mu_p t.$$

A similar result is obtained if we use  $\Psi^{(p)}(t) = \cos \mu_p t$ . Thus the terms

$$(20) \quad (A_p \sin \mu_p t + B_p \cos \mu_p t) \sin \left( p \frac{\pi x}{L} \right), \quad p = 1, 2, \dots,$$

will satisfy the difference equations (12) if  $\mu_p$  is such that

$$4 \sin^2 \frac{\mu_p \Delta t}{2} = \lambda^2 \zeta_p^2,$$

or

$$(21) \quad \sin \mu_p \frac{\Delta t}{2} = \pm \lambda \sin \left( \frac{\pi}{2} \frac{p}{J+1} \right), \quad p = 1, 2, \dots$$

These transcendental equations have real roots,  $\mu_p$ , for all  $p$ , iff  $\lambda = c \Delta t / \Delta x \leq 1$ .

A linear combination of the solutions in (20) yields

$$(22) \quad U(x, t) = \sum_{p=1}^{\infty} (A_p \sin \mu_p t + B_p \cos \mu_p t) \sin \left( p \frac{\pi x}{L} \right).$$

If this series converges, it is a solution of (12) and satisfies the boundary conditions (19). To satisfy the initial condition (13a) we must have

$$(23a) \quad \sum_{p=1}^{\infty} B_p \sin \left( p \frac{\pi x}{L} \right) = f(x);$$

while condition (14a) requires

$$(23b) \quad \begin{aligned} & \sum_{p=1}^{\infty} (A_p \sin \mu_p \Delta t + B_p \cos \mu_p \Delta t) \sin \left( p \frac{\pi x}{L} \right) \\ &= f(x) + \frac{\Delta t^2}{2} c^2 f_{xx}(x) \\ &= (1 - \lambda^2) f(x) + \lambda^2 \frac{f(x + \Delta x) + f(x - \Delta x)}{2}. \end{aligned}$$

From (11a) we see that (23a) is satisfied if  $B_p = b_p$ ,  $p = 1, 2, \dots$ , where the  $b_p$  are defined in (11b). From the identity  $1 - 2 \sin^2(\theta/2) = \cos \theta$  and (21), we have

$$\begin{aligned} \cos(\mu_p \Delta t) \sin \left( p \frac{\pi x}{L} \right) &= (1 - \lambda^2) \sin \left( p \frac{\pi x}{L} \right) \\ &+ \frac{\lambda^2}{2} \left\{ \sin \left[ p \frac{\pi(x + \Delta x)}{L} \right] + \sin \left[ p \frac{\pi(x - \Delta x)}{L} \right] \right\}. \end{aligned}$$

hence (23b) and (23a) yield

$$\sum_{p=1}^{\infty} A_p \sin \mu_p \Delta t \sin \left( p \frac{\pi x}{L} \right) = 0.$$

Clearly, this is satisfied by the choice  $A_p = 0$ . Thus (22) is the solution of the difference problem (12), (13a), (14), and (19), if  $A_p = 0$  and  $B_p = b_p$ . The series (22) converges if (11a) converges absolutely since the  $\mu_p$  are real (e.g., if  $f'(x)$  is continuous).

With the exact solution given by (11c), we obtain

$$(24) \quad |U(x, t) - u(x, t)| \leq \left| \sum_{p=1}^N b_p \left[ \cos \mu_p t - \cos \left( \frac{cp\pi}{L} t \right) \right] \sin \left( p \frac{\pi x}{L} \right) \right| \\ + \left| \sum_{p=N+1}^{\infty} b_p \cos \mu_p t \sin \left( p \frac{\pi x}{L} \right) \right| \\ + \left| \sum_{p=N+1}^{\infty} b_p \cos \left( \frac{cp\pi}{L} t \right) \sin \left( p \frac{\pi x}{L} \right) \right|.$$

By taking  $N$  sufficiently large, the last two sums can be made arbitrarily small for all  $\Delta x, \Delta t$  [since the corresponding series converge absolutely if  $f'(x)$  is continuous]. If  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$  while  $\lambda = c\Delta t/\Delta x \leq 1$  and  $(J + 1)\Delta x = L$ , we have from (21) for  $1 \leq p \leq N$ ,

$$\mu_p \rightarrow \frac{cp\pi}{L}.$$

Thus  $|U(x, t) - u(x, t)|$  can be made arbitrarily small. This proves convergence of the difference scheme, if  $\lambda \leq 1$ , for the mixed initial-boundary value problem, when  $f(x)$  has two continuous derivatives and  $g(x) \equiv 0$ . Convergence for the case  $g(x) \not\equiv 0$  can be shown in a similar way.

Now if  $\lambda \leq 1$ , convergence can be proved for the pure initial value problem, by making use of the notion of domain of dependence. That is, given an interval  $[a, b]$  and time  $T$ , convergence in  $S$ :  $\{(x, t) \mid a \leq x \leq b, 0 \leq t \leq T\}$ , can be shown by modifying the initial data only for  $x < a - (\Delta x/\Delta t)T$ , and  $x > b + (\Delta x/\Delta t)T$ . For this modified problem, the solutions of the differential equation and of the difference equations are unchanged in  $S$ . In fact, if the initial data have been modified so as to be periodic and odd about  $[a - (\Delta x/\Delta t)T - \delta, b + (\Delta x/\Delta t)T + \delta]$ , for some  $\delta > 0$ , then the above proof establishes convergence.

The proof of convergence does not generalize to equations with variable coefficients; furthermore, it does not provide an estimate of the error in terms of the interval size  $\Delta x$  or  $\Delta t$ ; neither does it provide a treatment of the effect of rounding errors. These defects are avoided in the analysis of the next subsection for the case of a first order system of equations.

### 3.3. Difference Methods for a First Order Hyperbolic System

We have seen in equations (5) through (7) that the initial value problem for the wave equation can be replaced by an equivalent first order system.

In fact, such systems arise more naturally in physical theories and applied mathematics than does the second order wave equation. We shall treat the simple system (5) which can be written in vector form as

$$(25a) \quad \frac{\partial \mathbf{u}}{\partial t} = cA \frac{\partial \mathbf{u}}{\partial x},$$

where

$$(25b) \quad \mathbf{u} \equiv \begin{pmatrix} u \\ v \end{pmatrix}, \quad A \equiv \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

Most of our methods and results are also applicable to more general *hyperbolic systems* of first order partial differential equations in two independent variables. [For example, such systems may be formulated as in (25a) with the square matrix  $A$  of order  $n$  having real eigenvalues and simple elementary divisors, i.e.,  $A$  is *diagonalizable*.] The initial conditions to be imposed can be written as

$$(26a) \quad \mathbf{u}(x, 0) = \mathbf{u}_0(x),$$

where for the problem posed in (5) and (6) we take

$$(26b) \quad \mathbf{u}_0(x) \equiv \begin{pmatrix} f(x) \\ G(x) \end{pmatrix}.$$

The solution of (25) subject to (26) is given in (7).

On the uniform net with spacing  $\Delta x$  and  $\Delta t$  we introduce the net functions  $U(x, t)$  and  $V(x, t)$ , or in vector form the vector net function

$$\mathbf{U}(x, t) \equiv \begin{pmatrix} U(x, t) \\ V(x, t) \end{pmatrix}.$$

Then as an approximation to the system (25) with  $\lambda = c\Delta t/\Delta x$ , we consider

$$(27a) \quad \mathbf{U}(x, t + \Delta t) = \frac{1}{2}[\mathbf{U}(x + \Delta x, t) + \mathbf{U}(x - \Delta x, t)] + \frac{\lambda}{2} A[\mathbf{U}(x + \Delta x, t) - \mathbf{U}(x - \Delta x, t)].$$

[It would be tempting to replace the first bracketed term on the right-hand side above by  $\mathbf{U}(x, t)$ , but as shown in Problem 1 that scheme is divergent.] If we subtract  $\mathbf{U}(x, t)$  from each side and divide by  $\Delta t$ , we can write (27) in the difference quotient notation

$$(27b) \quad \mathbf{U}_t(x, t) = cA\mathbf{U}_{\bar{x}}(x, t) + \frac{c}{2\lambda} \Delta x \mathbf{U}_{x\bar{x}}(x, t).$$

Thus our difference equations are obtained by adding a term of order  $\Delta x$  to the divergent approximation of (25). The discussion following equation (17c) may be considered a motivation for (27a).

An immediate advantage of the scheme in (27) can be seen by considering the case  $\lambda = 1$  in which the net diagonals and characteristics have the same slopes and so the numerical and analytical domains of dependence coincide. By writing the system in component form and using  $\Delta x = c\Delta t$ , we get from (27a)

$$(28) \quad \begin{aligned} U(x, t + \Delta t) &= \frac{1}{2}[U(x + c\Delta t, t) + U(x - c\Delta t, t)] \\ &\quad + \frac{1}{2}[V(x + c\Delta t, t) - V(x - c\Delta t, t)] \\ V(x, t + \Delta t) &= \frac{1}{2}[U(x + c\Delta t, t) - U(x - c\Delta t, t)] \\ &\quad + \frac{1}{2}[V(x + c\Delta t, t) + V(x - c\Delta t, t)] \end{aligned}$$

For the initial conditions

$$(29) \quad U(x, 0) = f(x), \quad V(x, 0) = G(x);$$

a comparison of (28) when  $t = 0$  with (7) when  $t = \Delta t$  shows that the numerical solution and the exact solution are *identical* on net points for which  $t = \Delta t$ . By considering the exact solution at this first time step as initial data, we find that the solution is also exact for net points with  $t = 2\Delta t$ . By induction, we can show that *the difference scheme (27) with  $\lambda = 1$  subject to the initial data (29) has a solution which is equal to the exact solution (7) of (25) and (26) at the points of the net.* (However, for higher order systems and variable coefficients, we do not get the exact solution for any fixed choice of  $\lambda$ . These results suggest the use of the largest value of  $\lambda$  for which the domain of dependence condition is satisfied.)

Let us consider the scheme (27) with  $\lambda$  arbitrary. From the considerations of domains of dependence we know that for  $\lambda > 1$ , the difference solution cannot generally converge to the exact solution. Therefore, we restrict the mesh ratio by  $0 < \lambda \leq 1$  and proceed to show that the approximate solution then converges to the exact solution (which is assumed sufficiently smooth). By using the exact solution  $\mathbf{u}(x, t)$  of (25) at the points of the net, we define the local truncation error  $\tau(x, t)$ ,

$$(30) \quad \begin{aligned} \tau(x, t) &\equiv \mathbf{u}_t(x, t) - cA\mathbf{u}_{\bar{x}}(x, t) - \frac{c}{2\lambda} \Delta x \mathbf{u}_{x\bar{x}}(x, t) \\ &= \left( \mathbf{u}_t - \frac{\partial \mathbf{u}}{\partial t} \right) - cA \left( \mathbf{u}_{\bar{x}} - \frac{\partial \mathbf{u}}{\partial x} \right) \\ &\quad - \frac{c}{2\lambda} \Delta x \left( \mathbf{u}_{x\bar{x}} - \frac{\partial^2 \mathbf{u}}{\partial x^2} \right) - \frac{c}{2\lambda} \Delta x \frac{\partial^2 \mathbf{u}}{\partial x^2} \\ &= \mathcal{O}(\Delta t + \Delta x). \end{aligned}$$

If we denote the error in the difference solution by

$$\mathbf{e}(x, t) = \mathbf{U}(x, t) - \mathbf{u}(x, t),$$

then from (27b) and (30) it follows that

$$\mathbf{e}_t(x, t) = cA\mathbf{e}_{\hat{x}}(x, t) + \frac{c}{2\lambda} \Delta x \mathbf{e}_{x\bar{x}}(x, t) - \boldsymbol{\tau}(x, t),$$

or as in (27a) this is equivalent to

$$(31) \quad \begin{aligned} \mathbf{e}(x, t + \Delta t) &= \frac{1}{2}(I + \lambda A)\mathbf{e}(x + \Delta x, t) \\ &\quad + \frac{1}{2}(I - \lambda A)\mathbf{e}(x - \Delta x, t) - \Delta t \boldsymbol{\tau}(x, t). \end{aligned}$$

Here we have introduced the identity matrix  $I$ .

Since  $A$  is symmetric, it can be diagonalized by an orthogonal matrix. We have, in fact,

$$(32) \quad PAP^* = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad P^{-1} = P^* = P \equiv \frac{1}{\sqrt{2}} \begin{pmatrix} 1 & 1 \\ 1 & -1 \end{pmatrix}.$$

Now let us introduce the vector net function

$$(33) \quad \mathbf{\epsilon}(x, t) \equiv \begin{pmatrix} \xi(x, t) \\ \eta(x, t) \end{pmatrix} = P\mathbf{e}(x, t)$$

and multiply (31) by  $P$  on the left to get

$$\begin{aligned} \mathbf{\epsilon}(x, t + \Delta t) &= \frac{1}{2}(I + \lambda PAP^*)\mathbf{\epsilon}(x + \Delta x, t) \\ &\quad + \frac{1}{2}(I - \lambda PAP^*)\mathbf{\epsilon}(x - \Delta x, t) - \Delta t P \boldsymbol{\tau}(x, t). \end{aligned}$$

By taking absolute values and using (32), we find in component form

$$\begin{aligned} |\xi(x, t + \Delta t)| &\leq \frac{1}{2}|1 + \lambda| \cdot |\xi(x + \Delta x, t)| + \frac{1}{2}|1 - \lambda| \cdot |\xi(x - \Delta x, t)| \\ &\quad + \frac{\Delta t}{\sqrt{2}} |\tau_1(x, t) + \tau_2(x, t)| \\ |\eta(x, t + \Delta t)| &\leq \frac{1}{2}|1 - \lambda| \cdot |\eta(x + \Delta x, t)| + \frac{1}{2}|1 + \lambda| \cdot |\eta(x - \Delta x, t)| \\ &\quad + \frac{\Delta t}{\sqrt{2}} |\tau_1(x, t) - \tau_2(x, t)|. \end{aligned}$$

Since  $0 < \lambda \leq 1$ , the absolute value signs can be removed from the factors  $|1 \pm \lambda|$ . Then with the definitions

$$(34) \quad E(t) \equiv \sup_x \|\mathbf{\epsilon}(x, t)\|, \quad \sigma(t) \equiv \sup_x \|\boldsymbol{\tau}(x, t)\|,$$

where the norm of a vector is the maximum absolute component, we deduce

$$E(t + \Delta t) \leq E(t) + \sqrt{2} \Delta t \sigma(t).$$

A recursive application of this inequality yields

$$(35) \quad E(t) \leq E(0) + \sqrt{2} \Delta t \sum_{\nu=1}^{t/\Delta t} \sigma(t - \nu \Delta t) \\ \leq E(0) + \sqrt{2} t \|\sigma(t)\|$$

where

$$\|\sigma(t)\| = \sup_{t' \leq t} \sigma(t') = \sup_{\substack{x \\ t' \leq t}} \|\tau(x, t')\|.$$

Now we recall that, from (32) and (33),  $\mathbf{e}(x, t) = P\mathbf{\epsilon}(x, t)$  and so

$$\begin{aligned} \|\mathbf{e}(x, t)\| &\leq \|P\| \cdot \|\mathbf{\epsilon}(x, t)\| \\ &\leq \sqrt{2} \|\mathbf{\epsilon}(x, t)\| \\ &\leq \sqrt{2} E(t). \end{aligned}$$

Thus it follows from (35) and the definitions of  $\mathbf{e}$  and  $\mathbf{E}$  that

$$(36) \quad \|\mathbf{U}(x, t) - \mathbf{u}(x, t)\| \leq \sqrt{2} \sup_x \|\mathbf{U}(x, 0) - \mathbf{u}(x, 0)\| + 2t \|\sigma(t)\|.$$

Note that the suprema on the right side need be taken only over points in the domain of dependence of the point  $(x, t)$ . By using the initial data (29) and the estimate (30) of the local truncation error, the above implies

$$(37) \quad \|\mathbf{U}(x, t) - \mathbf{u}(x, t)\| \leq t \mathcal{O}(\Delta t + \Delta x).$$

Thus, as was to be shown, *the difference solution of (27) and (29) converges to the exact solution of (25) and (26) as  $\Delta t \rightarrow 0$  and  $\Delta x \rightarrow 0$  for  $\lambda = c\Delta t/\Delta x \leq 1$ .* The convergence here is at least first order in  $\Delta t$  or  $\Delta x$ . In Problem 2, the numerical scheme (27c) is shown to be convergent if the rounding error is of the same order as the truncation error.

We remark that the scheme (27) is convergent for hyperbolic systems of order  $n$  in the form (25a), where  $A$  is diagonalizable (see Problems 3 and 4).

### PROBLEMS, SECTION 3

1. Show that the difference scheme  $W_t = W_{\hat{x}}$ , with constant  $\lambda = \Delta t/\Delta x$ , is divergent as an approximation to  $\partial w/\partial t = \partial w/\partial x$ .

[Hint: Show that  $e^{\beta t} e^{i\alpha x}$  is a solution of the difference equation, if  $i^2 = -1$  and

$$e^{\beta \Delta t} = 1 + i \frac{\Delta t}{\Delta x} \sin \alpha \Delta x.$$

Consider now the initial data

$$W(x, 0) = \sum_{r=0}^{\infty} 2^{-2r} \cos \frac{\pi}{2} 2^r x.$$

Show that  $W(0, t) \rightarrow \infty$  if  $\Delta x = 2^{-n}$  and  $n \rightarrow \infty$ . That is, set  $\alpha_r = \pi 2^{r-1}$  and

$$W(x, t) \equiv \operatorname{Re} \sum_{r=0}^{\infty} 2^{-2r} e^{\beta_r t} e^{i\alpha_r x}.$$

Show that the term  $r = n$  dominates the sum of all of the other terms as  $n \rightarrow \infty$ , in

$$W(0, t) \geq - \sum_{r=n+1}^{\infty} 2^{-2r} + \operatorname{Re} \sum_{r=0}^n 2^{-2r} \left(1 + i\lambda \sin \frac{\pi}{2} 2^{r-n}\right)^{t/\Delta t}.$$

2. Show that the difference scheme

$$(27c) \quad \mathbf{U}_t = cA\mathbf{U}_{\hat{x}} + \frac{c}{2\lambda} \Delta x \mathbf{U}_{x\hat{x}} + \rho(x, t),$$

$$\mathbf{U}(x, 0) = \mathbf{u}(x, 0) + \rho(x)$$

converges with error

$$\|\mathbf{U}(x, t) - \mathbf{u}(x, t)\| \leq t\mathcal{O}(\Delta t + \Delta x),$$

if the rounding errors  $\rho(x, t)$  and  $\rho(x)$  are at most of magnitude  $\mathcal{O}(\Delta t + \Delta x)$ .

3. Carry out the proof of convergence of scheme (27) for the case of a system of  $n$  equations (25a), where  $A$  is a constant matrix having a complete set of eigenvectors.

4. If  $A \equiv A(x, t)$  has a uniformly bounded matrix of real eigenvectors  $P(x, t)$ , with uniformly bounded inverse  $P^{-1}(x, t)$ , then show that (27) is a convergent scheme for (25a).

5. Given the difference scheme  $W_t = W_{\hat{x}}$  (shown in Problem 1 to be divergent if  $\lambda = \Delta t/\Delta x$  is constant), prove convergence if  $\lambda = \mu\Delta x$  with some constant  $\mu$ , for the periodic initial value problems

$$W(x, 0) = f(x) \equiv \sum_{n=-\infty}^{\infty} a_n e^{inx}$$

such that  $\sum_{n=-\infty}^{\infty} n|a_n| < \infty$ .

[Hint: Verify that the function

$$W(x, t) \equiv \sum_{n=-\infty}^{\infty} a_n \left(1 + i \frac{\Delta t}{\Delta x} \sin n\Delta x\right)^{t/\Delta t} e^{inx}$$

is defined by the series, satisfies the difference equation, and converges to  $f(x + t)$  as  $\Delta x \rightarrow 0$ , if  $\Delta t/\Delta x = \mu\Delta x$ . That is, show by using Lemma 0.1' of Chapter 8,

$$\begin{aligned} \left|1 + i \frac{\Delta t}{\Delta x} \sin n\Delta x\right|^{t/\Delta t} &\leq (1 + \mu^2 \Delta x^2 \sin^2 n\Delta x)^{t/(2\Delta t)} \\ &\leq e^{(\mu t/2) \sin^2 n\Delta x} \\ &\leq e^{\mu t/2}. \end{aligned}$$

Such a difference scheme is rather inefficient, since too many time steps are required.

#### 4. HEAT EQUATION

The initial value problem for the heat equation is: Find a continuous function  $u(x, t)$  that satisfies

$$(1a) \quad \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = 0 \quad t > 0;$$

$$(1b) \quad u(x, 0) = f(x) \quad -\infty < x < \infty.$$

The solution of this problem is found to be

$$(2) \quad u(x, t) = \int_{-\infty}^{\infty} \frac{e^{-(\xi-x)^2/4t}}{\sqrt{4\pi t}} f(\xi) d\xi.$$

Here we assume  $f(x)$  to be bounded and continuous, and then direct differentiation under the integral sign shows that (1a) is satisfied. Since

$$\int_{-\infty}^{\infty} e^{-y^2} dy = \sqrt{\pi}$$

we may write (2) as

$$u(x, t) = f(x) + \int_{-\infty}^{\infty} \frac{e^{-(\xi-x)^2/4t}}{\sqrt{4\pi t}} [f(\xi) - f(x)] d\xi,$$

and now let  $t \rightarrow 0$  from above. For all  $\xi \neq x$  we have

$$\lim_{t \rightarrow 0} \frac{e^{-(\xi-x)^2/4t}}{\sqrt{4\pi t}} = 0$$

and for  $\xi = x$ , the remaining factor in the integrand vanishes. Thus it is plausible that we could prove that the function given by (2) is continuous and satisfies the initial condition (1b).

Now from (2), we see that if  $f(x) > 0$  in an open interval  $(a, b)$  and  $f(x) \equiv 0$  outside  $(a, b)$ , then  $u(x, t) > 0$  for all  $x$  when  $t > 0$ . Thus we may say that *signals propagate with infinite speed* for the heat equation. Clearly, the form of the solution in (2) shows that *the domain of dependence* of a point  $(x, t)$  with  $t > 0$  is the entire  $x$ -axis (or initial line).

With the uniform net spacings  $\Delta x$ ,  $\Delta t$ , and net points  $D_\Delta$  in the half-space  $t > 0$  (see Subsection 3.1), we consider the difference equations

$$(3a) \quad U_t(x, t) - U_{xx}(x, t) = 0, \quad (x, t) \in D_\Delta.$$

In subscript notation with  $(x, t) = (x_j, t_n)$ , this can be written in the form

$$(3b) \quad U_{j,n+1} = (1 - 2\lambda)U_{j,n} + \lambda(U_{j+1,n} + U_{j-1,n}), \quad n = 0, 1, \dots$$

Here we have introduced the *mesh ratio*

$$(3c) \quad \lambda \equiv \frac{\Delta t}{\Delta x^2}.$$

The net function  $U(x, t)$  is also subject to initial conditions which, from (1b), we take as

$$(4) \quad U(x, 0) = f(x).$$

The net points used in the difference equation (3) have the *star* or *stencil* of Figure 1. The solution is easily evaluated by means of (3b) and we see that *the numerical interval of dependence* of a point  $(x, t)$  in the net is *the initial line segment*  $[x - (\Delta x/\Delta t)t, x + (\Delta x/\Delta t)t]$ . Thus in order to satisfy the domain of dependence condition of Subsection 3.1, which is again valid, we must have that  $\Delta t/\Delta x \rightarrow 0$  as  $\Delta t \rightarrow 0$  and  $\Delta x \rightarrow 0$ . Otherwise, the numerical interval of dependence of the difference equation (3) would not become arbitrarily large, and hence convergence could not occur in all cases.

If, as the net spacing goes to zero, the mesh ratio  $\lambda$  defined in (3c) is constant, then  $\Delta t/\Delta x = \lambda \Delta x \rightarrow 0$  and the domain of dependence condition is satisfied. We shall show, in fact, that if  $0 < \lambda \leq \frac{1}{2}$ , then the difference scheme (3) and (4) is convergent; but if  $\lambda > \frac{1}{2}$  *the difference solution does not generally converge to the exact solution*. As usual, the truncation error  $\tau(x, t)$  on  $D_\Delta$  is defined by writing for the exact solution  $u(x, t)$  of (1)

$$(5a) \quad u_t(x, t) - u_{xx}(x, t) \equiv \tau(x, t), \quad (x, t) \in D_\Delta.$$

By Taylor's theorem the truncation error can be expressed, assuming  $u$  to be sufficiently smooth, as

$$(5b) \quad \tau(x, t) = \frac{\Delta t}{2} \frac{\partial^2 \bar{u}}{\partial t^2} - \frac{\Delta x^2}{12} \frac{\partial^4 \bar{u}}{\partial x^4}.$$

With the definition

$$e(x, t) = U(x, t) - u(x, t)$$

we get from (5) and (3)

$$e_{j,n+1} = (1 - 2\lambda)e_{j,n} + \lambda(e_{j+1,n} + e_{j-1,n}) - \Delta t \tau_{j,n}.$$

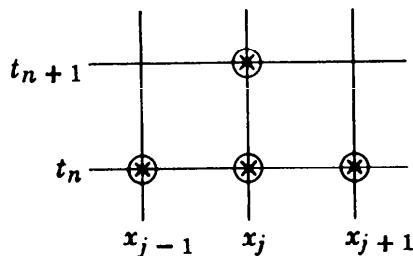


Figure 1. Net points of star for the explicit difference scheme (3).

If  $0 < \lambda \leq \frac{1}{2}$ , then  $(1 - 2\lambda) \geq 0$  and with the definitions

$$E_n \equiv \sup_j |e_{j,n}|, \quad \tau \equiv \sup_{j,n} |\tau_{j,n}|,$$

the above yields upon taking absolute values

$$\begin{aligned} |e_{j,n+1}| &\leq (1 - 2\lambda)|e_{j,n}| + \lambda(|e_{j+1,n}| + |e_{j-1,n}|) + \Delta t |\tau_{j,n}| \\ &\leq E_n + \Delta t \tau. \end{aligned}$$

Or since the right-hand side is now independent of  $j$ ,

$$E_{n+1} \leq E_n + \Delta t \tau.$$

Hence, by a recursive application

$$E_n \leq E_0 + n \Delta t \tau = E_0 + t_n \tau.$$

Thus we have deduced that

$$(6) \quad |u(x, t) - U(x, t)| \leq \sup_x |u(x, 0) - U(x, 0)| + t \mathcal{O}(\Delta t + \Delta x^2).$$

Therefore, by recalling (4),  $|u(x, t) - U(x, t)| \rightarrow 0$  as  $\Delta t \rightarrow 0$  and  $\Delta x \rightarrow 0$ , if  $\lambda = \Delta t / \Delta x^2 \leq \frac{1}{2}$ . The convergence demonstrated here is of order  $\mathcal{O}(\Delta x^2)$  since  $\Delta t = \lambda \Delta x^2$ .

To demonstrate the divergence of the difference scheme (3) when  $\lambda > \frac{1}{2}$ , we first construct explicit solutions of the difference equations. We try net functions of the exponential form

$$V^{(\alpha)}(x, t) = \operatorname{Re}(e^{i\alpha x - \omega t}).$$

Then

$$\begin{aligned} V_t^{(\alpha)} - V_{xx}^{(\alpha)} &= \left[ V^{(\alpha)}(x, t) \left( \frac{e^{-\omega \Delta t} - 1}{\Delta t} - \frac{e^{i\alpha \Delta x} - 2 + e^{-i\alpha \Delta x}}{\Delta x^2} \right) \right] \\ &= \left\{ V^{(\alpha)}(x, t) \frac{1}{\Delta t} \{e^{-\omega \Delta t} - [(1 - 2\lambda) + \lambda e^{i\alpha \Delta x} + \lambda e^{-i\alpha \Delta x}]\} \right\} \\ &= V^{(\alpha)}(x, t) \frac{1}{\Delta t} \left[ e^{-\omega \Delta t} - \left( 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2} \right) \right]. \end{aligned}$$

Now  $V^{(\alpha)}$  is a solution of the difference equations provided that  $\omega$  and  $\alpha$  satisfy

$$e^{-\omega \Delta t} = 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2}.$$

The initial conditions satisfied by  $V^{(\alpha)}$  are

$$(7a) \quad V^{(\alpha)}(x, 0) = \operatorname{Re} e^{i\alpha x} = \cos \alpha x$$

and the solution can be written, since  $e^{-\omega t} = (e^{-\omega \Delta t})^{t/\Delta t}$ ,

$$(7b) \quad V^{(\alpha)}(x, t) = \cos \alpha x \left( 1 - 4\lambda \sin^2 \frac{\alpha \Delta x}{2} \right)^{t/\Delta t}.$$

Clearly, for all  $\Delta x$  and  $\Delta t$  such that  $\lambda \leq \frac{1}{2}$  and real  $\alpha$ , it follows that the solution (7) satisfies  $|V^{(\alpha)}(x, t)| \leq 1$ . However, if  $\lambda > \frac{1}{2}$ , then for some  $\alpha$  and  $\Delta x$  we have  $|1 - 4\lambda \sin^2(\alpha \Delta x/2)| > 1$  and so  $|V^{(\alpha)}(0, t)|$  becomes arbitrarily large for sufficiently large  $t/\Delta t$ . We will capitalize on this *instability* (see next section) of the difference scheme (3), if  $\lambda > \frac{1}{2}$ , to construct a smooth initial condition for which divergence is easily demonstrated. Since the difference equations are linear and homogeneous, we may superpose solutions of the form (7) to get other solutions. With  $\alpha = \alpha_v = 2^v \pi$  and coefficients  $\beta_v > 0$ , we form

$$(8a) \quad V(x, t) = \sum_{v=0}^{\infty} \beta_v V^{(\alpha_v)}(x, t) \\ = \sum_{v=0}^{\infty} \beta_v \cos(2^v \pi x) \left(1 - 4\lambda \sin^2 \frac{2^v \pi \Delta x}{2}\right)^{t/\Delta t}.$$

The corresponding initial function

$$(8b) \quad V(x, 0) = f(x) = \sum_{v=0}^{\infty} \beta_v \cos(2^v \pi x),$$

has as many derivatives as we wish provided that  $\beta_v \rightarrow 0$  sufficiently fast. Now let  $\Delta x = 2^{-m}$  and  $\Delta t = \lambda 4^{-m}$  so that (8a) yields

$$\begin{aligned} V(0, t) &= \sum_{v=0}^{\infty} \beta_v \left[1 - 4\lambda \sin^2 \left(2^{v-m} \frac{\pi}{2}\right)\right]^{t/\Delta t} \\ &= \sum_{v=0}^m \beta_v \left[1 - 4\lambda \sin^2 \left(2^{v-m} \frac{\pi}{2}\right)\right]^{t/\Delta t} + \sum_{v=m+1}^{\infty} \beta_v. \end{aligned}$$

But

$$\sin^2 \left(2^{v-m} \frac{\pi}{2}\right) \leq \frac{1}{2} \quad \text{for } v = 0, 1, \dots, m-1,$$

and so the above yields, for  $\frac{1}{2} < \lambda \leq 1$ , with  $\beta_v > 0$ ,

$$\begin{aligned} |V(0, t)| &\geq - \sum_{v=0}^{\infty} \beta_v + \beta_m (4\lambda - 1)^{t/\Delta t} \\ &= -f(0) + \beta_m (4\lambda - 1)^{t 4^m / \lambda}. \end{aligned}$$

Now if the  $\beta_v$  are chosen as  $\beta_v = e^{-2^v}$ , then the initial function  $f(x)$  is a smooth (analytic) function and the estimate yields, for  $\frac{1}{2} < \lambda \leq 1$ ,

$$(8c) \quad |V(0, t)| \geq -V(0, 0) + e^{2^m [(t/\lambda) 2^m \ln(4\lambda - 1) - 1]}.$$

Thus, as  $m \rightarrow \infty$ , it follows that  $|V(0, t)|$  becomes unbounded, for any finite  $t > 0$ , since  $4\lambda - 1 > 1$ . Hence this difference solution cannot

converge to the solution of the corresponding smooth problem with initial data given by (8b). Thus, as was to be shown, the scheme (3) and (4) does not generally converge when  $\lambda > \frac{1}{2}$ . We say that the difference scheme (3) is *conditionally convergent* which means that the scheme is convergent only if  $\lambda$  satisfies some condition, i.e.,  $\lambda \leq \frac{1}{2}$ . In the next subsection, we will see that it is possible to construct *unconditionally convergent* schemes for the mixed initial-boundary value problem.

#### 4.1. Implicit Methods

To demonstrate implicit difference schemes, we consider mixed initial-boundary value problems for the inhomogeneous heat equation. That is,

$$(9a) \quad \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} = s(x, t) \quad 0 < x < L, \quad t > 0;$$

$$(9b) \quad u(x, 0) = f(x) \quad 0 \leq x \leq L;$$

$$(9c) \quad u(0, t) = g(t), \quad u(L, t) = h(t), \quad t > 0$$

The net spacing is now chosen such that

$$\Delta x = \frac{L}{J + 1}$$

and the net points in the interior of the half strip

$$D \equiv \{x, t \mid 0 \leq x \leq L, t \geq 0\}$$

we denote by  $D_\Delta$ ; i.e.,

$$D_\Delta \equiv \{x, t \mid x = j\Delta x, 1 \leq j \leq J; t = n\Delta t, n = 1, 2, \dots\}.$$

For a net function  $U(x, t)$ , we define the *implicit* difference equations

$$(10a) \quad U_t(x, t) - U_{xx}(x, t) = s(x, t), \quad (x, t) \in D_\Delta.$$

In subscript notation, again with  $(x, t) = (x_j, t_n)$ , these equations can be written as

$$(10b) \quad (1 + 2\lambda)U_{j,n} = U_{j,n-1} + \lambda(U_{j+1,n} + U_{j-1,n}) + \Delta t s_{j,n}; \\ n = 1, 2, \dots, 1 \leq j \leq J.$$

The only difference between (3a) and (10a) is the time difference quotient, which is forward in (3) and backward in (10). The star associated with (10) is shown in Figure 2. The initial and boundary data are specified in the obvious way

$$(11a) \quad U(x_j, 0) \equiv U_{j,0} = f(x_j), \quad 0 \leq x_j \leq L;$$

$$(11b) \quad U(0, t_n) \equiv U_{0,n} = g(t_n), \quad U(L, t_n) \equiv U_{J+1,n} = h(t_n), \quad t_n > 0.$$

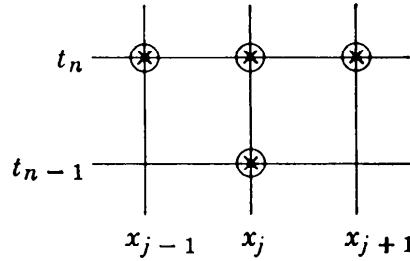


Figure 2. Net points of star for implicit difference scheme (10).

For each  $t = t_n$ , the equations in (10) and (11b) form a system of  $J + 2$  linear equations in the unknowns  $U_{j,n}$ ,  $0 \leq j \leq J + 1$ . However, since  $U_{0,n}$  and  $U_{J+1,n}$  are specified in (11b), it can be reduced to a system of order  $J$ . In fact, with the coefficient matrix  $A$  of order  $J$  defined by

$$(12) \quad A \equiv I + \lambda B \quad B \equiv \begin{pmatrix} 2 & -1 & & & & & 0 \\ -1 & 2 & -1 & & & & \\ & \ddots & \ddots & \ddots & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & \ddots & \ddots & \ddots & -1 \\ 0 & & & & -1 & 2 & \end{pmatrix}$$

and the  $J$ -dimensional vectors  $\mathbf{U}_n$ ,  $\mathbf{b}_n$ ,  $\mathbf{s}_n$ , and  $\mathbf{f}$  defined by

$$(13) \quad \mathbf{U}_n \equiv \begin{pmatrix} U_{1,n} \\ U_{2,n} \\ \vdots \\ U_{J,n} \end{pmatrix}, \quad \mathbf{b}_n \equiv \begin{pmatrix} U_{0,n} \\ 0 \\ \vdots \\ 0 \\ U_{J+1,n} \end{pmatrix} = \begin{pmatrix} g_n \\ 0 \\ \vdots \\ 0 \\ h_n \end{pmatrix},$$

$$\mathbf{s}_n \equiv \begin{pmatrix} s_{1,n} \\ s_{2,n} \\ \vdots \\ s_{J,n} \end{pmatrix}, \quad \mathbf{f} \equiv \begin{pmatrix} f_1 \\ f_2 \\ \vdots \\ f_J \end{pmatrix}$$

the systems (10) and (11) can be written as

$$(14) \quad A\mathbf{U}_n = \mathbf{U}_{n-1} + \lambda\mathbf{b}_n + \Delta t\mathbf{s}_n, \quad n = 1, 2, \dots; \quad \mathbf{U}_0 = \mathbf{f}.$$

For each  $n$ , a system with the same tridiagonal coefficient matrix  $A$  must be solved. Since  $\lambda > 0$ , it follows that the lemma of Subsection 3.2 in Chapter 2 applies. Thus, not only is  $A$  non-singular, but the solution of each system is easily obtained by evaluating two simple two term recursions. Of course, the factorization  $A = LU$  need only be done initially. [See equations (3.10) through (3.13) of Chapter 2.]

It is clear from (14) that the difference solution  $\mathbf{U}_n$  at any time  $t_n$  depends upon all components of the initial data,  $\mathbf{U}_0$ , for any value of  $\lambda$ . This is also clear from the form of the star corresponding to the difference equations (10). Thus *for any value of the mesh slope,  $\Delta t/\Delta x$* , the numerical domain of dependence is the entire initial line segment and hence *the domain of dependence condition is automatically satisfied* by the implicit difference scheme. We shall show that, in addition, *the implicit difference solution converges for all values of  $\lambda$*  to the exact solution. In other words, the scheme is *unconditionally convergent*.

The truncation error  $\tau(x, t)$  of the solution  $u(x, t)$  of (9) is defined for the difference scheme (10) as

$$(15a) \quad \tau(x, t) \equiv u_t(x, t) - u_{x\bar{x}}(x, t) - s(x, t), \quad (x, t) \in D_\Delta.$$

Since  $u(x, t)$  satisfies (9a), we obtain by the usual Taylor's series expansions, assuming sufficient differentiability of the solution,

$$(15b) \quad \tau(x, t) = \mathcal{O}(\Delta t + \Delta x^2).$$

Now from (10), (11), and (15) we get for the error,

$$e(x, t) \equiv U(x, t) - u(x, t),$$

the difference problem

$$(16a) \quad e_t(x, t) - e_{x\bar{x}}(x, t) = -\tau(x, t), \quad (x, t) \in D_\Delta;$$

$$(16b) \quad e(x, 0) = 0;$$

$$(16c) \quad e(0, t) = 0, \quad e(L, t) = 0.$$

In subscript notation (16a) yields

$$(1 + 2\lambda)e_{j,n} = e_{j,n-1} + \lambda(e_{j+1,n} + e_{j-1,n}) - \Delta t \tau_{j,n}, \\ n = 1, 2, \dots, 1 \leq j \leq J.$$

By taking absolute values and using  $E_n \equiv \max_j |e_{j,n}|$ ,  $\tau \equiv \sup_{n' \leq N} |\tau_{j,n'}|$ ,

and  $\lambda > 0$ , we get

$$(1 + 2\lambda)|e_{j,n}| \leq E_{n-1} + 2\lambda E_n + \Delta t \tau, \quad 1 \leq j \leq J, n \leq N.$$

Since the right-hand side is independent of  $j$  and  $e_{0,n} = e_{J+1,n} = 0$ , we may replace  $|e_{j,n}|$  by  $E_n$  to get

$$E_n \leq E_{n-1} + \Delta t \tau.$$

Thus by the usual recursion technique

$$E_n \leq E_0 + t_n \tau,$$

or

$$(17) \quad |u(x, t) - U(x, t)| \leq \max_x |u(x, 0) - U(x, 0)| + t \tau \\ \leq t \mathcal{O}(\Delta t + \Delta x^2), \quad t \leq N\Delta t \equiv T.$$

From this result we deduce unconditional convergence as  $\Delta t \rightarrow 0$  and  $\Delta x \rightarrow 0$ , i.e.,  $\lambda$  is arbitrary.

There are other implicit schemes which converge for arbitrary  $\lambda$  and one of them in particular has a local truncation error which is  $\mathcal{O}(\Delta t^2 + \Delta x^2)$ . We examine the family of schemes defined by

$$(18a) \quad U_t(x, t) = [\theta U_{x\bar{x}}(x, t) + (1 - \theta)U_{\bar{x}\bar{x}}(x, t - \Delta t)] \\ = \theta s(x, t) + (1 - \theta)s(x, t - \Delta t), \quad (x, t) \in D_\Delta.$$

Here  $\theta$  is a real parameter such that  $0 \leq \theta \leq 1$ . For  $\theta = 1$ , (18a) reduces to (10a); while for  $\theta = 0$ , (18a) is equivalent to (3). For any  $\theta \neq 0$ , the difference equations (18) are implicit. The boundary and initial data are as specified in (11). In subscript notation (18a) takes the form

$$(18b) \quad (1 + 2\theta\lambda)U_{j,n} - \theta\lambda(U_{j+1,n} + U_{j-1,n}) \\ = [1 - 2(1 - \theta)\lambda]U_{j,n-1} + (1 - \theta)\lambda(U_{j+1,n-1} + U_{j-1,n-1}) \\ + \Delta t[\theta s_{j,n} + (1 - \theta)s_{j,n-1}], \\ n = 1, 2, \dots, 1 \leq j \leq J.$$

By using the matrices and vectors in (12) and (13), the system (18) and (11) can be written as

$$\mathbf{U}_0 = \mathbf{f},$$

$$(19) \quad (I + \theta\lambda B)\mathbf{U}_n = [I - (1 - \theta)\lambda B]\mathbf{U}_{n-1} + \lambda[\theta \mathbf{b}_n + (1 - \theta)\mathbf{b}_{n-1}] \\ + \Delta t[\theta \mathbf{s}_n + (1 - \theta)\mathbf{s}_{n-1}], \quad n = 1, 2, \dots$$

These systems can be solved by factoring the tridiagonal matrix  $I + \theta\lambda B$ . Clearly, for  $\theta \neq 0$ , the domain of dependence condition is satisfied for arbitrary  $\Delta t/\Delta x$ .

The truncation error  $\tau(x, t)$  now depends upon the parameter  $\theta$ . The

usual Taylor's series expansion about  $(x, t)$  yields, if the solution  $u(x, t)$  of (9) has enough derivatives and we make use of (9a) to simplify,

$$\begin{aligned}
 (20) \quad \tau(x, t) &= u_t(x, t) - [\theta u_{xx}(x, t) + (1 - \theta)u_{xx}(x, t - \Delta t)] \\
 &\quad - [\theta s(x, t) + (1 - \theta)s(x, t - \Delta t)] \\
 &= \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - s(x, t) \\
 &\quad - \frac{\Delta t}{2} \left[ \frac{\partial^2 u}{\partial t^2} - 2(1 - \theta) \left( \frac{\partial^3 u}{\partial t \partial x^2} + \frac{\partial s}{\partial t} \right) \right] \\
 &\quad + \mathcal{O}[(\Delta t)^2 + (\Delta x)^2] \\
 &= \frac{(1 - 2\theta)\Delta t}{2} \frac{\partial^2 u}{\partial t^2} + \mathcal{O}(\Delta t^2 + \Delta x^2).
 \end{aligned}$$

Thus for the special case  $\theta = \frac{1}{2}$ , the truncation error is  $\mathcal{O}(\Delta t^2 + \Delta x^2)$ . [In this case all the difference quotients in (18) are centered about  $(x, t + \Delta t/2)$  and the difference method is called the *Crank-Nicolson* scheme.] For arbitrary  $\theta$ , the truncation error is  $\mathcal{O}(\Delta t + \Delta x^2)$ , as in the explicit and purely implicit cases. With the notation  $e \equiv U - u$  we obtain from (20), (18), (11), and (9b and c)

$$\begin{aligned}
 (21a) \quad e_t(x, t) - [\theta e_{xx}(x, t) + (1 - \theta)e_{xx}(x, t - \Delta t)] \\
 &= -\tau(x, t), \quad (x, t) \in D_\Delta;
 \end{aligned}$$

$$(21b) \quad e(x, 0) = 0;$$

$$(21c) \quad e(0, t) = 0, \quad e(L, t) = 0.$$

Let us write (21) in vector form by using the matrix  $B$  of (12) and the vectors

$$\mathbf{e}_n \equiv \begin{pmatrix} e_{1,n} \\ e_{2,n} \\ \vdots \\ e_{J,n} \end{pmatrix}, \quad \boldsymbol{\tau}_n \equiv \begin{pmatrix} \tau_{1,n} \\ \tau_{2,n} \\ \vdots \\ \tau_{J,n} \end{pmatrix},$$

to get  $\mathbf{e}_0 = 0$  and

$$(22) \quad (I + \theta\lambda B)\mathbf{e}_n = [I - (1 - \theta)\lambda B]\mathbf{e}_{n-1} - \Delta t \boldsymbol{\tau}_n, \quad n = 1, 2, \dots$$

Since  $\lambda > 0$  and  $I + \theta\lambda B$  is non-singular, we may multiply by the inverse of this matrix to get

$$(23a) \quad \mathbf{e}_n = C\mathbf{e}_{n-1} + \Delta t \boldsymbol{\sigma}_n, \quad n = 1, 2, \dots,$$

where we have introduced

$$(23b) \quad C \equiv I - (I + \theta\lambda B)^{-1}\lambda B, \quad \boldsymbol{\sigma}_n \equiv -(I + \theta\lambda B)^{-1}\boldsymbol{\tau}_n.$$

A recursive application of (23a) yields

$$(24) \quad \mathbf{e}_n = C^n \mathbf{e}_0 + \Delta t \sum_{v=1}^n C^{v-1} \boldsymbol{\sigma}_v.$$

Upon taking norms of this representation of the error, we get

$$(25) \quad \begin{aligned} \|\mathbf{e}_n\| &\leq \|C\|^n \cdot \|\mathbf{e}_0\| + \Delta t \sum_{v=1}^n \|C\|^{v-1} \cdot \|\boldsymbol{\sigma}_v\| \\ &\leq \|C\|^n \cdot \|\mathbf{e}_0\| + \Delta t \frac{1 - \|C\|^n}{1 - \|C\|} \cdot \max_{1 \leq v \leq n} \|\boldsymbol{\sigma}_v\|. \end{aligned}$$

[This could have been deduced directly from (23a) by first taking the norm and then applying the recursion.]

Let us use a special norm in (25) for which we are able to compute  $\|C\|$ , i.e.,  $\|\mathbf{x}\| = \left( \sum_{i=1}^n |x_i|^2 \right)^{\frac{1}{2}}$ . Then, since  $B$  is symmetric, it follows that  $C$  in (23b) is symmetric and by (1.11) of Chapter 1,

$$\|C\| = \rho(C) \equiv \max_j |\gamma_j(C)|,$$

where  $\gamma_j(C)$  is an eigenvalue of  $C$ . That is, the spectral radius of  $C$  is the corresponding natural norm. The eigenvalues of  $B$  are easily obtained. We note that the matrix  $B$  is related to the matrix  $H$  in (1.17b). Using  $L_J$  of (1.14b) we have  $B = 2I - (L_J + L_J^T)$  and the calculations in (1.23)–(1.25) are applicable. Specifically the eigenvalues  $\beta_j$  of  $B$  are found to be

$$(26a) \quad \beta_j = 4 \sin^2 \left( \frac{\pi}{2} \frac{j}{J+1} \right), \quad j = 1, 2, \dots, J;$$

corresponding to the eigenvectors

$$(26b) \quad \mathbf{y}^{(j)} = \begin{pmatrix} \sin [1j\pi/(J+1)] \\ \sin [2j\pi/(J+1)] \\ \vdots \\ \sin [Jj\pi/(J+1)] \end{pmatrix}, \quad j = 1, 2, \dots, J.$$

Thus the eigenvalues,  $\gamma_j$ , of  $C$  defined in (23b) are

$$(27) \quad \gamma_j = 1 - \frac{\lambda \beta_j}{1 + \theta \lambda \beta_j}.$$

In order that

$$\rho(C) \equiv \max_j |\gamma_j| < 1,$$

we must have  $-1 < \gamma_j < 1$ . Since  $\beta_j > 0$ , it follows that  $\gamma_j < 1$  for all  $\lambda > 0$  and  $\theta \geq 0$ . Now  $\gamma_j > -1$  is equivalent to

$$(1 - 2\theta)\lambda\beta_j < 2,$$

and this is satisfied for all  $\lambda > 0$  if  $\theta \geq \frac{1}{2}$ . Thus we have shown that

$$(28a) \quad \|C\| < 1 \text{ for all } \lambda > 0 \text{ if } \theta \geq \frac{1}{2}.$$

On the other hand, for  $0 \leq \theta < \frac{1}{2}$ , we must have  $\lambda < 2/[(1 - 2\theta)\beta_j]$ . Or since  $0 < \beta_j < 4$  for  $j = 1, 2, \dots, J$ , this implies

$$(28b) \quad \|C\| < 1 \text{ for } \lambda \leq \frac{1}{2(1 - 2\theta)} \text{ if } 0 \leq \theta < \frac{1}{2}.$$

Under either of the conditions (28), we obtain from (25) and (23b), since

$$\|\sigma_v\| \leq \|(I + \theta\lambda B)^{-1}\| \cdot \|\tau_v\| \leq \frac{1}{1 + \theta\lambda\beta_1} \|\tau_v\| \leq \|\tau_v\|,$$

that

$$(29) \quad \begin{aligned} \|\mathbf{e}_n\| &\leq \|\mathbf{e}_0\| + \frac{\Delta t}{1 - \|C\|} \max_{v \leq n} \|\tau_v\| \\ &= \|\mathbf{e}_0\| + \frac{\Delta t}{1 - \|C\|} \mathcal{O}[(\theta - \frac{1}{2})\Delta t + \Delta t^2 + \Delta x^2]. \end{aligned}$$

To examine the convergence properties as  $\Delta x \rightarrow 0$  and  $\Delta t \rightarrow 0$ , we note that for small  $\Delta x$

$$\beta_1 = \left(\frac{\pi}{L}\Delta x\right)^2 + \mathcal{O}(\Delta x^4), \quad \beta_J = 4 - \left(\frac{\pi}{L}\Delta x\right)^2 + \mathcal{O}(\Delta x^4).$$

It is easily established that  $1 - [x/(1 + \theta x)]$  is a decreasing function of  $x$  and an increasing function of  $\theta$  for  $x > 0$  and  $\theta \geq 0$ . Thus

$$(30) \quad \|C\| = \max(\gamma_1, |\gamma_J|),$$

and for small  $\Delta x$

$$(31) \quad \begin{aligned} \gamma_1 &= 1 - \lambda \left(\frac{\pi}{L}\Delta x\right)^2 + \mathcal{O}(\lambda\Delta x^4) \\ &= 1 - \Delta t \frac{\pi^2}{L^2} + \mathcal{O}(\lambda\Delta x^4). \end{aligned}$$

In case (28b), for any  $\theta$  in  $0 \leq \theta < \frac{1}{2}$  and  $\lambda \leq 1/[2(1 - 2\theta)]$ , we get an upper bound for  $|\gamma_J|$  by picking the largest value of  $\lambda$ ,

$$(32) \quad \begin{aligned} |\gamma_J| &\leq \left| 1 - \frac{\beta_J}{2(1 - 2\theta) + \theta\beta_J} \right| \\ &\leq 1 - (\frac{1}{2} - \theta) \left(\frac{\pi}{L}\Delta x\right)^2 + \mathcal{O}(\Delta x^4). \end{aligned}$$

Hence, in case (28b) we find from (30), (31), and (32) that

$$1 - \|C\| \geq \begin{cases} \mathcal{O}(\Delta t), & \text{or} \\ \mathcal{O}(\Delta x^2). \end{cases}$$

Therefore (29) yields for  $0 \leq \theta < \frac{1}{2}$ ,  $0 < \lambda \leq 1/2(1 - 2\theta)$ ,

$$(33) \quad \|e_n\| \leq \|e_0\| + \max(1, \lambda)\mathcal{O}[(\theta - \frac{1}{2})\Delta t + \Delta t^2 + \Delta x^2].$$

Finally, in case (28a),  $\theta \geq \frac{1}{2}$  and  $\lambda$  arbitrary, we get an upper bound for  $|\gamma_J|$  by picking the smallest value of  $\theta$ ,

$$\begin{aligned} (34) \quad |\gamma_J| &\leq \left| 1 - \frac{\lambda\beta_J}{1 + \frac{1}{2}\lambda\beta_J} \right| \\ &\leq \left| 1 - \frac{4\lambda}{1 + 2\lambda} + \frac{\lambda}{(1 + 2\lambda)^2} \left( \frac{\pi}{L} \Delta x \right)^2 + \mathcal{O}(\lambda\Delta x^4) \right|. \end{aligned}$$

The last inequality is most useful in the case of very large  $\lambda$ , i.e.,

$$\begin{aligned} |\gamma_J| &\leq \frac{4\lambda}{1 + 2\lambda} - 1 - \frac{\lambda}{(1 + 2\lambda)^2} \left( \frac{\pi}{L} \Delta x \right)^2 + \mathcal{O}(\lambda\Delta x^4) \\ &\leq 1 - \frac{1}{\lambda} + o\left(\frac{1}{\lambda}\right), \quad \lambda \gg 1. \end{aligned}$$

Therefore,

$$1 - \|C\| \geq \begin{cases} \mathcal{O}(\Delta t), \\ \mathcal{O}(1/\lambda), \quad \theta \geq \frac{1}{2}. \end{cases}$$

Hence (29) becomes for  $\theta \geq \frac{1}{2}$ ,  $\lambda$  arbitrary,

$$(35) \quad \|e_n\| \leq \|e_0\| + \max(1, \lambda\Delta t)\mathcal{O}[(\theta - \frac{1}{2})\Delta t + \Delta t^2 + \Delta x^2].$$

Inequality (35) indicates that with the choice  $\theta = \frac{1}{2}$ ,  $\lambda\Delta t = \text{constant}$ , the error is bounded by

$$(36) \quad \|e_n\| \leq \|e_0\| + \mathcal{O}(\Delta x^2).$$

Of course, this error bound is of the same magnitude as the error estimate for the explicit scheme, but for a much larger time step. That is, even though the number of operations required to solve the implicit equations for one time step is of the order of twice the number of operations required for one time step of the explicit scheme, there is a tremendous saving in labor when we choose  $\lambda\Delta t = \text{constant}$  for the Crank-Nicolson scheme.

**PROBLEMS, SECTION 4**

**1.** Given  $u(x, t)$  continuous in

$$\bar{D} \equiv D \cup C_1 \cup C_2,$$

where

$$D \equiv \{x, t \mid 0 < x < L, 0 < t < T\},$$

$$C_1 \equiv \{x, t \mid 0 < x < L, t = T\},$$

$$C_2 \equiv \{x, t \mid 0 \leq x \leq L, t = 0\};$$

$$0 = x, \quad 0 < t \leq T;$$

$$x = L, 0 < t \leq T\}.$$

If, in  $D \cup C_1$ ,  $u(x, t)$  has a continuous second derivative with respect to  $x$  and a continuous first derivative with respect to  $t$ , such that

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2},$$

then

$$\max_{x \in \bar{D}} u(x, t)$$

is attained at some point of  $C_2$ . (This is a weak form of the *maximum principle* satisfied by the solutions of the heat equation.)

[Hint: For any  $\epsilon > 0$ , set

$$v(x, t) \equiv u(x, t) - \epsilon t.$$

Clearly,  $v(x, t)$  cannot attain its maximum at any point  $P$  of  $D \cup C_1$ , since otherwise

$$\frac{\partial v(P)}{\partial t} \geq 0,$$

hence

$$\frac{\partial^2 v}{\partial x^2} \equiv \frac{\partial v}{\partial t} + \epsilon,$$

would be positive at  $P$ . Therefore,  $v(x, t)$  attains its maximum *only* on  $C_2$ . But since  $\epsilon > 0$  could be arbitrarily small, this implies that  $u(x, t)$  must attain its maximum on  $C_2$ .]

(Note: By considering

$$W(x, t) \equiv -u(x, t),$$

we can establish the *minimum principle*,

$$\min_{x \in \bar{D}} u(x, t)$$

is attained at some point in  $C_2$ .)

**2.** Verify that the solution  $U_{j,n}$  of (10a and b) and (11a and b) satisfies the *maximum principle*

$$V_n \leq \max \{V_0 + n\Delta t S_n, G_n, H_n\}$$

where

$$V_n \equiv \max_j |U_{j,n}|, \quad n = 0, 1, 2, \dots,$$

$$S_n \equiv \max_{\substack{0 \leq t \leq n\Delta t \\ 0 \leq x \leq L}} |s(x, t)|,$$

$$G_n \equiv \max_{0 \leq t \leq n\Delta t} |g(t)|,$$

$$H_n \equiv \max_{0 \leq t \leq n\Delta t} |h(t)|.$$

[Hint: Modify the argument that estimates  $E_n$ , after equations (16a, b, and c).]

3. Formulate and prove a maximum principle for the explicit scheme

$$(10a') \quad \begin{aligned} U_t(x, t) - U_{xx}(x, t) &= s(x, t), \\ (x, t) \in D_\Delta, \quad &\text{with } \lambda \leq \frac{1}{2}. \end{aligned}$$

Use the auxiliary conditions (11a and b).

4. Prove convergence of the finite difference solutions  $U(x, t)$  to the solution of the mixed initial-boundary value problem in (9a, b, and c) with  $s(x, t) \equiv 0$ , under the *weak compatibility condition*

$$f(0) = g(0); f(L) = h(0),$$

and with  $f(x)$ ,  $g(t)$  and  $h(t)$  that are continuous.

[Hint: Uniformly approximate  $f(x)$ ,  $g(t)$ , and  $h(t)$  by polynomials  $f_m(x)$ ,  $g_m(t)$ , and  $h_m(t)$  satisfying the *strong compatibility condition*,

$$\begin{aligned} f_m''(0) &= g_m'(0), & f_m^{(iv)}(0) &= g_m''(0); \\ f_m''(L) &= h_m'(0), & f_m^{(iv)}(L) &= h_m''(0). \end{aligned}$$

That is, assume there exist corresponding smooth solutions  $u_m(x, t)$  and, for any given  $(\Delta x, \Delta t)$ , difference solutions  $U_m(x, t)$  as well as the continuous solution  $u(x, t)$  and the difference solution  $U(x, t)$ . Then estimate

$$\begin{aligned} u(x, t) - U(x, t) &= [u(x, t) - u_m(x, t)] \\ &\quad + [u_m(x, t) - U_m(x, t)] \\ &\quad + [U_m(x, t) - U(x, t)], \end{aligned}$$

by using the maximum principles for the outermost bracketed terms to fix an  $m$  for which they contribute at most  $\epsilon$  for all  $(\Delta x, \Delta t)$ . Next, pick  $(\Delta x, \Delta t)$  sufficiently small so that the middle bracketed term is at most  $\epsilon$  ]

## 5. GENERAL THEORY: CONSISTENCY, CONVERGENCE, AND STABILITY

The apparently scattered results of the preceding sections can be related by a simple general theory. A more complete and rigorous development

could be given with the aid of the simplest notions of functional analysis, which we forego.

A partial differential equation can be represented symbolically as

$$(1a) \quad L(u) = f(P), \quad P \in D;$$

with the convention that only terms involving the dependent variable  $u$  are included on the left-hand side of (1) and that all inhomogeneous terms are included in  $f$  [i.e.,  $f$  is a function only of the independent variables,  $P \equiv (t, x, y, \dots)$ ]. The domain in which (1a) is to be satisfied is denoted by  $D$ . The set of points on which boundary and/or initial data are prescribed is denoted by  $C$ . The conditions to be satisfied by  $u$  on  $C$  can be represented as

$$(1b) \quad B(u) = g(P), \quad P \in C.$$

Here  $B$  may not be a differential operator, but (1b) merely represents the conditions imposed on various parts of  $C$ . For example, conditions (4.9b) and (4.9c) would both be incorporated in (1b) for the mixed initial-boundary value problem of (4.9). We shall only consider problems (1) for which a unique and smooth solution  $u$  exists for any data in some class of smooth functions  $\{f, g\}$  (smooth means “sufficiently” differentiable).

Let us consider a net for the independent variables of the problem (1) with spacing:  $\Delta t, \Delta x, \Delta y, \dots$ . Certain of these net points, say those *interior* to  $D$  will be denoted by the set  $D_\Delta$ . Similarly, *boundary* net points,  $C_\Delta$ , will also be defined. There are various ways in which this can be done, depending upon the difference method employed. Obviously net points lying on  $C$  may be included in  $C_\Delta$ , but frequently we may also wish to include the points of intersection of  $C$  with the net lines. (In fact, for some problems, net points outside of  $C$  are included in  $C_\Delta$  and points outside  $D$  are included in  $D_\Delta$ , but we shall not dwell on these possibilities in the present discussion.)

At the points of  $D_\Delta + C_\Delta$  a difference approximation  $U$  is defined as the solution of some set of difference equations. These may be indicated symbolically as

$$(2a) \quad L_\Delta(U) = f(P), \quad P \in D_\Delta,$$

which is to approximate (1a); and the boundary difference approximations are indicated by

$$(2b) \quad B_\Delta(U) = g(P), \quad P \in C_\Delta.$$

Again the notation may imply different relations over different parts of  $C_\Delta$ , say as in (4.11). Of course, it is desired that the difference solution  $U$  of (2) should be a close approximation to the solution  $u$  of (1) at corresponding points of  $D_\Delta + C_\Delta$  for all data that are sufficiently smooth.

Furthermore, the difference solution should be uniquely defined by (2) and its numerical evaluation should be possible without significant loss of accuracy due, say, to roundoff errors. To study these questions the three notions of *consistency*, *convergence*, and *stability* of the difference schemes are introduced. It is then easily shown that *for consistent schemes, stability implies convergence*. We begin with the definition of

**CONSISTENCY.** Let  $\phi(t, x, y, \dots)$  be any function with “sufficiently many” continuous partial derivatives in  $D + C$ . For each such function and every point  $P \in D_\Delta$ , let

$$(3a) \quad \tau\{\phi(P)\} \equiv L(\phi(P)) - L_\Delta(\phi(P));$$

and for each point  $P \in C_\Delta$  let

$$(3b) \quad \beta\{\phi(P)\} \equiv B(\phi(P)) - B_\Delta(\phi(P)).$$

Then the difference problem (2) is consistent with problem (1) if

$$(3c) \quad \|\tau\{\phi\}\| \rightarrow 0, \quad \|\beta\{\phi\}\| \rightarrow 0,$$

when  $\Delta t \rightarrow 0$ ,  $\Delta x \rightarrow 0$ ,  $\Delta y \rightarrow 0, \dots$ , in some manner, and  $\|\cdot\|$  represents norms in the appropriate sets  $D_\Delta$  and  $C_\Delta$ . We call  $\tau\{\phi\}$  and  $\beta\{\phi\}$  the local truncation errors.

If (3c) is satisfied only when some particular relationship between the  $\Delta t, \Delta x, \Delta y, \dots$  is maintained (i.e., say provided that  $\Delta t/\Delta x \rightarrow 0$  as  $\Delta t \rightarrow 0$  and  $\Delta x \rightarrow 0$ ), then we say that the difference formulation is *conditionally consistent*. For example, with the heat equation operator:

$$L(u) \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}$$

the ordinary explicit scheme of (4.3a) can be written in terms of the difference operator

$$L_\Delta(U) \equiv U_t - U_{x\bar{x}}$$

which is “unconditionally” consistent with  $L(u)$ . However, the *Dufort-Frankel* explicit scheme employs the difference operator

$$L'_\Delta(U) \equiv \frac{1}{2}(U_t + U_{\bar{t}}) - \left( U_{x\bar{x}} - \frac{\Delta t^2}{\Delta x^2} U_{tt} \right),$$

which is consistent with  $L(u)$  only if  $\Delta t/\Delta x \rightarrow 0$  with the net spacing. In fact, if  $\Delta t/\Delta x \equiv c =$  a fixed constant, then the difference operator  $L'_\Delta(U)$  is consistent with the hyperbolic operator

$$L'(u) \equiv c^2 \frac{\partial^2 u}{\partial t^2} + \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2}.$$

These latter results follow from the simple calculation

$$\begin{aligned}
 \tau\{\phi\} &\equiv L(\phi) - L'_\Delta(\phi) \\
 &= \left[ \frac{\partial\phi}{\partial t} - \frac{1}{2}(\phi_t + \phi_{tt}) \right] - \left( \frac{\partial^2\phi}{\partial x^2} - \phi_{xx} \right) \\
 &\quad - \frac{\Delta t^2}{\Delta x^2} \left( \phi_{tt} - \frac{\partial^2\phi}{\partial t^2} \right) - \frac{\Delta t^2}{\Delta x^2} \frac{\partial^2\phi}{\partial t^2} \\
 &= \left( \frac{\Delta t^2}{6} \frac{\partial^3\phi'}{\partial t^3} \right) + \left( \frac{\Delta x^2}{12} \frac{\partial^4\phi''}{\partial x^4} \right) - \left( \frac{\Delta t^4}{12\Delta x^2} \frac{\partial^4\phi'''}{\partial t^4} \right) - \frac{\Delta t^2}{\Delta x^2} \frac{\partial^2\phi}{\partial t^2}.
 \end{aligned}$$

In the above consideration, we have neglected to mention that initial data must also be prescribed at  $t = \Delta t$  for  $L'_\Delta$  to become an explicit scheme. In many cases, say the mixed problem (4.9) where (1b) represents (4.9b and c) and (2b) represents (4.11a and b), we have  $\beta\{\phi\} \equiv 0$  and so only the difference approximation to the differential equation determines consistency. On the other hand, if (1b) represents initial conditions like (3.2) for the wave equation, then (2b) represents some approximation like (3.13) or alternatively like (3.13a) and (3.14b). In the first case, we obtain  $\beta\{\phi\} = \mathcal{O}(\Delta t)$  and in the second case

$$\beta\{\phi\} = \mathcal{O}\left[\Delta t^2 + \Delta t\Delta x^2 + \Delta t\left(\frac{\partial^2\phi}{\partial t^2} - c^2 \frac{\partial^2\phi}{\partial x^2}\right)\right].$$

Here we have an example in which the order of the local truncation error is increased for special functions (i.e., solutions of the wave equation).

In practice, we will not work with the exact solution of (2a and b), because of rounding operations. Hence we will consider the solutions  $W$  defined on  $D_\Delta + C_\Delta$  which satisfy the modified equations

$$(2c) \quad L_\Delta(W) = f(P) + \rho(P) \quad P \in D_\Delta,$$

$$(2d) \quad B_\Delta(W) = g(P) + \sigma(P) \quad P \in C_\Delta.$$

The functions  $\rho(P)$  and  $\sigma(P)$  represent the error introduced in solving (2a and b) approximately. We will refer to  $\rho(P)$  and  $\sigma(P)$  as *rounding errors*.

We turn now to the definition of

**CONVERGENCE.** *Let  $u$  be the solution of problem (1), and let  $U$  be the difference solution of problem (2). The difference solution is convergent to the exact solution iff*

$$\|u(P) - U(P)\| \rightarrow 0$$

for all  $P \in D_\Delta + C_\Delta$  when  $\Delta t \rightarrow 0$ ,  $\Delta x \rightarrow 0$ ,  $\Delta y \rightarrow 0, \dots$ , in some manner, and  $\|\cdot\|$  represents a norm in  $D_\Delta + C_\Delta$ .

If the difference solution is convergent for all data in some wide class of smooth functions  $\{f, g\}$ , we call the corresponding *difference scheme convergent*. Notice, however, that this notion is quite distinct from that of consistency and, in fact, a scheme may readily be consistent but not convergent. The schemes (3.17a and c) which are consistent approximations to (3.8a) furnish two such examples.

Care must be taken in observing that a difference scheme may be convergent for a class,  $\mathcal{F}$ , of smooth functions  $\{f, g\}$ , but not convergent for a larger class  $\mathcal{G} \supset \mathcal{F}$ . For example, consider the Cauchy problem given by

$$(4a) \quad L(u) \equiv \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} = 0, \quad t > 0,$$

$$B(u) \equiv u(x, 0) = e^{i\alpha x};$$

and the corresponding difference problem

$$(4b) \quad L_\Delta(U) \equiv U_t - U_{\hat{x}} = 0$$

$$B_\Delta(U) \equiv U(x, 0) = e^{i\alpha x}.$$

We easily verify that for any real  $\alpha$ , the solutions  $u$  and  $U$  of (4a) and (4b), respectively, are

$$(5a) \quad u(x, t) = e^{i\alpha(t+x)},$$

and

$$(5b) \quad U(x, t) = \left(1 + i \frac{\Delta t}{\Delta x} \sin \alpha \Delta x\right)^{t/\Delta t} e^{i\alpha x}.$$

But if  $|\alpha| \leq M$ , we have, for  $0 \leq t \leq T$  and all  $x$ , the uniform convergence

$$(6) \quad \lim_{\Delta x, \Delta t \rightarrow 0} U(x, t) = e^{i(\alpha t + \alpha x)} = u(x, t).$$

In other words, the scheme  $L_\Delta(U) = 0$  which we have shown to be divergent, in general, in Problem 3.1, is convergent when the initial data are chosen from the class of finite trigonometric sums, i.e., for any initial data of the form

$$u(x, 0) = \sum_{j=1}^N \beta_j e^{i\alpha_j x}.$$

The reader should observe that even if the domain of dependence condition is violated by the difference scheme (4b) (i.e.,  $\lambda = \Delta t / \Delta x > 1$ ), the solution (5b) will still converge to that in (5a). Thus for this *special class of trigonometric data* the *difference scheme (4b) is unconditionally convergent*.

We recall that the actual evaluation of a numerical scheme produces approximate solutions, say  $W$ , which satisfy slight modifications of (2a and b), say (2c and d). Hence, in practice, we are interested primarily in schemes for which  $\|W - u\| \rightarrow 0$ , when  $\|\rho\| \rightarrow 0$  and  $\|\sigma\| \rightarrow 0$ , as  $\Delta t \rightarrow 0$ ,  $\Delta x \rightarrow 0$ , . . . . (The norms  $\|\cdot\|$  are defined respectively for the sets  $D_\Delta + C_\Delta$ ,  $D_\Delta$  and  $C_\Delta$ , and the data  $\{f, g\}$  are to belong to some wide class of functions.) We say that such schemes have *convergent approximate solutions*.

When a convergent scheme is known, we are assured that difference solutions of arbitrarily good accuracy *exist*. When a scheme with convergent approximate solutions is known we are assured that difference solutions of good accuracy can be *computed!* But in either case, it is important that an a priori estimate of the error can be evaluated, preferably in terms of the data and the mesh spacing. For this and other purposes we introduce the concept of

**STABILITY.** *A difference scheme determined by linear difference operators  $L_\Delta(\cdot)$  and  $B_\Delta(\cdot)$  is stable if there exists a finite positive quantity  $K$ , independent of the net spacing, such that*

$$(7) \quad \|U\| \leq K(\|L_\Delta(U)\| + \|B_\Delta(U)\|)$$

*for all net functions  $U$  defined on  $D_\Delta + C_\Delta$ . (The norms  $\|\cdot\|$  are, as usual, defined for net functions on  $D_\Delta + C_\Delta$ ,  $D_\Delta$  and  $C_\Delta$  respectively.) If (7) is valid for all net spacings, then the linear difference scheme  $\{L_\Delta, B_\Delta\}$  is unconditionally stable; if (7) holds for some restricted family of net spacings in which  $\Delta t, \Delta x, \Delta y, \dots$ , may all be made arbitrarily small, then  $\{L_\Delta, B_\Delta\}$  is conditionally stable.*

Clearly by this definition, stability of a difference scheme is a property independent of any differential equation problem. We have restricted this definition to linear difference schemes as they are the only ones treated in this chapter. However, a more general definition can be given which reduces to the above for linear problems. (This is an obvious restatement of the definition given in Section 5 of Chapter 8 for ordinary differential equations.) Briefly, if  $L_\Delta$  and  $B_\Delta$  are the difference operators in question, *they are stable if for every pair of net functions  $U$  and  $V$  defined on  $D_\Delta + C_\Delta$ , there is a  $K > 0$  independent of the net spacing such that*

$$\|U - V\| \leq K(\|L_\Delta(U) - L_\Delta(V)\| + \|B_\Delta(U) - B_\Delta(V)\|).$$

If  $L_\Delta$  and  $B_\Delta$  are linear, then this reduces to the previous definition applied to  $(U - V)$ .

The factor  $K$  in the definitions of stability may depend upon the dimensions of the domain  $D$  containing  $D_\Delta$ . We have already proved the stability

of various difference schemes in the previous sections. For example, with the Laplace difference approximation (1.4), consider the difference equations (1.5). The corresponding difference operators are  $L_\Delta \equiv -\Delta_\delta$  and  $B_\Delta U \equiv U$ . By applying Theorem 1.2 we deduce that, for arbitrary  $\Delta x$  and  $\Delta y$ ,

$$\begin{aligned}\|U\| &\equiv \max_{D_\delta + C_\delta} |U| \leq K(\max_{C_\delta} |U| + \max_{D_\delta} |\Delta_\delta U|) \\ &= K(\|B_\Delta U\| + \|L_\Delta U\|),\end{aligned}$$

where  $K \equiv \max(1, a^2/2)$ . Thus the difference scheme used in (1.5) is unconditionally stable.

Next, consider the (hyperbolic) system of difference equations defined in (3.27). We define  $L_\Delta$  by

$$L_\Delta(\mathbf{U}(x, t)) \equiv \mathbf{U}_t(x, t) - cA\mathbf{U}_{\bar{x}}(x, t) - \frac{c}{2\lambda} \Delta x \mathbf{U}_{\bar{x}\bar{x}}(x, t),$$

and the initial data are to be given by specifying

$$B_\Delta(\mathbf{U}(x, 0)) \equiv \mathbf{U}(x, 0).$$

(The generalization to vectors  $\mathbf{U}$ ,  $\mathbf{f}$ , and  $\mathbf{g}$  is taken for granted.) By using these definitions in (2) we have the difference problem

$$L_\Delta(\mathbf{U}(x, t)) = \mathbf{f}(x, t), \quad (x, t) \text{ in } D_\Delta; \quad \mathbf{U}(x, 0) = \mathbf{g}(x).$$

However, this is just the problem posed in (3.31) for  $\mathbf{e}(x, t)$  where  $\tau$  replaces  $\mathbf{f}$  and  $\mathbf{e}(x, 0)$  replaces  $\mathbf{g}(x)$ . Thus, as in the derivation of (3.35) and (3.36), we deduce for the above difference problem that if  $\lambda \equiv c\Delta t/\Delta x \leq 1$ , then with the maximum norms over the appropriate sets

$$\|\mathbf{U}(x, t)\| \leq K(\|\mathbf{g}\| + \|\mathbf{f}\|) = K(\|B_\Delta \mathbf{U}\| + \|L_\Delta \mathbf{U}\|),$$

where  $K \equiv \max(\sqrt{2}, 2t)$ . Hence, conditional stability is established (i.e., for  $c\Delta t \leq \Delta x$ ) and we note that the constant  $K$  grows with the time interval included in  $D$ .

Finally, consider the explicit difference equations for the heat equation, which we write as,

$$L_\Delta U(x, t) \equiv U_t(x, t) - U_{\bar{x}\bar{x}}(x, t) = f(x, t).$$

If, initially, we take

$$B_\Delta U(x, 0) \equiv U(x, 0) = g(x),$$

then exactly as in the derivation of (4.6) from (4.3) and (4.4) we get, provided  $\lambda = \Delta t/\Delta x^2 \leq \frac{1}{2}$ ,

$$\|U(x, t)\| \equiv \max_x |U(x, t)| \leq K(\|f(x)\| + \max_{t' \leq t} \|s(x, t')\|),$$

where  $K \equiv \max(1, t)$ . Again we have conditional stability, for  $\Delta t \leq \Delta x^2/2$ , and the constant  $K$  grows with the time. The special choice  $s(x, t) \equiv 0$  and  $f(x) \equiv V(x, 0)$  given by (4.8b), for the above difference problem shows that *this explicit difference scheme is unstable for fixed  $\lambda > \frac{1}{2}$* . The purely implicit difference equations given in (4.10) with initial and boundary data specified as in (4.11) are easily shown, by the methods used in (4.15) through (4.17), to form an unconditionally stable difference scheme.

The basic result connecting the three concepts which we have introduced in this section may be stated as

**THEOREM 1.** *Let  $L_\Delta$  and  $B_\Delta$  be linear difference operators which are stable and consistent with  $L$  and  $B$  on some family of nets in which  $\Delta t, \Delta x, \Delta y, \dots$ , may be made arbitrarily small. Then the difference solution  $U$  of (2) is convergent to the solution  $u$  of (1).*

*Proof.* For each point  $P \in D_\Delta$  and for any of the above family of nets, we obtain by subtracting (1a) from (2a)

$$\begin{aligned} 0 &= L_\Delta(U(P)) - L(u(P)) \\ &= [L_\Delta(U(P)) - L_\Delta(u(P))] + [L_\Delta(u(P)) - L(u(P))]. \end{aligned}$$

From the assumed linearity of  $L_\Delta$  and the definition of the local truncation error,  $\tau\{\phi\}$ , we then have

$$(8a) \quad L_\Delta(U - u) = \tau\{u(P)\}, \quad P \in D_\Delta.$$

In an analogous manner (1b) and (2b) imply

$$(8b) \quad B_\Delta(U - u) = \beta\{u(P)\}, \quad P \in C_\Delta.$$

However, the difference operations in (8) have been assumed stable on the family of nets employed here. Thus it follows that for the net function  $(U - u)$ ,

$$(9) \quad \|U - u\| \leq K(\|\tau\{u\}\| + \|\beta\{u\}\|).$$

Now by the assumed consistency we may let  $\Delta t \rightarrow 0, \Delta x \rightarrow 0, \Delta y \rightarrow 0, \dots$ , in such a manner that  $\|\tau\| \rightarrow 0$  and  $\|\beta\| \rightarrow 0$ . Then, obviously,  $\|U - u\| \rightarrow 0$  and convergence is demonstrated. ■

It should be recalled, in the above proof, that the solution  $u$  of (1) is to have as many continuous derivatives as are required for the derivation of consistency. We then see from (9) that the error in the difference solution is estimated in terms of the local truncation errors. With little change in the proof, Theorem 1 is applicable if  $L_\Delta$  and  $B_\Delta$  are non-linear stable difference operators. (It should also be observed that the

linearity of  $L$  and  $B$  have not been assumed in the above proof. Their linearity follows from the required consistency with linear difference operators.)

### 5.1. Further Consequences of Stability

The stability of a linear difference scheme comes very close to insuring that computations with the proposed scheme are practical. More precisely, what we are assured is that stable linear difference equations *have a unique solution* and that, at least in principle, the growth of roundoff errors is bounded. In this case, the scheme has convergent approximate solutions.

Let the difference problem (2a and b) be linear. Then the difference equations form a linear system whose order is equal to the number of net points in  $D_\Delta + C_\Delta$ . Now we *assume* that the number of unknowns and equations are equal. Having made this important assumption (which in particular cases is easily verified) we may, in order to show that (2a and b) has a solution, either show that some coefficient matrix is non-singular or, equivalently, show that the corresponding homogeneous problem has only the trivial solution. However, from the assumed stability of  $L_\Delta$  and  $B_\Delta$  we get from (2a and b)

$$\|U\| \leq K(\|f\| + \|g\|).$$

It follows that the system has only the trivial solution if  $f \equiv g \equiv 0$ . Thus the unique solvability of the linear difference problem is a simple consequence of stability.

The consideration of the effect of roundoff errors is also quite simple. If by  $W(P)$  we represent the numbers actually obtained in numerically approximating the solution of (2a and b), then

$$(10a) \quad L_\Delta W = f(P) + \rho(P) \quad P \in D_\Delta;$$

$$(10b) \quad B_\Delta W = g(P) + \sigma(P) \quad P \in C_\Delta.$$

Here  $\rho(P)$  and  $\sigma(P)$  represent the effects of rounding, which cause  $W$  to be in error, and hence not quite satisfy the system (2a and b). As in the proof of Theorem 1, we now derive by means of the linearity of  $L_\Delta$  and  $B_\Delta$

$$L_\Delta(W - u) = \tau\{u(P)\} + \rho(P) \quad P \in D_\Delta;$$

$$B_\Delta(W - u) = \beta\{u(P)\} + \sigma(P) \quad P \in C_\Delta.$$

From the assumed stability of the difference problem it now follows that

$$(11) \quad \|W - u\| \leq K(\|\tau\| + \|\beta\| + \|\rho\| + \|\sigma\|).$$

Thus we have shown that *a stable and consistent linear scheme has convergent approximate solutions*.

To maintain accuracy consistent with the local truncation errors, the local roundoffs  $\rho$  and  $\sigma$  should be of the same order in the net spacing. The quantities  $\rho$  and  $\sigma$  introduced in (10) are usually the actual rounding errors committed in computing, divided by some multiple of the net spacing. This is because one does not compute with the equations in the form (2), but rather with multiples of these equations in which the coefficients are bounded as the net spacing vanishes. So the actual rounding errors should be reduced like the truncation error times a power of the net spacing.

The definition of stability that we have given is considerably more restrictive than is required to prove convergence in many cases. In fact, the “stability constant”  $K$  may be allowed to depend upon the net spacing and be unbounded as, say,  $\Delta t \rightarrow 0$ . But if in this case

$$(12) \quad \lim_{\Delta t \rightarrow 0} (\Delta t)^p K = 0$$

for some  $p > 0$ , then convergence still follows if  $\|\tau\| + \|\beta\| = \mathcal{O}(\Delta t^p)$ . Convergent approximate solutions are also obtained if the norms of the rounding errors,  $\|\rho\|$  and  $\|\sigma\|$ , are required to be at least  $\mathcal{O}(\Delta t^p)$ . [When a condition of the form (12) holds for all  $p \geq p_0 > 0$  but not for  $p < p_0$ , the scheme is frequently said to be *weakly stable*.] In addition, as we have seen in the examples, many proofs of stability yield inequalities of the form

$$(13) \quad \|U\| \leq K_0 \|L_\Delta U\| + K_1 \|B_\Delta U\|,$$

which then yield stability with the constant  $K = \max(K_0, K_1)$ . However, if for example  $K_1 \rightarrow 0$  as the net spacing vanishes, then (13) would imply convergence if only  $L_\Delta$  were consistent with  $L$  but not necessarily  $B_\Delta$  with  $B$ . Thus *it is possible to have convergence without consistency* provided a stronger form of stability holds. In the case of such stronger stability it is clear that the error bound (11) can be replaced by

$$(14) \quad \|W - u\| \leq K_0(\|\tau\| + \|\rho\|) + K_1(\|\beta\| + \|\sigma\|).$$

Thus a poorer approximation of the boundary conditions need not affect the overall accuracy if, as the net spacing vanishes,

$$K_1(\|\beta\| + \|\sigma\|) \approx K_0(\|\tau\| + \|\rho\|).$$

## 5.2. The von Neumann Stability Test

There is an important special class of difference schemes for which a simple algebraic criterion always furnishes a necessary and sometimes even a sufficient condition for stability. Simply stated the difference schemes

in question should have constant coefficients and correspond to pure initial value problems with periodic initial data.

We shall illustrate this theory first by considering rather general *explicit* difference schemes of the form

$$(15) \quad \mathbf{U}(t + \Delta t, x) = \sum_{j=-m}^m C_j \mathbf{U}(t, x + j\Delta x); \\ 0 \leq t \leq T - \Delta t, \quad 0 \leq x \leq 2\pi.$$

Here  $\mathbf{U}(t, x)$  is, say, a  $p$ -dimensional vector to be determined on some net with spacing  $\Delta x, \Delta t$  and the  $C_j \equiv C_j(\Delta x, \Delta t)$  are matrices of order  $p$  which are independent of  $x, t$ , but, in general, depend upon  $\Delta x$  and  $\Delta t$ . Initial data are also specified by, say,

$$(16) \quad \mathbf{U}(0, x) = \mathbf{g}(x),$$

where  $\mathbf{g}(x + 2\pi) = \mathbf{g}(x)$ . The difference equations (15) employ data at  $2m + 1$  net points on level  $t$  in order to compute the new vector at one point on level  $t + \Delta t$ . We use the assumed periodicity of  $\mathbf{U}(t, x)$  in order to evaluate (15) for  $x$  near 0 or  $2\pi$ . With no loss in generality then, assume that  $m\Delta x < 2\pi$ , and  $\mathbf{U}(t, x)$  is defined for all  $x$  in  $0 \leq x \leq 2\pi$ , and  $t = 0, \Delta t, 2\Delta t, \dots$  (see Problem 2).

Since  $\mathbf{U}(t, x)$  is to be periodic in  $x$  and the  $C_j$  are constants, we can formally construct Fourier series solutions of (15). That is,  $\mathbf{U}(t, x)$  is of the form

$$(17) \quad \mathbf{U}(t, x) = \sum_{k=-\infty}^{\infty} \mathbf{V}(t, k) e^{ikx}.$$

Upon recalling the orthogonality over  $[0, 2\pi]$  of  $e^{ikx}$  and  $e^{iqx}$  for  $k \neq q$ , we find that this series satisfies (15) iff

$$(18a) \quad \mathbf{V}(t + \Delta t, k) = G(k, \Delta x, \Delta t) \mathbf{V}(t, k)$$

where

$$(18b) \quad G(k, \Delta x, \Delta t) \equiv \sum_{j=-m}^m C_j(\Delta x, \Delta t) e^{ijk\Delta x}, \quad |k| = 0, 1, 2, \dots$$

From (16) and (17), it follows that the  $\mathbf{V}(0, k)$  are just the Fourier coefficients of the initial data,  $\mathbf{g}(x)$ ; i.e.,

$$\mathbf{V}(0, k) = \frac{1}{2\pi} \int_0^{2\pi} \mathbf{g}(x) e^{-ikx} dx.$$

Repeated application of (18a) now yields

$$(19) \quad \mathbf{V}(t, k) = G^n(k, \Delta x, \Delta t) \mathbf{V}(0, k), \quad n = t/\Delta t.$$

The matrices  $G(k, \Delta x, \Delta t)$ , of order  $p$  as defined in (18b) are called the *amplification matrices* of the scheme (15), since they determine the growth of the Fourier coefficients of the solutions of (15).

Because  $\mathbf{U}(t, x)$  is defined for  $0 \leq x \leq 2\pi$ , we may introduce as a norm

$$(20) \quad \|\mathbf{U}(t)\| \equiv \left\{ \frac{1}{2\pi} \int_0^{2\pi} |\mathbf{U}(t, x)|^2 dx \right\}^{1/2}, \quad t = 0, \Delta t, 2\Delta t, \dots,$$

which for each  $t$  is called the  $\mathcal{L}^2$ -norm of  $\mathbf{U}(t, x)$ . (By  $|\mathbf{U}|$ , we denote the Euclidean norm of the vector  $\mathbf{U}$ .) However, by the Parseval equality (see Theorem 5.3 of Chapter 5), or directly by using (17) in (20), we have

$$(21) \quad \|\mathbf{U}(t)\| = \left\{ \sum_{k=-\infty}^{\infty} |\mathbf{V}(t, k)|^2 \right\}^{1/2}.$$

One of the main reasons for using the  $\mathcal{L}^2$ -norm is that it can be simply related, as above, to the sum of the squares of the Fourier coefficients. Then from (19) and (21), we conclude that

$$(22) \quad \begin{aligned} \|\mathbf{U}(t)\| &\leq \max_k \|G^n(k, \Delta x, \Delta t)\| \left\{ \sum_{k=-\infty}^{\infty} |\mathbf{V}(0, k)|^2 \right\}^{1/2} \\ &= (\max_k \|G^n(k, \Delta x, \Delta t)\|) \|\mathbf{U}(0)\|, \quad t = n\Delta t, n = 1, 2, \dots \end{aligned}$$

The matrix norm  $\|G^n\|$  to be used in (22) is, of course, any norm compatible with the Euclidean vector norm  $|\mathbf{V}|$ . As previously observed in Section 1 of Chapter 1 the natural norm induced by this vector norm is the smallest such compatible matrix norm and so we shall employ it here. Thus with no loss in generality let  $\|G^n\|$  be the spectral norm of  $G^n$  [see the definition in (1.11) of Chapter 1 and the discussion preceding Lemma 1.2 thereof].

Let us say, for the present, that *the difference scheme (15) and (16) is stable (in the  $\mathcal{L}^2$ -norm) iff there exists a constant  $K$ , independent of the net spacing, such that*

$$(23) \quad \|\mathbf{U}(t)\| \leq K \|\mathbf{U}(0)\|, \quad 0 \leq t \leq T,$$

*for all solutions  $\mathbf{U}(t, x)$  of (15) and (16) for all  $\mathbf{g}(x)$  with finite  $\mathcal{L}^2$ -norm.* But then we see from (22) that stability is a consequence of the uniform boundedness of the powers of the amplification matrices. To be more precise we introduce the definition: the family of matrices

$$G^n(k, \Delta x, \Delta t) \quad \left\{ \begin{array}{l} m\Delta x \leq 2\pi \\ 0 < n\Delta t \leq T \\ k = 0, \pm 1, \pm 2, \dots \end{array} \right.$$

is *uniformly bounded* if there exists a constant  $K$  independent of  $k$ ,  $\Delta x$  and  $\Delta t$  such that

$$(24) \quad \|G^n(k, \Delta x, \Delta t)\| \leq K \quad \begin{cases} \text{for } m\Delta x \leq 2\pi \\ 0 < n\Delta t \leq T \\ |k| = 0, 1, 2, \dots \end{cases}$$

Now we have the simple

**THEOREM 2.** *The difference equations (15) are stable in the  $\mathcal{L}^2$ -norm iff the amplification matrices (18b) satisfy the uniform boundedness condition (24).*

*Proof.* As indicated above, (24) and (22) imply (23) thus showing the sufficiency. On the other hand, if (24) is not satisfied for any finite  $K$ , then (23) cannot be satisfied for any finite  $K$ , for all  $\mathbf{U}(0, x)$ . That is, given any  $K$ , a single Fourier term of (17) will be a solution of (15), if (18) is satisfied, and can be chosen so that (23) is violated. ■

Having established the importance of uniform boundedness we now state a simple necessary condition for stability due to *von Neumann*.

**THEOREM 3.** *If the scheme (15) is stable (in the  $\mathcal{L}^2$ -norm) then there exists a constant  $M$ , independent of the net spacing, such that*

$$(25) \quad \rho(G(k, \Delta x, \Delta t)) \leq 1 + M\Delta t, \quad \text{for } k = 0, \pm 1, \pm 2, \dots$$

*Proof.* If (15) is stable, then by Theorem 2 the uniform boundedness condition (24) holds. But upon recalling Lemma 1.2 of Chapter 1, we have

$$\begin{aligned} \rho^n(G(k, \Delta x, \Delta t)) &= \rho(G^n(k, \Delta x, \Delta t)) \\ &\leq \|G^n(k, \Delta x, \Delta t)\| \\ &\leq K, \quad |k| = 0, 1, 2, \dots, 0 < n\Delta t \leq T, \quad m\Delta x \leq 2\pi. \end{aligned}$$

With no loss in generality we may take  $K \geq 1$  and thus for  $T = n\Delta t$ ,

$$\rho(G(k, \Delta x, \Delta t)) \leq K^{1/n} = K^{\Delta t/T} \leq 1 + K \frac{\Delta t}{T}, \quad \Delta t \leq T.$$

The last inequality is established in Problem 3 and with  $M \equiv K/T$  the result (25) follows. ■

We call (25) the *von Neumann condition* and Theorem 3 shows that it is necessary for stability in the  $\mathcal{L}^2$ -norm. In some cases it may also be a sufficient condition. For instance, if the amplification matrices  $G(k, \Delta x, \Delta t)$  are Hermitian, then  $\|G^n\| = \|G\|^n = \rho^n(G)$  and Theorem 2 implies that the von Neumann condition is sufficient for stability. In any event, one should

always try to compute the eigenvalues of  $G(k, \Delta x, \Delta t)$  to rule out possibly unstable schemes. If (25) is valid only for special nets, say with  $\Delta x = \phi(\Delta t)$ , then the scheme may only be conditionally stable (see Problems 4 and 5).

Difference schemes of the form (15) can be applied to quite general classes of partial differential equation problems. For example, systems of the form (3.25a) can be treated provided the matrix  $cA$  is constant and a uniformly spaced net is used. More generally, we can treat systems of the form

$$(26) \quad L\mathbf{u}(t, x) \equiv \frac{\partial \mathbf{u}(t, x)}{\partial t} - \mathcal{A}\mathbf{u}(t, x) = 0$$

where  $\mathcal{A}$  is some differential operator with respect to the  $x$  variable and has constant coefficients. Thus the heat equation (4.1a) and other higher order systems are included.

The previous analysis is easily extended to more general difference schemes. The most obvious such generalization is to implicit schemes which may be written as

$$(27) \quad \sum_{|j| \leq m} B_j \mathbf{U}(t + \Delta t, x + j\Delta x) = \sum_{|j| \leq m} C_j \mathbf{U}(t, x + j\Delta x), \\ 0 \leq t \leq T - \Delta t.$$

Here the  $B_j$  are matrices of order  $p$ , independent of  $x$  and  $t$ , but dependent, in general, on  $\Delta x$  and  $\Delta t$ . We now need only change the definition of the amplification matrices from (18b) to

$$(28) \quad G(k, \Delta x, \Delta t) \equiv \left( \sum_{|j| \leq m} B_j e^{ijk\Delta x} \right)^{-1} \left( \sum_{|j| \leq m} C_j e^{ijk\Delta x} \right), \\ |k| = 0, 1, 2, \dots$$

and of course require that the indicated inverse exists (see Problem 4). We can treat difference schemes with more than two time levels but then the amplification matrices are of higher order, say of order  $pq$  for  $q + 1$  time levels. Extensions to more independent variables are not difficult.

We recall that the stability definition (23) which has been used in this subsection is not the same as that in (7), which is employed in the basic Theorem 1. However, for the class of equations, say of the form (26), for which difference schemes of the form (15) or (27) are appropriate, we can show that (23) [or (24)] is equivalent to (7). Specifically, we have

**THEOREM 4.** *Let  $B_\Delta \equiv I$ , the identity operator, and let  $L_\Delta$  be defined by*

$$(29) \quad \Delta t L_\Delta \mathbf{U}(t, x) \equiv \mathbf{U}(t + \Delta t, x) - \sum_{|j| \leq m} C_j \mathbf{U}(t, x + j\Delta x).$$

*Then for difference problems of the form (15) and (16) the definition (23) of  $\mathcal{L}^2$ -stability is equivalent to that in (7) with appropriate norms.*

*Proof.* The appropriate norms to employ in (7) are, in terms of the  $\mathcal{L}^2$ -norm (20),

$$\|\mathbf{U}\| \equiv \sup_{0 \leq t_n \leq T} \|\mathbf{U}(t_n)\|, \quad \|L_\Delta \mathbf{U}\| \equiv \sup_{0 < t_n \leq T} \|L_\Delta \mathbf{U}(t_n)\|,$$

$$\|B_\Delta \mathbf{U}\| \equiv \|\mathbf{U}(0)\|.$$

Then trivially it follows that (7) implies (23) for all solutions  $\mathbf{U}$  of (15) and (16); that is, for all  $\mathbf{U}(t, x)$  satisfying

$$L_\Delta \mathbf{U} = 0 \text{ on } D_\Delta, \quad \text{i.e., } 0 < t \leq T;$$

$$B_\Delta \mathbf{U} \equiv \mathbf{U}(0, x) = \mathbf{g}(x) \text{ on } C_\Delta, \quad \text{i.e., } t = 0.$$

Now let  $\mathbf{U}(t, x)$  be *any* function with convergent Fourier series (17), *not necessarily a solution of (15)*. Insert the series on the right-hand side of (29), multiply the result by  $(1/2\pi)e^{-ikx}$ , and integrate over  $[0, 2\pi]$ . We obtain with the definition (18b),

$$\mathbf{V}(t + \Delta t, k) = G(k, \Delta x, \Delta t) \mathbf{V}(t, k) + \frac{\Delta t}{2\pi} \int_0^{2\pi} e^{-ikx} L_\Delta \mathbf{U}(t, x) dx.$$

By applying this result recursively in  $t$ , with the notation  $G \equiv G(k, \Delta x, \Delta t)$  and  $n\Delta t = t$ , it follows that

$$\mathbf{V}(t, k) = G^n \mathbf{V}(0, k) + \frac{\Delta t}{2\pi} \sum_{v=0}^{n-1} G^v \int_0^{2\pi} e^{-ikx} L_\Delta \mathbf{U}(t - v\Delta t, x) dx,$$

$$0 < n\Delta t \leq T, |k| = 0, 1, 2, \dots$$

Upon taking the Euclidean norm of this vector equation, we have

$$|\mathbf{V}(t, k)| \leq \max_{v \leq n} \|G^v(k, \Delta x, \Delta t)\|$$

$$\times \left\{ |\mathbf{V}(0, k)| + \Delta t \sum_{v=0}^{n-1} \left| \frac{1}{2\pi} \int_0^{2\pi} e^{-ikx} L_\Delta \mathbf{U}(t - v\Delta t, x) dx \right| \right\}.$$

Finally, by using this result in (21) (see Problem 6), we get with the aid of the Schwarz inequality, the inequalities

$$2ab \leq a^2 + b^2 \quad \text{and} \quad a + b \geq \sqrt{a^2 + b^2},$$

the estimate

$$(30) \quad \|\mathbf{U}(t)\| \leq \sqrt{2} \sup_{\substack{v \leq n \\ |k| < \infty}} \|G^v(k, \Delta x, \Delta t)\| \{ \|\mathbf{U}(0)\| + t \max_{\tau \leq t} \|L_\Delta \mathbf{U}(\tau)\| \},$$

$$\text{for } n\Delta t \leq T, |k| = 0, 1, 2, \dots$$

Here we have introduced the  $\mathcal{L}^2$ -norm of  $L_\Delta \mathbf{U}(t, x)$ , as in (20), and used the Parseval equality to deduce that

$$\|L_\Delta \mathbf{U}(\tau)\|^2 = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} \left| \int_0^{2\pi} e^{-ikx} L_\Delta \mathbf{U}(\tau, x) dx \right|^2.$$

Now let the scheme (15) be stable in the sense of (23). Then by Theorem 2 the amplification matrices  $G^v(k, \Delta x, \Delta t)$  are uniformly bounded, as in (24), and we find from (30) that

$$\|\mathbf{U}\| \leq K' \{\|\mathbf{U}(0)\| + \|L_\Delta \mathbf{U}\|\}$$

where  $K' \equiv \sqrt{2} K \max(1, T)$ . But this holds for all (sufficiently smooth) functions  $\mathbf{U}(t, x)$  in  $0 \leq x \leq 2\pi$ ,  $0 \leq t \leq T$ , and so the stability defined in (7) holds for the difference operator  $L_\Delta$  of (29). ■

We conclude with the remark that the condition of stability thus far has only been shown to be sufficient for convergence. But, by using elementary ideas of functional analysis, Lax has proved that for linear well-posed initial value problems, the *consistent difference schemes* (15) or (27) are stable iff the schemes are convergent!

## PROBLEMS, SECTION 5

**1.\*** With the notation  $\sum_{j=-n}^n'$  and  $x_k$  of Subsection 5.1 of Chapter 5, set  $h = \Delta x$  and verify that the  $2\pi$  periodic solution  $W(x, t)$  of  $W_t = W_x$ , with  $W(x_k, 0) = f(x_k)$  for  $k = 0, \pm 1, \dots, \pm n$ , satisfies

$$\begin{aligned} \frac{1}{2n} \sum_{k=-n}^n' |W(x_k, t)|^2 &= \sum_{j=-n}^n' \left| b_j \left( 1 + i \frac{\Delta t}{\Delta x} \sin j\Delta x \right)^{t/\Delta t} \right|^2 \\ &\leq e^{\mu t} \sum_{j=-n}^n' |b_j|^2 = \frac{e^{\mu t}}{2n} \sum_{k=-n}^n' |W(x_k, 0)|^2, \end{aligned}$$

where

$$b_j = \frac{\Delta x}{2\pi} \sum_{k=-n}^n' f(x_k) e^{-ijx_k} \quad \text{and} \quad \Delta t = \mu \Delta x^2.$$

That is, if we define  $\|\cdot\|$  by

$$\|W(t)\|^2 = \frac{1}{2n} \sum_{k=-n}^n' |W(x_k, t)|^2,$$

we have shown that

$$\|W(t)\| \leq e^{\mu t/2} \|W(0)\|$$

and the difference scheme is stable for  $\Delta t = \mu \Delta x^2$  for any constant  $\mu$  (see Problem 3.5).

**2.** Explain how if  $\mathbf{U}(0, x)$  is given, (15) may be used to define  $\mathbf{U}(\Delta t, x)$ , for  $0 \leq x \leq 2\pi$ —even though (15) is a difference equation.

**3.** Verify the inequality, with  $K \geq 1$ ,

$$K^x \leq 1 + Kx, \quad \text{for } 0 \leq x \leq 1.$$

[Hint: Study  $f(x) \equiv e^{x \ln K}$ , show that  $f'(x) > 0$ ,  $f''(x) > 0$ .]

**4.** (a) Show that the amplification matrices for the Crank-Nicolson like scheme (4.18) are the scalars

$$G(k, \Delta x, \Delta t) = \frac{1 - (1 - \theta)2\lambda(1 - \cos k\Delta x)}{1 + \theta 2\lambda(1 - \cos k\Delta x)}, \quad \lambda = \frac{\Delta t}{\Delta x^2},$$

(b) Since this is a case where  $G$  is Hermitian, determine by means of the von Neumann condition the restrictions on  $\lambda$  for stability for any  $\theta$  in  $0 \leq \theta \leq 1$ . Compare your results with (4.28).

**5.** (a) Find the amplification matrices of the scheme (3.27). Verify that the von Neumann condition is satisfied when the Courant condition,  $\lambda \equiv c\Delta t/\Delta x < 1$ , is satisfied.

(b) Apply the von Neumann test to the divergent scheme

$$\mathbf{U}(x, t + \Delta t) = \mathbf{U}(x, t) + \frac{\lambda}{2} A[\mathbf{U}(x + \Delta x, t) - \mathbf{U}(x - \Delta x, t)],$$

$$A = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}.$$

**6.** If  $a(k, n) = b(k) + c \sum_{\nu=0}^{n-1} d(k, \nu)$ , for real numbers  $a, b, c$ , and  $d$ , show that

$$[a(k, n)]^2 \leq 2 \left\{ [b(k)]^2 + c^2 \left[ \sum_{\nu=0}^{n-1} d(k, \nu) \right]^2 \right\}.$$

Hence, show that

$$[a(k, n)]^2 \leq 2 \left\{ [b(k)]^2 + nc^2 \sum_{\nu=0}^{n-1} [d(k, \nu)]^2 \right\}.$$

[Hint: Apply Schwarz' inequality:  $(\sum 1 \cdot d)^2 \leq (\sum 1)(\sum d^2)$ .]

# Bibliography

## GENERAL TEXTS

- Hildebrand, F. B., *Introduction to Numerical Analysis*, McGraw-Hill Book Co., New York, 1956.
- Householder, Alston S., *Principles of Numerical Analysis*, McGraw-Hill Book Co., New York, 1953.
- John, F., *Advanced Numerical Methods* (lecture notes), New York University, Courant Institute of Mathematical Sciences, New York, 1956. Revised as *Lectures on Numerical Analysis*, Gordon and Breach, New York, to be published.
- Lanczos, Cornelius, *Applied Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1956.
- Milne, W. E., *Numerical Calculus*, Princeton University Press, Princeton, N.J., 1949.
- Todd, John (editor), *Survey of Numerical Analysis*, McGraw-Hill Book Co., New York, 1962.

## SPECIAL REFERENCES FOR CHAPTERS 1, 2, 3, 4

- Faddeev, D. K., and V. N. Faddeeva, *Computational Methods of Linear Algebra* (translated by R. C. Williams), W. H. Freeman and Co., San Francisco, 1963.
- Forsythe, G. E., and P. Henrici, "The cyclic Jacobi method for computing the principal vectors of a complex matrix," *Trans. Amer. Math. Soc.*, **94**, 100–115 (1958).
- Francis, J. G. F., "The *QR* transformation—a unitary analogue to the *LR* transformation," *Computer Journal*, **4**, 265–271, 332–345 (1961/62).
- Givens, W., "A method of computing eigenvalues and eigenvectors suggested by classical results on symmetric matrices," *Appl. Math. Ser. Natl. Bur. Stand.*, **29**, 117–122 (1953).
- , "Numerical computation of the characteristic values of a real symmetric matrix," *Oak Ridge Natl. Lab. Report ORNL 1574*, 1954.
- Householder, Alston, *The Theory of Matrices in Numerical Analysis*, Blaisdell Publishing Co., New York, 1964.
- Ostrowski, A. M., *Solution of Equations and Systems of Equations*, Academic Press, New York, 1960.

- Ostrowski, A. M., "On the convergence of the Rayleigh quotient iteration for the computation of the characteristic roots and vectors, I VI," *Arch. Rational Mech. Anal.*, **1**, 233-241 (1958); **2**, 423-428 (1959), **3**, 325-340, 341-367, 472-481 (1959); **4**, 153-165 (1960).
- Parlett, B., "The development and use of methods of *LR* type," *SIAM Review*, **6**, 275-295 (1964).
- Rutishauser, H., "Solution of eigenvalue problems with the *LR*-transformation," *Appl. Math. Ser. Natl. Bur. Stand.*, **49**, 47-81 (1958).
- Wilkinson, J. H., *Rounding Errors in Algebraic Processes*, Prentice-Hall, Englewood Cliffs, N.J., 1963.
- , *The Algebraic Eigenvalue Problem*, Oxford University Press, London, 1965.

### SPECIAL REFERENCES FOR CHAPTERS 5, 6, 7

- Achieser, N. I., *Theory of Approximation* (translated by C. J. Hyman), Frederick Ungar Publishing Co., New York, 1956.
- Davis, P. J., *Interpolation and Approximation*, Blaisdell Publishing Co., New York, 1965.
- Krylov, V. I., *Approximate Calculation of Integrals* (translation by A. H. Stroud), The Macmillan Co., New York, 1962.
- Langer, R. E. (editor), *On Numerical Approximation*, University of Wisconsin Press, Madison, Wisc., 1959.
- Lorentz, G. G., *Bernstein Polynomials*, University of Toronto Press, Toronto, 1953.
- Milne-Thomson, L. M., *Calculus of Finite Differences*, Macmillan and Co., London, 1933.
- Steffensen, J. F., *Interpolation*, Chelsea Publishing Co., New York, 1927.
- Stiefel, E., "Note on Jordan elimination, linear programming and Tchebycheff approximation," *Numer. Math.*, **2**, 1-17 (1960).
- Stroud, A. H., "A bibliography on approximate integration," *Math. Comp.*, **15**, 52-80 (1961).
- Szegö, G., *Orthogonal Polynomials*, American Mathematical Society, New York, 1959.

### SPECIAL REFERENCES FOR CHAPTER 8

- Dahlquist, G., "Convergence and stability in the numerical integration of ordinary differential equations," *Math. Scand.*, **4**, 33-53 (1956).
- Fox, L., *Numerical Solution of Two-Point Boundary Problems*, Clarendon Press, Oxford, 1957.
- Henrici, P., *Discrete Variable Methods in Ordinary Differential Equations*, John Wiley and Sons, New York, 1962.
- , *Error Propagation for Difference Methods*, John Wiley and Sons, New York, 1963.
- Milne, W. E., *Numerical Solution of Differential Equations*, John Wiley and Sons, New York, 1953.

### SPECIAL REFERENCES FOR CHAPTER 9

- Collatz, L., *The Numerical Treatment of Differential Equations*, 3rd ed., Springer-Verlag, Berlin, 1960.

- Courant, R., K. O. Friedrichs, and H. Lewy, "Über die partiellen Differenzengleichungen der mathematischen Physik," *Math. Ann.*, **100**, 32–74 (1928). (Translation by P. Fox: "On the partial difference equations of mathematical physics," New York University, Courant Institute of Mathematical Sciences, Report NYO-7689, 1956.)
- Douglas, J., Jr., "On the relation between stability and convergence in the numerical solution of linear parabolic and hyperbolic differential equations," *SIAM Journal*, **4**, 20–37 (1956).
- , "Survey of numerical methods for parabolic differential equations," *Advances in Computers*, **2**, Academic Press, New York, 1961.
- Forsythe, G. E., and W. R. Wasow, *Finite-Difference Methods for Partial Differential Equations*, John Wiley and Sons, New York, 1960.
- Kreiss, H. O., "On difference approximations of the dissipative type for hyperbolic differential equations," *Comm. Pure Appl. Math.*, **17**, 335–353 (1964).
- Lax, P., "Numerical solution of partial differential equations," *Amer. Math. Monthly*, **72**, No. 2, part II, 74–84 (1965).
- Peaceman, D. W. and H. H. Rachford, Jr., "The numerical solution of parabolic and elliptic differential equations," *SIAM Journal*, **3**, 28–41 (1955).
- Richtmyer, R. D., *Difference Methods for Initial Value Problems*, Interscience, New York, 1957.
- Varga, R. S., *Matrix Iterative Analysis*, Prentice-Hall, Englewood Cliffs, N.J., 1962.



# Index

- Acceleration, of iterative methods, 73–81  
parameter(s), 122, 468 ff., 476 ff.  
    optimal, 470  
procedure, 120, 151
- Adams method, 388  
    improved, 394
- Aitken,  $\delta^2$ -method, 102–108, 260  
     $\delta^2$ -process, 109, 152, 374  
    iterative interpolation, 259
- Algorithm, 21  
    convergent, 1, 23
- Alternative principle, 29, 422, 436
- Altman, 83
- Analytic(ity), 279  
    domain of, 279
- Anti-symmetric, 266
- A posteriori (error) estimate, 26, 46–49, 140
- A posteriori test, 123
- A priori (error) estimate, 26, 37–46, 140, 169, 448, 462
- Backward error analysis, 38
- Bairstow's method, 131 ff.
- Balanced method, 390, 397
- Bauer, 139
- Bernoulli's method, 128 ff., 133, 158
- Bernstein polynomials, 183 ff., 192
- Bessel's inequality, 197, 203, 238
- Best approximation, 221 ff., 477  
    trigonometric, 240 ff.
- Binomial expansion, 184
- Biorthogonal, 137
- Biorthogonalization, 154
- Bisection method, 128
- Boundary conditions, 421 ff., 443, 483
- Boundary points, *see* Net points
- Boundary value problem, 421 ff.  
    linear, 421 ff.
- Cauchy, problem, 479 ff.  
    Schwarz inequality, 5, 146, 219, 220  
    sequence, 88
- Centered difference, approximation, 293  
    method, 377 ff.
- Centroid, 357
- Characteristic(s), 479 ff.  
    directions, 482  
    equation, 129, 134  
        homogeneous, 483  
    form, 481 ff.  
    polynomial, 2, 407  
    slope, 487  
    variables, 481 ff.
- Chebyshev, 224, 267  
    approximations, 211  
    polynomials, 81, 152, 203, 209 ff., 214, 219, 221, 226 ff., 236
- Chopping, 18
- Chord method, 97, 98, 113, 437
- Christoffel-Darboux relation, 205, 333
- Christoffel numbers, 333, 335
- Codiagonal elements, 164

- Compatibility condition, strong, 514  
weak, 514
- Complete linear space, 194
- Component, 136
- Composite rules, 337 ff.
- Computing problem, 21
- Comrie, 280
- Condition number, 27, 37
- Conjugate transpose, 10
- Consistent, 365, 368, 411, 514, 516  
conditionally, 516
- Contracting mapping, 85
- Convergence, 517  
in the mean, 197, 239  
of difference solutions, 491 ff., 514 ff.  
properties, 365
- Convergence factor, 119  
asymptotic, 91
- Convergent, approximate solutions, 519  
conditionally, 490, 505  
method, 412  
unconditionally, 505, 508
- Corrector, 386 ff.
- Courant, 445  
condition, 489  
Friedrichs-Lewy condition, 489
- Crank-Nicolson scheme, 509, 530
- Crout reduction, 51
- Cubature formula, 353 ff.  
composite, error, 362
- Cyclic, alternating direction method,  
477  
Jacobi scheme, 163  
parameters, 80, 476 ff.
- Dahlquist, 417
- Deflation methods, 152 ff.
- Degree of precision, *see* Precision
- De la Vallée-Poussin, 223
- $\delta^2$ -method, 102–108
- $\delta^2$ -process, 109, 152, 158
- Dependence, domain of, 480, 485 ff.  
501  
interval of, 480  
numerical domain of, 487
- Detecting isolated errors, 263
- Difference equation(s), 129, 365  
homogeneous, 406  
linear, 405 ff.
- Difference methods, consistency, 410 ff.,  
514 ff.
- Difference methods, convergence,  
410 ff., 514 ff.  
convergent, 518  
stability, 410 ff., 514 ff., 519
- Difference quotients, 445  
backward, 445  
centered, 445  
forward, 445
- Difference(s), average value of, 274  
centered, 272, 283  
forward, 260 ff.  
modified second, 274  
Newton's backward formula, 283  
operators, calculus of, 281 ff.
- Diffusion equation, 443, 501 ff.
- Direct methods, 427 ff.
- Dirichlet problem, 446
- Discontinuous integrands, 346 ff.
- Discrete orthonormality, 215
- Discretization error, 368
- Distance from a set, 142
- Divergent method, 379 ff.
- Divided differences, 246 ff., 296
- Domain of dependence condition, 489  
*see also* Dependence
- Double precision, 52
- Dufort-Frankel, 516
- Duhamel's principle, 409
- Eigenfunctions, 209, 434 ff., 458 ff.
- Eigenpairs, 135
- Eigen-problems, 434 ff.
- Eigenvalue-eigenvector problem, 134–  
175
- Eigenvalue(s), 134 ff., 209, 434 ff.,  
458 ff.  
localizing, 135  
principal, 147  
problems, 421 ff., 434 ff., 455 ff.
- Eigenvector(s), 134 ff., 169  
left, 137  
orthonormal, 11  
right, 137  
row, 137
- Elimination, block, 59–61, 463  
group, 59–61
- Equivalent systems, 29
- Error estimates, *a posteriori*; *see A*  
*a posteriori*  
*a priori*; *see A priori*
- Error factor, 267

- Error propagation, 91  
 Euclidean algorithm, 127  
 Euler-Cauchy method, 366 ff., 383, 395, 405  
     convergence, 376  
     modified, 388, 394, 402, 405  
 Euler-Maclaurin summation formula, 287, 288, 340  
 Everett, *see* Interpolation polynomial(s)  
 Explicit schemes, 365, 385, 487  
 Exponent, 18  
 External points, 444  
 Extrapolation, 73  
     to zero mesh width, 374, 383
- Factorial polynomial, 265  
 Factorization methods, 52–61, 171 ff  
 False position, 98, 99–102  
 Fibonacci sequence, 101  
 Fike, 139  
 Finite difference methods, 427 ff.  
     eigenvalue problem, 455 ff.  
     error estimate, 429  
 Finite jump discontinuities, 346  
 First order method, 91  
 Floating-point arithmetic, 17  
 Forsythe, 163  
 Forward differences, 260 ff.  
 Fourier, 239  
     coefficients, 239, 484, 524  
     series, 237 ff.  
 Francis, 173  
 Franklin, 143  
 Friedrichs, 445  
 Functional, analysis, 1  
     iteration, 386  
 Fundamental set of solutions, 406  
 Gauss-Jordan elimination, 50  
 Gauss-Seidel, *see* Iteration; Line  
 Gershgorin, 135  
 Givens, 140, 160, 168  
     transformation, 164 ff.  
 Goldstine, 49  
 Gram polynomials, 214  
 Gram-Schmidt orthonormalization method, 199, 212, 218  
 Grid, 364, 444  
 Haar property, 240  
 Hadamard, 21, 444
- Heat conduction equation, 443, 501 ff.  
 Henrici, 163  
 Hermite, interpolation, 192  
     polynomials, 221, 351  
 Hermitian, 70  
     transpose, 10  
 Hessenberg form, 160, 170  
 Heun's method, 402, 405  
 Higher order equations, 418 ff.  
 Hilbert segments, 196, 217  
 Homogeneous problem, 422  
 Householder, 160  
     method, 165 ff.  
 Hyman, 170  
 Hyperbolic type, 482
- Ideal method, 366  
 Implicit schemes, 365, 385, 392, 505 ff.  
 Infinite integrand, 346 ff.  
 Infinite integration limits, 350  
 Initial conditions, 443 ff., 483  
 Initial value methods, 424 ff.  
     problem, 364 ff., 422, 443, 479 ff., 501 ff.  
 Inner product, 19, 134, 199, 212  
 Instability, 504 ff.  
 Interior points, *see* Net points  
 Interpolation formulae, centered, 270 ff.  
 Interpolation polynomial(s), 187 ff., 264 ff.  
     divergence of, 275  
     error, 189 ff., 248, 265 ff., 275, 296  
     Everett's form, 273, 280, 319  
     Gaussian (forward) form, 271  
     Hermite, 208  
     Newton, 246 ff.  
     osculatory, 192 ff., 255  
     remainder, 265  
 Interpolatory, 302  
 Inverse iteration, 152 ff.  
 Iteration(s), alternating direction, 475 ff.  
     block, 72, 471 ff.  
     extrapolated, 79  
     functional, 85 ff.  
     Gauss-Seidel, 66, 71, 465 ff., 470  
     Gauss-Seidel, accelerated, 470  
     higher order, 94  
     interpolated, 79  
     inverse, 157

- Iteration(s), Jacobi, 64, 72, 464, 467  
 line, 471 ff.  
 simultaneous, 64, 464  
 successive, 66
- Iterative interpolation, 102  
 inverse, 259  
 linear, 258 ff.
- Iterative methods, 61–84, 85–133, 463 ff.
- Jacobian, 113, 420
- Jacobi, method, 160 ff.  
 polynomials, 335
- Jordan, 50  
 form, 2
- Kantorovich, 119
- Kernel, 198, 205
- Kronecker delta, 27
- Kublanovskaja, 173
- Kutta's method, 402, 405
- Lagrange, interpolation coefficients, 189 ff., 218, 264 ff.  
 interpolation polynomial, 189 ff., 218, 246  
 multipliers, 322
- Laguerre polynomials, 221, 351
- Laplace equation, 443, 445 ff.
- Laplacian, 445
- Lattice, 364, 444
- Lax, 529
- Least squares approximation, 194 ff., 237 ff.
- Lebesgue integral, 194
- Legendre polynomials, 202, 206, 218 ff.
- Leibnitz' rule, 106
- Lemniscates, 279
- Lewy, 445
- Liebmann method, 465
- Line, accelerated successive, 473  
 Gauss-Seidel, 473  
 Gauss-Seidel, accelerated, 474  
 iterations, 471 ff.  
 Jacobi method, 471 ff.
- Linear independence, 199
- Linearly independent, 406
- Lipschitz, condition, 86, 183  
 constant, 85, 405  
 continuity, 22  
 continuous, 364, 413
- Lower Hessenberg form, 167
- Majorizing set, 378
- Mantissa, 18
- Matrix, amplification, 525  
 approximate inverse, 27  
 augmented, 29  
 block-tridiagonal, 58  
 circulant, 175  
 complex conjugate transpose, 137  
 convergent, 14, 63  
 diagonalizable, 496  
 formulation, 452 ff.  
 Hermitian, 12, 137, 140  
 Hessenberg, 49, 160, 170  
 Hilbert segment, 196, 217  
 identity, 27  
 ill-conditioned, 37  
 inverse, 27  
 Jacobi, 55  
 lower triangular, 31  
 permutation, 33  
 perturbed, 137  
 residual, 47  
 singular, 169  
 sparse, 156  
 symmetric, 151, 436  
 transformations, 159 ff.  
 triangular, 31  
 tridiagonal, 55, 164 ff.  
 uniformly bounded, 526  
 unitary, 139, 144  
 upper triangular, 31  
 well-conditioned, 37  
 zero, 14
- Maximum absolute column sum, 10
- Maximum absolute row sum, 9
- Maximum norm, 4, 9, 178
- Maximum principle, 431, 439, 447 ff., 461, 513
- Mean convergence, 197, 239
- Mean square error, 196
- Measure zero, 194
- Mesh, 364, 444  
 ratio, 490, 501  
 widths, 364
- Midpoint rule, 240, 316, 323  
 composite, 343, 344
- Milne's method, 388, 394
- Minimizing sequence, 244
- Minimum principle, 513
- Minkowski's inequality, 6

- Mixed initial-boundary value problem, 483 ff.
- Möbius transformations, 77
- Modulus of continuity, 183, 343
- Moulton's method, 388
- Multiple integrals, 352 ff.  
composite formulae, 361 ff.
- Multiplier(s), 33, 35
- Multipoint methods, 102
- Multistep methods, 384 ff.
- Multivariate interpolation, 294 ff.
- Nesting procedure, 124
- Net, 364, 444  
function, 365  
points, boundary, 444, 515  
interior, 444, 515  
slope, 487  
spacing, 364, 444  
spacing change, 393  
uniform, 367
- Neville's iterated interpolation, 259
- Newton-Cotes formula, 308 ff., 354, 391  
closed, 308 ff., 316, 337  
error, 310 ff., 316  
open, 313, 316
- Newton-Raphson method, 133
- Newton's identities, 324  
interpolation polynomial, 246 ff., 283, 296  
method, 84, 97–99, 113, 115–119, 126, 131, 427, 433, 441
- Nodes, *see* Quadrature
- Normal system, 195, 196, 211
- Norm(s), 1–17, 176 ff., 211  
compatible, 8  
essentially strict, 220  
Euclidean, 4, 10, 12, 525  
induced, 8  
 $\mathcal{L}^2$ , 525  
maximum, 4, 9, 178, 221, 240, 450  
natural, 8  
operator, 8  
 $p$ , 4, 139  
semi-, 17, 177 ff.  
spectral, 12  
strict, 181, 211  
uniform, 4
- Notation, 444 ff., 448  
subscript, 445, 448
- $n$ th order method, 96
- Null space, 29, 145
- Numerical differentiation, 288 ff.  
integration formula, 300 ff.  
quadrature, 300 ff.
- $o(1)$ , 109
- One-step methods, 395 ff., 419
- Open formula, 392
- Operational counts, 34
- Operations, 35
- Operator, centered difference, 283  
derivative, 281  
difference, 281, 439  
displacement, 281  
identity, 281  
Laplace, 445 ff.  
linear, 301  
linear interpolation, 461 ff.  
method, 281 ff.  
shift, 282
- Ops., 35
- Optimal parameter, 74, 470
- Optimal (parameter) value, 151
- Orthogonal, 137  
functions, 196 ff.  
polynomials, 203 ff.
- Orthogonality relations, 435
- Orthogonalization method, 152 ff.
- Orthonormal functions, 197 ff., 202
- Osculating polynomial, 192 ff.
- Overflow, 18
- Over-relaxation, 73
- Parseval's equality, 197, 203, 238, 244, 525
- Peaceman and Rachford, 475
- Perturbations, 43, 140
- Picard iteration, 85
- Pivot, element, 32  
maximal, 34  
maximal column, 34, 169
- Pivoting, maximal column, 45  
partial, 45
- Pointwise convergence, 205 ff.
- Pointwise error, 412
- Poisson equation, 446 ff.
- Polygon method, 367 ff.
- Positive definite, 17, 49, 457  
systems, 70
- Power method, 147

- Precision, degree of, 301, 311, 312, 327 ff., 336, 359, 360, 401
- Predictor, 86 ff.
- Predictor-corrector method, 386 ff., 419  
error estimates, 388 ff.
- Properly posed, 92, 444
- $p$ th order method, 82
- QR (factorization) method, 169, 173
- Quadratic convergence, 113
- Quadrature, coefficients, 300, 353  
continuous functions, 341 ff.  
error, 300 ff., 320  
interpolatory, 303
- formula, composite, 302, 336 ff.  
composite error, 338  
simple, 302, 336, 384 ff., 400 ff.
- Gauss-Chebyshev, 334 ff., 350
- Gauss-Hermite, 351
- Gauss-Laguerre, 352
- Gaussian, 327 ff.  
error, 329  
interpolatory, 303 ff.  
nodes, 300, 331, 353  
periodic functions, 340 ff.  
points, 300  
uniform coefficients, 319 ff.  
weighted, 331 ff., 349  
weighted interpolatory, 332  
weighted Gaussian, 333, 351
- Rachford, 475
- Randomness, 320
- Rank, 29  
column, 29  
row, 29
- Rate of convergence, 64, 91, 148, 464, 467, 470, 472
- Rational function, 176
- Rayleigh quotient, 142
- Rayleigh-Ritz, 461
- Regula Falsi, 101, 260  
classical, 102
- Remainder term, 265
- Residual, 69  
correction, 68
- Richardson's deferred approach to the limit, 374, 383
- Rolle's theorem, 190, 289
- Romberg's method, 344
- Root condition, 412
- Roots of unity, 456
- Rotations, two dimensional, 160 ff.
- Rounding, 18, 91  
errors, 517  
unbiased, 18
- Roundoff, 94  
average, 321  
contamination, 153  
error, 38, 264, 319 ff., 374 ff., 451, 516  
accumulated, 320  
initial, 375  
local, 375  
mean-accumulated, 321  
mean-square error, 322, 336  
root mean-square error, 322  
statistical notions, 320  
uncorrelated, 321
- Runge-Kutta methods, 402, 405
- Runge's example, 191, 275
- Rutishauser, 172
- Scale, 46
- Schur, 3, 157
- Schwarz' inequality, *see* Cauchy-Schwarz
- Second order method, 94
- Semi-norm, 17, 177 ff.
- Separation of variables, 359, 455 ff., 458, 483 ff., 493
- Separation property, 168
- Shooting methods, 424 ff.
- Signal, 479 ff., 501
- Simpson's rule, 287, 316, 339
- Single-step methods, 395 ff., 402
- Singular, integrals, 346 ff.  
matrix, 169
- Smooth functions, 515
- Spectral radius, 10, 63, 510
- Splitting(s), 62, 74  
family of, 75
- Stability, 24, 365, 413, 514 ff., 519, 522 ff.  
test, 523 ff.
- Stable, conditionally, 519  
in the  $\mathcal{L}^2$ -norm, 525  
unconditionally, 519  
weakly, 523
- Star, 446, 502
- Starring, 137
- Stationary values, 459 ff.

- Steffensen's method, 103
- Stencil, 446, 502
- Stirling's formula, 267, 279
- Sturm-Liouville problems, 434
- Sturm sequence(s), 126, 168
- Subdivision, 364
- Summation by parts, 213
- Symmetric, 266
- Symmetric functions, 247
  - elementary, 324
- Synthetic division, 125, 131
- Taylor's series, finite, 397 ff.
- Threshold(s), 163
  - scheme, 163
- Total error, 375
- Trapezoidal rule, 239, 316, 318, 339
  - end corrections, 343
  - extrapolation of, 344
  - periodic functions, 340 ff.
- Triangle inequality, 4, 177, 219
- Triangular, array, 297
  - decomposition, 52
- Tridiagonal form, 164 ff.
- Trigonometric, approximation, 229 ff.
  - interpolation, 230 ff.
  - least squares approximation, 237 ff.
  - sum, 229
- Trigonometric functions, discrete ortho-normality, 215
- Truncation error, local, 368, 387, 395, 411, 449, 490, 497, 507, 508, 516
  - order, 412
- Underflow, 18
- Undetermined coefficients, method of, 292, 315, 317, 333, 356 ff., 380
- Unit ball, 7
- Unstable, 366, 504 ff., 521
  - schemes, 382
- Vandermonde determinant, 129, 188, 193, 242, 291, 317
- Varga, 478
- Variational equation, 427, 441
- Variational principles, 435, 459 ff.
- Vector, compound, 454
  - orthonormal, 149
  - residual, 48, 140
  - zero, 4
- Vector space, complex, 2
  - linear, 406
- Volume, 321
- von Neumann, 49, 523 ff.
  - condition, 526
- Wave, 479 ff.
  - equation, 443, 479 ff.
- Weierstrass, 7, 180
  - approximation theorem, 183 ff., 230
- Weight function, 202, 331
- Weighted least squares approximation, 202 ff.
- Well-posed, computing problem, 1, 22
  - problem, 23, 27, 444
- Well-posed(ness), 139 ff.
- Wilkinson, 27, 44, 49, 140, 172, 174