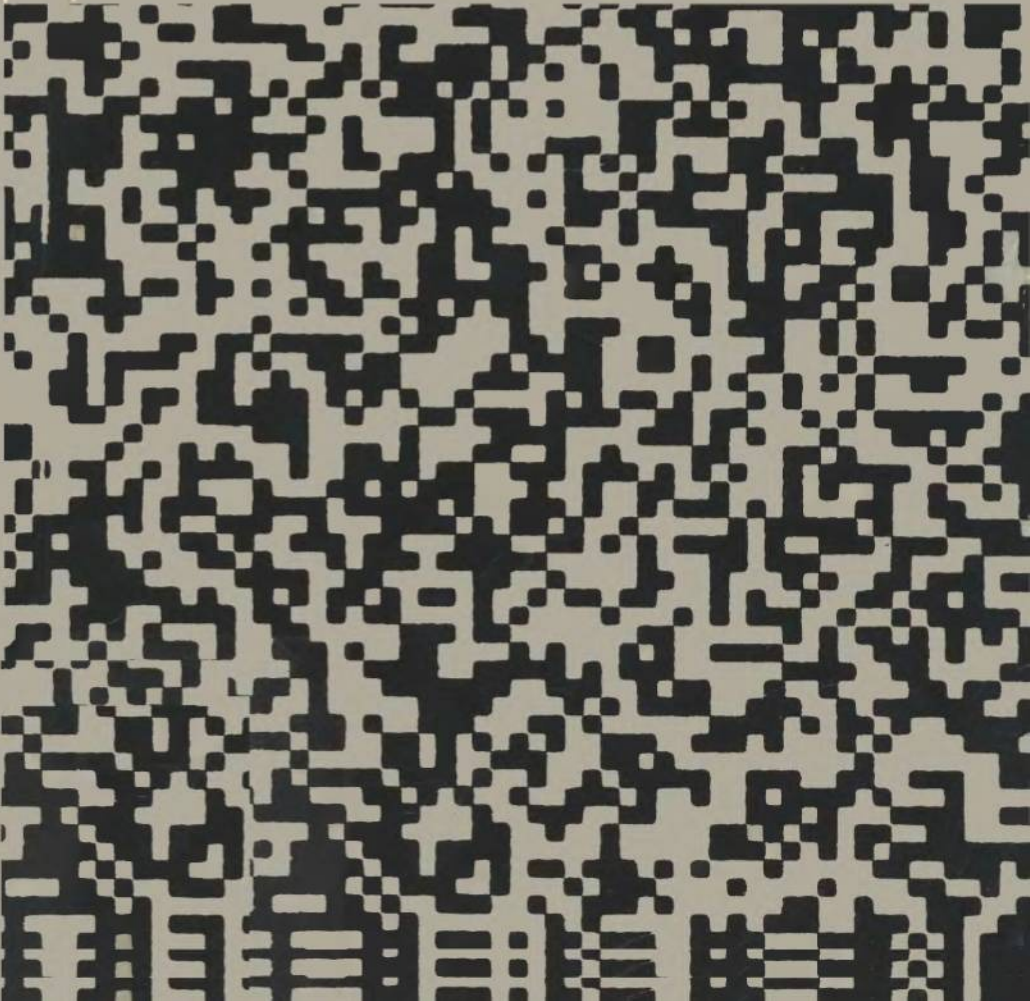
OXFORD APPLIED MATHEMATICS
AND COMPUTING SCIENCE SERIES

Numerical Solution of Partial Differential Equations: Finite Difference Methods

G.D. SMITH

Third Edition



SMITH
Numerical Solution of Partial Differential Equations:
Finite Difference Methods
Third Edition
OXFORD

Oxford Applied Mathematics and Computing Science Series

General Editors
J. N. Buxton, R. F. Churchhouse, and A. B. Tayler

G. D. SMITH
Brunel University

Numerical solution of partial differential equations

FINITE DIFFERENCE METHODS

THIRD EDITION

CLARENDON PRESS · OXFORD

OXFORD APPLIED MATHEMATICS AND COMPUTING SCIENCE SERIES

G. D. Smith: *Numerical Solution of Partial Differential Equations* (*Third Edition*)

J. R. Ullmann: *A Pascal Database Book*

S. Barnett and R. G. Cameron: *Introduction to Mathematical Control Theory* (*Second Edition*)

A. B. Tayler: *Mathematical Models in Applied Mechanics*

I. Anderson: *A First Course in Combinatorial Mathematics*

D. W. Jordan and P. Smith: *Nonlinear Ordinary Differential Equations* (*Second Edition*)

D. S. Jones: *Elementary Information Theory*

B. Carré: *Graphs and Networks*

A. J. Davies: *The Finite Element Method*

W. E. Williams: *Partial Differential Equations*

R. G. Garside: *The Architecture of Digital Computers*

J. C. Newby: *Mathematics for the Biological Sciences*

R. Hill: *A First Course in Coding Theory*

P. Baxardall and H. Liebeck: *Vector Calculus*

Oxford University Press, Walton Street, Oxford OX2 6DP
Oxford New York Toronto
Delhi Bombay Calcutta Madras Karachi
Petaling Jaya Singapore Hong Kong Tokyo
Nairobi Dar es Salaam Cape Town
Melbourne Auckland
and associated companies in
Berlin Ibadan

Oxford is a trade mark of Oxford University Press

Published in the United States
by Oxford University Press, New York

First edition published in Oxford Mathematical Handbooks 1965
Second edition 1978, reprinted (with corrections) 1979, 1984
Third edition 1985, reprinted 1986, reprinted (with corrections) 1987, 1989 (twic

British Library Cataloguing in Publication Data

Smith, G. D.
Numerical solution of partial differential
equations: finite difference methods.—3rd ed.
—(Oxford applied mathematics and computing
science series)
1. Differential equations, Partial—Numerical
solutions
I. Title
515.3'53 QA374

ISBN 0-19-859641-3
ISBN 0-19-859650-2 (pbk)

Typeset and printed by The Universities Press (Belfast) Ltd

Preface to the third edition

As with the two previous editions this book has been written primarily as a textbook for students with no previous knowledge of numerical methods whatsoever, and because of its purpose the various methods of solution and analysis have been illustrated by worked examples. In addition, every chapter, except the first, includes exercises with solutions that often extend the theory as well as amplify points in the text.

It is intended for students taking degree courses in mathematics, physics and engineering, and is self-contained in that the basic calculus and matrix algebra needed for the development of the subject will be known to most second-year students. It is also hoped that the book will continue to be useful to those postgraduate students who need a rapid and uncomplicated introduction to this field of study.

With regard to changes, the new Chapter 2 is essentially a combination of the main ideas of Chapters 2 and 3 of the second edition; I did this because students, both undergraduate and postgraduate, said they preferred a more immediate mixture of method and analysis. The section on stability has been completely rewritten and based on the Lax–Richtmyer definition which ensures convergence when the difference equations are consistent and stable. This has necessitated an introductory section on norms.

Chapter 3 can be omitted by those who need only a quick introduction to all three types of partial differential equations because it deals mainly with an alternative derivation of difference equations approximating parabolic equations. It also includes miscellaneous methods for non-linear equations, the improvement of accuracy, the analytical solution of difference equations, and a new section on an eigenvalue–eigenvector solution that is used to give a new method of approximation particularly suited to large values of t.

Chapter 4 contains the standard work on hyperbolic equations and has been enlarged only slightly by the application of new work done in Chapter 3.

Chapter 5 on elliptic equations and iterative methods remains unchanged, except that my teaching experience has led to a simpler presentation of the theoretical work associated with the SOR method.

Finally, it gives me very great pleasure to thank all those who have been of help to me in the preparation of the three editions. Mr E. L. Albasiny and Dr D. W. Martin of the National Physical Laboratory for a stimulating set of lectures they gave on numerical methods for partial differential equations at Brunel in 1960, Professor J. Crank for criticising and improving the first edition; Dr I. Parker for computing many of the numerical examples when a research student; Professor L. Fox and Dr J. Gregory for helpful suggestions concerning the second edition, and particularly Dr N. Papamichael for his criticisms of my initial draft and his permission to use parts of his M.Sc. dissertation on consistent orderings. With regard to the third edition, I am deeply grateful to Dr E. H. Twizell and Dr A. Q. M. Khaliq, Brunel University, for their work on L_0-stable methods, and to Mr D. Drew, Brunel University, for very recent work on semi-discrete semi-analytic approximation methods.

Brunel University G. D. S.
December 1984

Contents

3. PARABOLIC EQUATIONS: ALTERNATIVE DERIVATION OF DIFFERENCE EQUATIONS AND MISCELLANEOUS TOPICS

4. HYPERBOLIC EQUATIONS AND CHARACTERISTICS

5. ELLIPTIC EQUATIONS AND SYSTEMATIC ITERATIVE METHODS

Notation

$i=1(1)n$ $\quad$ i varies from 1 to n by intervals of 1, i.e. $i=1, 2, 3, \ldots, n-1, n$.

$S=\{a_1, a_2, \ldots, a_n\}$ $\quad$ S is the set whose members are $a_1, a_2, \ldots, a_n$.

$a_r \in S$ $\quad$ a_r is a member of the set S.

$S=\{(x, y): 0 \leqslant x \leqslant a, 0 \leqslant y \leqslant b\}$ $\quad$ S is the set of ordered pairs (x, y) such that $0 \leqslant x \leqslant a$ and $0 \leqslant y \leqslant b$.

$\mathbf{A}=[a_{ij}]$ $\quad$ $\mathbf{A}$ is the matrix whose element in the ith row and jth column is a_{ij}.

$\mathbf{I}_N$ $\quad$ The unit matrix of order N.

$\mathbf{A}^T$ $\quad$ The transpose of matrix $\mathbf{A}$, i.e. $\mathbf{A}^T=[a_{ji}]$.

$\mathbf{A}^{-1}$ $\quad$ The inverse of matrix $\mathbf{A}$, i.e., $\mathbf{A}\mathbf{A}^{-1}=\mathbf{I}$.

$\bar{\mathbf{A}}$ $\quad$ The matrix whose elements are the complex conjugates of the corresponding elements of matrix $\mathbf{A}$.

$\mathbf{x}=[x_1, x_2, \ldots, x_n]^T$ $\quad$ The column vector whose components are $x_1, x_2, \ldots, x_n$.

$|\lambda|$ $\quad$ The modulus of λ.

$\max_i |\lambda_i|, i=1(1)n.$ $\quad$ The maximum of $|\lambda_1|, |\lambda_2|, \ldots, |\lambda_n|$.

$\rho(\mathbf{A})$ $\quad$ The spectral radius of matrix $\mathbf{A}$, which is the maximum of the moduli of its eigenvalues λ_i, $i=1(1)n$.

1 Introduction and finite-difference formulae

The mathematical formulation of most problems in science involving rates of change with respect to two or more independent variables, usually representing time, length or angle, leads either to a partial differential equation or to a set of such equations. Special cases of the two dimensional second-order equation

$$a\frac{\partial^2\phi}{\partial x^2}+b\frac{\partial^2\phi}{\partial x\,\partial y}+c\frac{\partial^2\phi}{\partial y^2}+d\frac{\partial\phi}{\partial x}+e\frac{\partial\phi}{\partial y}+f\phi+g=0,$$

where a, b, c, d, e, f, and g may be functions of the independent variables x and y and of the dependent variable ϕ, occur more frequently than any other because they are often the mathematical form of one of the conservation principles of physics.

For reasons that are given in Chapter 4 this equation is said to be *elliptic* when $b^2-4ac<0$, *parabolic* when $b^2-4ac=0$, and *hyperbolic* when $b^2-4ac>0$.

Two-dimensional elliptic equations

These equations, of which the best known are Poisson's equation

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+g=0$$

and Laplace's equation

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}=0,$$

are generally associated with equilibrium or steady-state problems. For example, the velocity potential for the steady flow of incompressible non-viscous fluid satisfies Laplace's equation and is the mathematical way of expressing the idea that the rate at which such fluid enters any given region is equal to the rate at which it leaves it. Similarly, the electric potential V associated with a two-dimensional electron distribution of charge density ρ

satisfies Poisson's equation $\partial^2V/\partial x^2+\partial^2V/\partial y^2+\rho/\varepsilon=0$, where ε is the dielectric constant. This is the partial differential equation form of the well-known theorem by Gauss which states that the total electric flux through any closed surface is equal to the total charge enclosed.

The *analytical* solution of a two-dimensional elliptic equation is a function of the space co-ordinates x and y which satisfies the partial differential equation *at every point* of the area S inside a plane *closed* curve C and satisfies certain conditions *at every point* on this boundary curve C (Fig. 1.1). The function ϕ, for instance, from which we can calculate the displacements and shear stresses within a long solid elastic cylinder in a state of torsion satisfies

$$\frac{\partial^2\phi}{\partial x^2}+\frac{\partial^2\phi}{\partial y^2}+2=0$$

at every point of a right cross-section, and has a constant value round the perimeter of the cross-section. Similarly, the steady motion of incompressible viscous fluid through a straight uniform tube can be found from a function that satisfies Laplace's equation at every point of the cross-section and equals $\frac{1}{2}(x^2+y^2)$ at each point on the boundary.

The condition that the dependent variable must satisfy round the boundary curve C is termed the boundary condition.

To the present, only a limited number of special types of

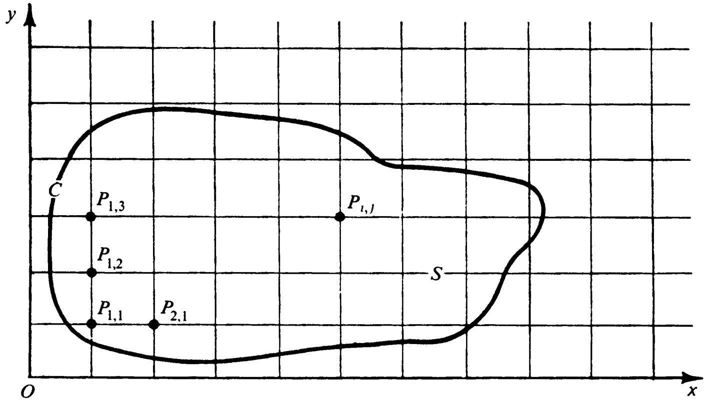


Fig. 1.1

elliptic equations have been solved analytically and the usefulness of these solutions is further restricted to problems involving shapes for which the boundary conditions can be satisfied. This not only eliminates all problems with boundary curves that are undefined in terms of equations, but also many for which the boundary conditions are too difficult to satisfy even though the equations for the boundary curves are known. In such cases approximation methods, whether analytical or numerical in character, are the only means of solution, apart from the use of analogue devices. Analytical approximation methods often provide extremely useful information concerning the character of the solution for critical values of the dependent variables but tend to be more difficult to apply than the numerical methods, and will not be discussed in this book. Of the numerical approximation methods available for solving differential equations those employing finite-differences or finite elements are more frequently used and more universally applicable than any other, although finite elements are not considered in this book. Before outlining these methods however, the reader should be aware of the manner in which the term 'approximation method' is used. Finite-difference methods are approximate in the sense that derivatives *at a point* are approximated by difference quotients *over a small interval*, i.e., $\partial\phi/\partial x$ is replaced by $\delta\phi/\delta x$ where δx is small and y is constant, but the solutions are *not* approximate in the sense of being crude estimates. The data of the problems of technology are invariably subject to errors of measurement, besides which, all arithmetical work is limited to a finite number of significant figures and contains rounding errors, so even analytical solutions provide only approximate numerical answers. Finite-difference methods generally give solutions that are either as accurate as the data warrant or as accurate as is necessary for the technical purposes for which the solutions are required. In both cases a finite-difference solution is as satisfactory as one calculated from an analytical formula. In future, all non-analytical approximation methods will be called numerical methods

They are not of course restricted to problems for which no analytical solutions can be found. The numerical evaluation of an analytical solution is often a laborious task, as can be seen by inspecting the solution of the torsion problem for a rectangular cross-section defined by $x = \pm a$, $y = \pm b$, namely

$$\phi = b^2 - y^2 - 32b^2\pi^{-3} \sum_{n=0}^{\infty} \frac{(-1)^n}{(2n+1)^3} \operatorname{sech} \frac{(2n+1)\pi a}{2b} \times \cosh \frac{(2n+1)\pi x}{2b} \cos \frac{(2n+1)\pi y}{2b},$$

and numerical methods generally provide adequate numerical solutions more simply and efficiently. This is certainly so with finite-difference methods for solving partial differential equations.

In these methods (Fig. 1.1), the area of integration of the elliptic equation, i.e. the area S bounded by the closed curve C, is overlayed by a system of rectangular meshes formed by two sets of equally spaced lines, one set parallel to Ox and the other parallel to Oy, and an approximate solution to the differential equation is found at the points of intersection $P_{1,1}, P_{1,2}, \ldots, P_{i,j}, \ldots$ of the parallel lines, which points are called mesh points. (Other terms in common use are pivotal, nodal, grid, or lattice points.) This solution is obtained by approximating the partial differential equation over the area S by n *algebraic* equations involving the values of ϕ at the n mesh points internal to C. The approximation consists of replacing each derivative of the partial differential equation at the point $P_{i,j}$ (say) by a finite-difference approximation in terms of the values of ϕ at $P_{i,j}$ and at neighbouring mesh points and boundary points, and in writing down for each of the n internal mesh points the algebraic equation approximating the differential equation. This process clearly gives n algebraic equations for the n unknowns $\phi_{1,1}$, $\phi_{1,2}, \ldots \phi_{i,j}, \ldots$. Accuracy can usually be improved either by increasing the number of mesh points or by including 'correction terms' in the approximations for the derivatives.

Parabolic and hyperbolic equations

Problems involving time t as one independent variable lead usually to parabolic or hyperbolic equations.

The simplest parabolic equation, $\partial U/\partial t = \kappa \partial^2 U/\partial x^2$, derives from the theory of heat conduction and its solution gives, for example, the temperature U at a distance x units of length from one end of a thermally insulated bar after t seconds of heat conduction. In such a problem the temperatures at the *ends* of a bar of length l (say) are often known for all time. In other words, the *boundary conditions* are known. It is also usual for the temperature distribution along the bar to be known at some particular instant. This instant is usually taken as zero time and the temperature distribution is called the *initial condition.* The solution gives U for values of x between 0 and l and values of t from zero to infinity. Hence the area of integration S in the x–t plane (Fig. 1.2), is the infinite area bounded by the x-axis and the parallel lines $x=0, x=l$. This is described as an *open* area because the boundary curves marked C do not constitute a closed boundary in any finite region of the x–t plane.

Applications of finite-difference methods of solution to parabolic equations are no different from their application to elliptic equations in so far as the integration of the differential equation over S is approximated by the solution of algebraic equations. The structure of the algebraic equations is different however in that it propagates the solution forward from one time row to the next in a step-by-step fashion.

Hyperbolic equations generally originate from vibration problems, or from problems where discontinuities can persist in time, such as with shock waves, across which there are discontinuities in speed, pressure and density. The simplest hyperbolic equation is the one-dimensional wave equation $\partial^2 U/\partial t^2 = c^2 \partial^2 U/\partial x^2$, giving, for example, the transverse displacement U at a distance x

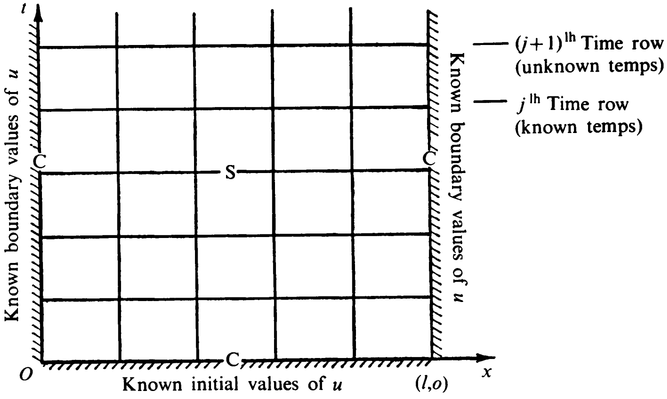


Fig. 1.2

from one end of a vibrating string of length l after a time t. As the values of U at the ends of the string are usually known for all time (the boundary conditions) and the shape and velocity of the string are prescribed at zero time (the initial conditions), it is seen (Fig. 1.2), that the solution is similar to that of a parabolic equation in that the calculation of U for a given x and t, $(0 \leqslant x \leqslant l)$, entails integration of the equation over the open area S bounded by the open curve C. Although hyperbolic equations can be solved numerically by finite-difference methods, those involving only two independent variables, x and t say, are often dealt with by the method of characteristics, especially if the initial conditions and/or boundary conditions involve discontinuities. This method finds special curves in the x–t plane, called characteristic curves, along which the solution of the *partial* differential equation is reduced to the integration of an *ordinary* differential equation. This ordinary equation is generally integrated by numerical methods.

In conclusion, it is worth noting that whereas changes to the shape of the area of integration or to the boundary and initial conditions of partial differential equations often make their analytical solutions impossible, such changes do not fundamentally affect finite-difference methods although they sometimes necessitate rather complicated modifications to the methods.

Finite-difference approximations to derivatives

When a function U and its derivatives are single-valued, finite and continuous functions of x, then by Taylor's theorem,

$$U(x+h) = U(x) + hU'(x) + \tfrac{1}{2}h^2U''(x) + \tfrac{1}{6}h^3U'''(x) + \dots \tag{1.1}$$

and

$$U(x-h) = U(x) - hU'(x) + \tfrac{1}{2}h^2u''(x) - \tfrac{1}{6}h^3U'''(x) \dots . \tag{1.2}$$

Addition of these expansions gives

$$U(x+h) + U(x-h) = 2U(x) + h^2U''(x) + O(h^4), \tag{1.3}$$

where $O(h^4)$ denotes terms containing fourth and higher powers of h. Assuming these are negligible in comparison with lower powers of h it follows that,

$$U''(x) = \left(\frac{d^2U}{dx^2}\right)_{x=x} \simeq \frac{1}{h^2}\{U(x+h) - 2U(x) + U(x-h)\}, \quad (1.4)$$

with a leading error on the right-hand side of order h^2.

Subtraction of eqn (1.2) from eqn (1.1) and neglect of terms of order h^3 leads to

$$U'(x) = \left(\frac{dU}{dx}\right)_{x=x} \simeq \frac{1}{2h}\{U(x+h) - U(x-h)\}, \quad (1.5)$$

with an error of order h^2.

Equation (1.5) clearly approximates the slope of the tangent at P by the slope of the chord AB, and is called a *central-difference* approximation. We can also approximate the slope of the tangent at P by either the slope of the chord PB, giving the *forward-difference* formula,

$$U'(x) \simeq \frac{1}{h}\{U(x+h) - U(x)\}, \quad (1.6)$$

or the slope of the chord AP giving the *backward-difference* formula

$$U'(x) \simeq \frac{1}{h}\{U(x) - U(x-h)\}. \quad (1.7)$$

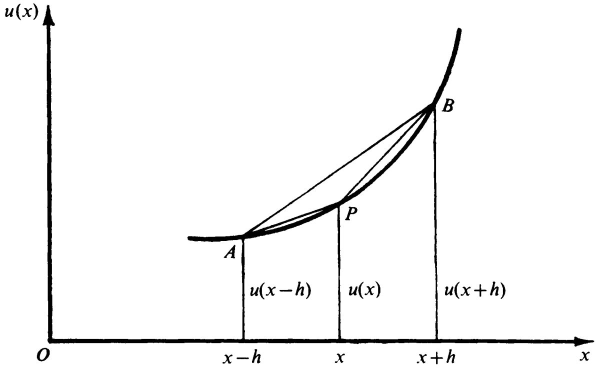


Fig. 1.3

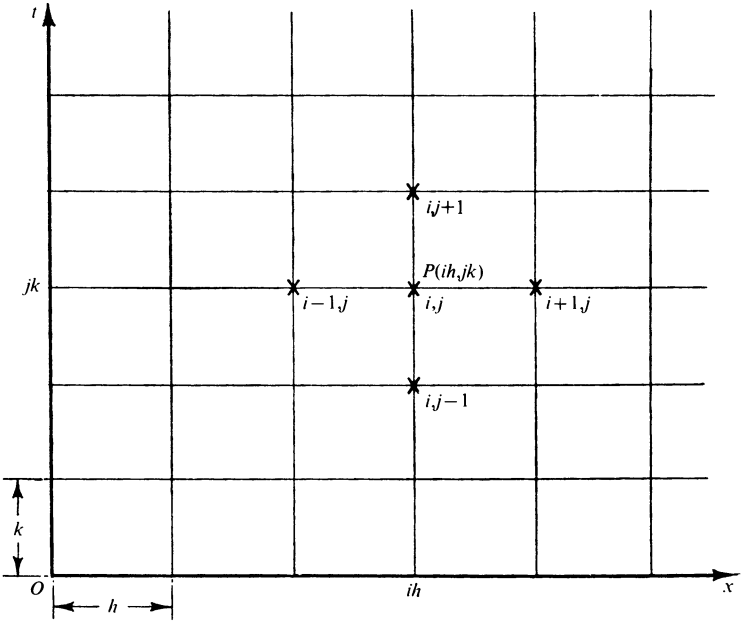


Fig. 1.4

Both (1.6) and (1.7) can be written down immediately from eqns (1.1) and (1.2) respectively, assuming second and higher powers of h are negligible. This shows that the leading errors in these forward and backward-difference formulae are both $O(h)$.

Notation for functions of several variables

Assume U is a function of the independent variables x and t. Subdivide the x–t plane into sets of equal rectangles of sides $\delta x = h$, $\delta t = k$, by equally spaced grid lines parallel to Oy, defined by $x_i = ih$, $i = 0, \pm 1, \pm 2, \ldots$, and equally spaced grid lines parallel to Ox, defined by $t_j = jk$, $j = 0, 1, 2, \ldots$, as shown in Fig. 1.4.

Denote the value of U at the representative mesh point $P(ih, jk)$ by

$$U_P = U(ih, jk) = U_{i,j}.$$

Then by eqn (1.4),

$$\left(\frac{\partial^2 U}{\partial x^2}\right)_P = \left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j} \simeq \frac{U\{(i+1)h, jk\} - 2U\{ih, jk\} + U\{(i-1)h, jk\}}{h^2}.$$

i.e.

$$\frac{\partial^2 U}{\partial x^2} \simeq \frac{U_{i+1,j} - 2U_{i,j} + U_{i-1,j}}{h^2}, \tag{1.8}$$

with a leading error of order h^2. Similarly,

$$\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j} \simeq \frac{U_{i,j+1} - 2U_{i,j} + U_{i,j-1}}{k^2}, \tag{1.9}$$

with a leading error of order k^2.

With this notation the forward-difference approximation for $\partial U/\partial t$ at P is

$$\frac{\partial U}{\partial t} \simeq \frac{U_{i,j+1} - U_{i,j}}{k}, \tag{1.10}$$

with a leading error of $O(k)$.

2 Parabolic equations: finite difference methods, convergence, and stability

Transformation to non-dimensional form

The computational stage of all numerical methods for solving problems of any complexity generally involves a great deal of arithmetic. It is usual therefore to arrange, whenever possible, for one solution to suffice for a variety of different problems. This can be done by expressing all equations in terms of non-dimensional variables. Then all problems with the same non-dimensional mathematical formulation can be dealt with by means of one solution. For example, the oscillation of a pendulum in a viscous medium and the discharge of electricity from a capacitance through a resistance and inductance are different problems physically, but identical mathematically when expressed in terms of non-dimensional variables. The problems need not, of course, be dimensionally different, but merely variations of the same type of problem, as we would have with the calculation of the periods of oscillation of springs of different lengths l supporting different masses m and having different stiffnesses s. A single solution of the corresponding non-dimensional equation would allow us to solve a wide variety of spring problems because a single parameter ξ, say, would replace some combination of l, m, and s.

This non-dimensionalizing process is illustrated below with the parabolic equation

$$\frac{\partial U}{\partial T}=\kappa\frac{\partial^2 U}{\partial X^2},\quad \kappa \text{ constant}, \tag{2.1}$$

the solution of which gives the temperature U at a distance X from one end of a thin uniform rod after a time T. (This assumes the rod is heat-insulated along its length so that temperature changes occur through heat conduction along its length and heat transfer at its ends.) Let L represent the length of the rod and U_0 some particular temperature such as the maximum or minimum temperature at zero time. Put

$$x=\frac{X}{L}\quad\text{and}\quad u=\frac{U}{U_0}.$$

Then

$$\frac{\partial U}{\partial X}=\frac{\partial U}{\partial x}\frac{\mathrm{d}x}{\mathrm{d}X}=\frac{\partial U}{\partial x}\frac{1}{L}$$

and

$$\frac{\partial^2 U}{\partial X^2}=\frac{\partial}{\partial X}\left(\frac{\partial U}{\partial X}\right)=\frac{\partial}{\partial x}\left(\frac{1}{L}\frac{\partial U}{\underset{\partial x}{\partial \cancel{X}}}\right)\frac{\mathrm{d}x}{\mathrm{d}X}=\frac{1}{L^2}\frac{\partial^2 U}{\partial x^2},$$

so eqn (2.1) transforms to

$$\frac{\partial(uU_0)}{\partial T}=\frac{\kappa}{L^2}\frac{\partial^2(uU_0)}{\partial x^2},$$

i.e.

$$\frac{1}{\kappa L^{-2}}\frac{\partial u}{\partial T}=\frac{\partial^2 u}{\partial x^2}.$$

Writing $t=\kappa T/L^2$ and applying the function of a function rule to the left side yields

$$\frac{\partial u}{\partial t}=\frac{\partial^2 u}{\partial x^2} \tag{2.2}$$

as the non-dimensional form of (2.1).

It should be noted that the number representing the length of the rod is 1.

An explicit method of solution

By eqns (1.10) and (1.8) one finite-difference approximation to

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2} \tag{2.3}$$

is

$$\frac{u_{i,j+1}-u_{i,j}}{k}=\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2},$$

where u is the exact solution of the approximating difference equations,

$$x_i = ih, \quad (i = 0, 1, 2, \ldots),$$

and

$$t_j = jk, \quad (j = 0, 1, 2, \ldots).$$

This can be written as

$$u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{i,j} + ru_{i+1,j}, \tag{2.4}$$

where $r = \delta t/(\delta x)^2 = k/h^2$, and gives a formula for the unknown 'temperature' $u_{i,j+1}$ at the $(i, j+1)$th mesh point in terms of known 'temperatures' along the jth time-row (Fig. 2.1). Hence we can calculate the unknown pivotal values of u along the first time-row, $t = k$, in terms of known boundary and initial values along $t = 0$, then the unknown pivotal values along the second time-row in terms of the calculated pivotal values along the first, and so on. A formula such as this which expresses *one* unknown pivotal value directly in terms of known pivotal values is called an explicit formula.

Example 2.1

As a numerical example let us solve (2.4) given that the ends of the rod are kept in contact with blocks of melting ice and that the

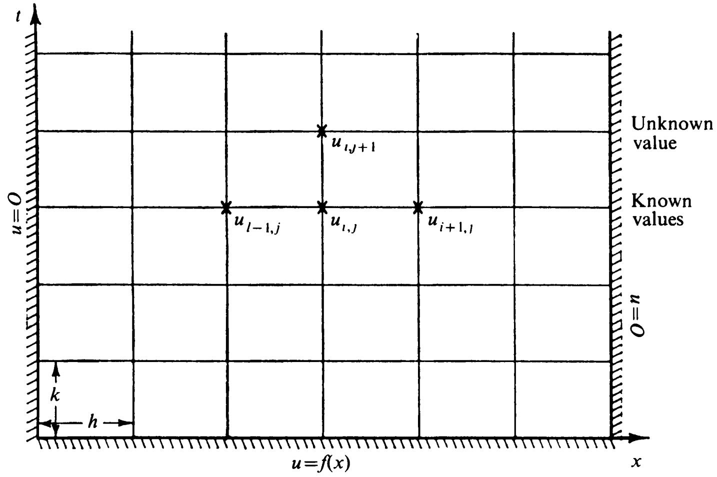


Fig. 2.1

initial temperature distribution in non-dimensional form is

$$\begin{aligned}&\text{(a)}\ U=2x &&0\leqslant x\leqslant\tfrac{1}{2},\\ &\text{(b)}\ U=2(1-x), &&\tfrac{1}{2}\leqslant x\leqslant 1.\end{aligned}\tag{2.5}$$

In other words, we are seeking a numerical solution of $\partial U/\partial t=\partial^2 U/\partial x^2$ which satisfies

(i) $U=0$ at $x=0$ and 1 for all $t>0$. (The boundary condition.)

(ii) $\left.\begin{array}{l}U=2x \text{ for } 0\leqslant x\leqslant\tfrac{1}{2}\\ \text{and } U=2(1-x) \text{ for } \tfrac{1}{2}\leqslant x\leqslant 1.\end{array}\right\}$ $t=0$. (The initial condition.)

(This initial temperature distribution could be obtained by heating the centre of the rod for a long time and keeping the ends in contact with the ice.)

For $\delta x=h=\frac{1}{10}$, the initial values and boundary values are as shown in Table 2.1. The problem is symmetric with respect to $x=\frac{1}{2}$ so we need the solution only for $0\leqslant x\leqslant\frac{1}{2}$.

Case 1

Take $\delta x=h=\frac{1}{10}$, $\delta t=k=\frac{1}{1000}$, so $r=k/h^2=\frac{1}{10}$. Equation (2.4) then reads as

$$u_{i,j+1}=\tfrac{1}{10}(u_{i-1,j}+8u_{i,j}+u_{i+1,j}).\tag{2.6}$$

For pencil and paper calculations the relationship between these four function values is represented very conveniently by the 'molecule' in Fig. 2.2. The numbers in the 'atoms' are the multipliers of the function values at the corresponding mesh points.

Application of eqn (2.6) to the data of Table 2.1 is shown in Table 2.2, and readers are recommended to check some of the

TABLE 2.1

	$x=0$	0.1	0.2	0.3	0.4	0.5	0.6	x
$j=0$	0	0.2	0.4	0.6	0.8	1.0	0.8	
$j=1$	0							
$j=2$	0							
$j=3$	0							
$j=4$	0							
t								

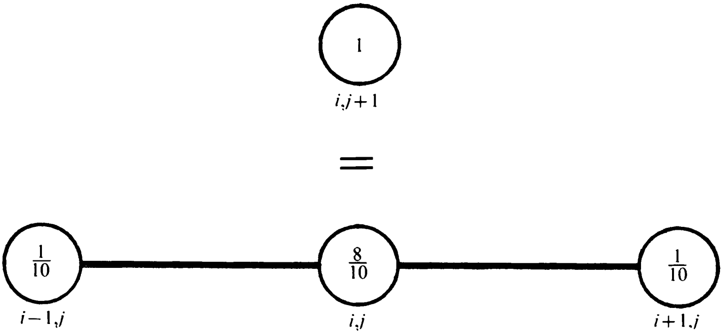


Fig. 2.2

calculations, remembering that the values of U at $x=\frac{4}{10}$ and $\frac{6}{10}$ are equal because of symmetry. (Increasing values of t, i.e. of j, are shown moving downwards for convenience of calculation.) As examples,

$$u_{5,1}=\tfrac{1}{10}\{0.8+(8\times 1)+0.8\}=0.9600.$$

$$u_{4,2}=\tfrac{1}{10}\{0.6+(8\times 0.8)+0.96\}=0.7960.$$

The analytical solution of the partial differential equation satisfying these conditions is

$$U=\frac{8}{\pi^2}\sum_{n=1}^{\infty}\frac{1}{n^2}(\sin\tfrac{1}{2}n\pi)(\sin n\pi x)\exp(-n^2\pi^2 t).$$

TABLE 2.2

	$i=0$ $x=0$	$i=1$ 0.1	$i=2$ 0.2	$i=3$ 0.3	$i=4$ 0.4	$i=5$ 0.5	$i=6$ 0.6
$(j=0)\,t=0.000$	0	0.2000	0.4000	0.6000	0.8000	1.0000	0.8000
$(j=1)$ 0.001	0	0.2000	0.4000	0.6000	0.8000	0.9600	0.8000
$(j=2)$ 0.002	0	0.2000	0.4000	0.6000	0.7960	0.9280	0.7960
$(j=3)$ 0.003	0	0.2000	0.4000	0.5996	0.7896	0.9016	0.7896
$(j=4)$ 0.004	0	0.2000	0.4000	0.5986	0.7818	0.8792	0.7818
$(j=5)$ 0.005	0	0.2000	0.3999	0.5971	0.7732	0.8597	0.7732
⋮							
$(j=10)$ 0.01	0	0.1996	0.3968	0.5822	0.7281	0.7867	0.7281
⋮							
$(j=20)$ 0.02	0	0.1938	0.3781	0.5373	0.6486	0.6891	0.6486

TABLE 2.3

	Finite-difference solution ($x = 0.3$)	Analytical solution ($x = 0.3$)	Difference	Percentage error
$t = 0.005$	0.5971	0.5966	0.0005	0.08
$t = 0.01$	0.5822	0.5799	0.0023	0.4
$t = 0.02$	0.5373	0.5334	0.0039	0.7
$t = 0.10$	0.2472	0.2444	0.0028	1.1

Comparison of this solution with the finite-difference one at $x = 0.3$, as given above, shows that the finite-difference solution is reasonably accurate. The percentage error is the difference of the solutions expressed as a percentage of the analytical solution of the partial differential equation.

The comparison at $x = 0.5$ is not quite so good because of the discontinuity in the initial value of $\partial U/\partial x$, from $+2$ to -2, at this point (eqn 2.5). Inspection of Table 2.4 shows, however, that the effect of this discontinuity dies away as t increases.

It can be proved analytically that when the boundary values are constant the effect of discontinuities in initial values and initial derivatives upon the solution of a parabolic equation decreases as t increases.

An examination of Tables 2.19 and 2.21 given in Exercise 1 at the end of this chapter shows that the same finite-difference solution for a problem in which the initial function and all its derivatives are continuous is very close indeed to the solution of the partial differential equation.

Richtmyer, reference 25, has shown for this particular finite-difference scheme that when the initial function and its first $(p-1)$ derivatives are continuous and the pth derivative ordinar-

TABLE 2.4

	Finite-difference solution ($x = 0.5$)	Analytical solution ($x = 0.5$)	Difference	Percentage error
$t = 0.005$	0.8597	0.8404	0.0193	2.3
$t = 0.01$	0.7867	0.7743	0.0124	1.6
$t = 0.02$	0.6891	0.6809	0.0082	1.2
$t = 0.10$	0.3056	0.3021	0.0035	1.2

TABLE 2.5

	$i=0$ $x=0$	1 0.1	2 0.2	3 0.3	4 0.4	5 0.5	6 0.6
$T=0.000$	0	0.2000	0.4000	0.6000	0.8000	1.0000	0.8000
0.005	0	0.2000	0.4000	0.6000	0.8000	0.8000	0.8000
0.010	0	0.2000	0.4000	0.6000	0.7000	0.8000	0.7000
0.015	0	0.2000	0.4000	0.5500	0.7000	0.7000	0.7000
0.020	0	0.2000	0.3750	0.5500	0.6250	0.7000	0.6250
⋮							
0.100	0	0.0949	0.1717	0.2484	0.2778	0.3071	0.2778

ily discontinuous (i.e. changes by finite jumps), then the difference between the solution of the partial differential equation and a convergent solution of the difference equation is of order $(\delta t)^{(p+2)/(p+4)}$, for small δt.

In this example, $p=1$, so the difference is of order $(\delta t)^{\frac{3}{5}}$. As $(0.001)^{\frac{3}{5}}=0.016$, it is seen that the finite-difference solution is actually better than the estimate indicates, a feature common to most error estimates. When all the derivatives are continuous, $p \to \infty$, and the error is of order δt.

Case 2

Take $\delta x = h = \frac{1}{10}$, $\delta t = k = \frac{5}{1000}$, so $r = k/h^2 = 0.5$. Then eqn (2.4) gives

$$u_{i,j+1} = \tfrac{1}{2}(u_{i-1,j} + u_{i+1,j}), \tag{2.7}$$

and the solution obtained by applying this finite-difference equation to the boundary and initial values is recorded in Table 2.5.

TABLE 2.6

	Finite-difference solution ($x=0.3$)	Analytical solution ($x=0.3$)	Difference	Percentage error
$t=0.005$	0.6000	0.5966	0.0034	0.57
$t=0.01$	0.6000	0.5799	0.0201	3.5
$t=0.02$	0.5500	0.5334	0.0166	3.1
$t=0.1$	0.2484	0.2444	0.0040	1.6

It is seen that this finite-difference solution is not quite as good an approximation to the solution of the partial differential equation as the previous one; nevertheless it would be adequate for most technical purposes.

Case 3

Take $\delta x = \frac{1}{10}$, $\delta t = \frac{1}{100}$, so $r = \delta t/(\delta x)^2 = 1$. Then eqn (2.4) gives

$$u_{i,j+1} = u_{i-1,j} - u_{i,j} + u_{i+1,j}, \tag{2.9}$$

and the solution of this finite-difference scheme is as below.

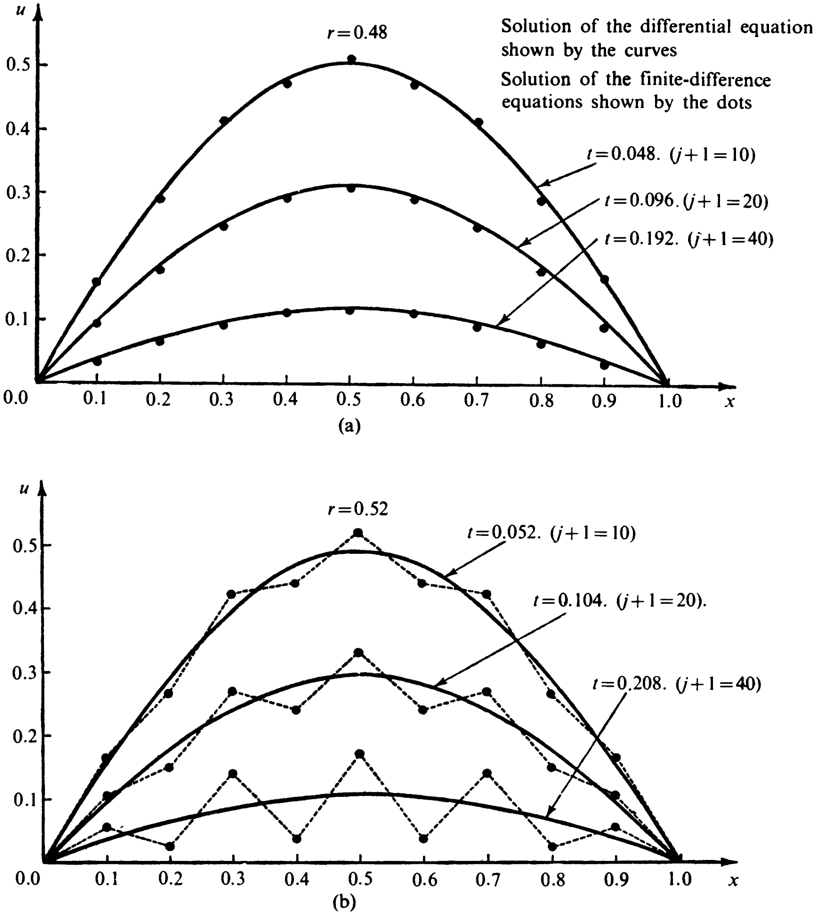
TABLE 2.7

	$i=0$ $x=0$	1 0.1	2 0.2	3 0.3	4 0.4	5 0.5	6 0.6
$t=0.00$	0	0.2	0.4	0.6	0.8	1.0	0.8
0.01	0	0.2	0.4	0.6	0.8	0.6	0.8
0.02	0	0.2	0.4	0.6	0.4	1.0	0.4
0.03	0	0.2	0.4	0.2	1.2	−0.2	1.2
0.04	0	0.2	0.0	1.4	−1.2	2.6	−1.2

Considered as a solution of the partial differential equation this is obviously meaningless, although it is, of course, the correct solution of eqn (2.9) with respect to the initial values and boundary values given.

These three cases clearly indicate that the value of r is important and it will be proved later that this explicit method is valid only for $0 < r \leqslant \frac{1}{2}$. The conditions that must be satisfied for a valid solution are dealt with both descriptively and analytically later in this chapter under the headings of convergence, stability, and consistency. Any reader who would prefer to have an introduction to these concepts at this stage could do so by reading the descriptive treatments of these topics as they are independent of the remainder of this chapter.

The graphs opposite compare the analytical solution of the partial differential equation (shown as continuous curves) with the finite-difference solution (shown by dots) for values of r just below and above $\frac{1}{2}$, and the same number of time-steps.



(a)

(b)

Fig. 2.3

Crank–Nicolson implicit method

Although the explicit method is computationally simple it has one serious drawback. The time step $\delta t = k$ is necessarily very small because the process is valid only for $0 < k/h^2 \leqslant \frac{1}{2}$, i.e. $k \leqslant \frac{1}{2}h^2$, and $h = \delta x$ must be kept small in order to attain reasonable accuracy. Crank and Nicolson (1947) proposed, and used, a method that reduces the total volume of calculation and is valid (i.e., convergent and stable) for all finite values of r. They considered the partial differential equation as being satisfied at the midpoint $\{ih, (j+\frac{1}{2})k\}$ and replaced $\partial^2U/\partial x^2$ by the mean of its finite-difference approximations at the jth and $(j+1)$th time-levels. In other words they approximated the equation

$$\left(\frac{\partial U}{\partial t}\right)_{i,j+\frac{1}{2}} = \left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j+\frac{1}{2}}$$

by

$$\frac{u_{i,j+1}-u_{i,j}}{k} = \frac{1}{2}\left\{\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{h^2} + \frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{h^2}\right\},$$

giving

$$-ru_{i-1,j+1}+(2+2r)u_{i,j+1}-ru_{i+1,j+1} = ru_{i-1,j}+(2-2r)u_{i,j}+ru_{i+1,j}, \tag{2.10}$$

where $r = k/h^2$.

In general, the left side of eqn (2.10) contains three unknown, and the right side three known, pivotal values of u (Fig. 2.4).

If there are N internal mesh points along each time row then for $j=0$ and $i=1, 2, \ldots, N$, eqn (2.10) gives N simultaneous equations for the N unknown pivotal values along the first

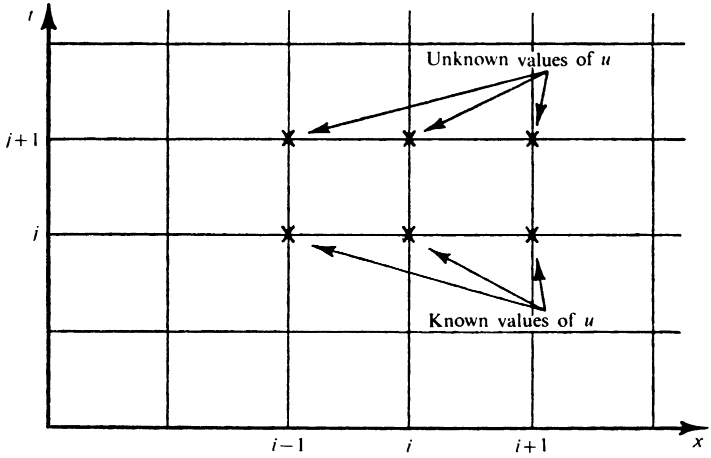


Fig. 2.4

time-row in terms of known initial and boundary values. Similarly, $j=1$ expresses N unknown values of u along the second time-row in terms of the calculated values along the first, etc. A method such as this where the calculation of an unknown pivotal value necessitates the solution of a set of simultaneous equations is described as an *implicit* one.

For readers familiar with finite-difference notation, the Crank–Nicolson method approximates the partial differential equation at the point $\{ih, (j+\frac{1}{2})k\}$ by

$$\frac{1}{k}\,\delta_t u_{i,j+\frac{1}{2}} = \frac{1}{2h^2}\left\{\delta_x^2 u_{i,j+1} + \delta_x^2 u_{i,j}\right\},$$

where the subscripts t and x denote differencing in the t- and x-directions respectively. Relative to the point $\{ih, (j+\frac{1}{2})k\}$, both $\partial U/\partial t$ and $\partial^2 U/\partial x^2$ have been replaced by central-difference approximations. This tends to reduce the errors introduced by the approximations.

Example 2.2

Use the Crank–Nicolson method to calculate a numerical solution of the previous worked example, namely,

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0<x<1, t>0,$$

where (i) $U=0$, $x=0$ and 1, $t\geqslant 0$,
(ii) $U=2x$, $0\leqslant x\leqslant\frac{1}{2}$, $t=0$,
(iii) $U=2(1-x)$, $\frac{1}{2}\leqslant x\leqslant 1$, $t=0$.

Take $h=\frac{1}{10}$. Although the method is valid for all finite values of $r=k/h^2$, a large value will yield an inaccurate approximation for $\partial U/\partial t$. A suitable value is $r=1$ and has the advantage of making the coefficient of $u_{i,j}$ zero in (2.10). Then $k=\frac{1}{100}$ and (2.10) reads as

$$-u_{i-1,j+1} + 4u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j}. \tag{2.11}$$

The computational molecule corresponding to eqn (2.11) is shown in Fig. 2.5. Denote $u_{i,j+1}$ by u_i $(i=1, 2, \ldots, 9)$. For this problem, because of symmetry, $u_6=u_4$, $u_7=u_3$, etc. (Fig. 2.6).

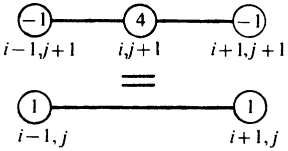


Fig. 2.5

The values of u for the first time step then satisfy

$$
\begin{aligned}
-0+4u_1-u_2&=0\quad+0.4,\\
-u_1+4u_2-u_3&=0.2+0.6,\\
-u_2\;+4u_3-u_4&=0.4\;+0.8,\\
-u_3+4u_4-u_5&=0.6+1.0,\\
-2u_4+4u_5\qquad&=0.8+0.8.
\end{aligned}
$$

As indicated in the next section these are easily solved by systematic eliminations to give

$$u_1=0.1989,\ u_2=0.3956,\ u_3=0.5834,\ u_4=0.7381,\ u_5=0.7691.$$

Hence the equations for the pivotal values of u along the next time row are

$$
\begin{aligned}
-0+4u_1-u_2&=0+0.3956,\\
-u_1+4u_2-u_3&=0.1989+0.5834,\\
-u_2+4u_3-u_4&=0.3956+0.7381,\\
-u_3+4u_4-u_5&=0.5834+0.7691,\\
-2u_4+4u_5\qquad&=2\times0.7381
\end{aligned}
$$

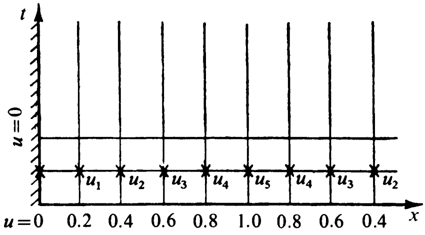


Fig. 2.6

TABLE 2.8

	$x=0$	0.1	0.2	0.3	0.4	0.5
$t=0.00$	0	0.2	0.4	0.6	0.8	1.0
$t=0.01$	0	0.1989	0.3956	0.5834	0.7381	0.7691
$t=0.02$	0	0.1936	0.3789	0.5400	0.6461	0.6921
.	0					
.	.					
.	.					
$t=0.10$	0	0.0948	0.1803	0.2482	0.2918	0.3069
Analytical solution $t=0.10$	0	0.0934	0.1776	0.2444	0.2873	0.3021

The solution of these equations is given in Table 2.8 together with figures comparing the finite-difference solution at $t=0.1$ with the solution of the partial differential equation. The numerical solution is clearly a good one.

Table 2.9 below displays both solutions at $x=0.5$ for various values of t. A glance at Table 2.4 shows that in this example the accuracy of this implicit method over the time-range taken is about the same as for the explicit method which uses ten times as many time-steps.

As mentioned previously the greatest difference between the two solutions occurs at $x=0.5$ because of the ordinary discontinuity in the initial value of $\partial U/\partial x$ at this point. A glance at Table 2.25 in Exercise 3 at the end of the chapter shows that this difference is less for an initial function that is continuous together with its derivatives.

Although the Crank–Nicolson method for $\partial U/\partial t=\partial^2 U/\partial x^2$ is stable for all positive values of r in the sense that the solution and all errors eventually tend to zero as j tends to infinity, it will be shown later that large values for r, such as 40, can introduce

TABLE 2.9

	Finite-difference solution ($x=0.5$)	Analytical solution ($x=0.5$)	Difference	Percentage error
$t=0.01$	0.7691	0.7743	−0.0052	−0.7
$t=0.02$	0.6921	0.6809	+0.0112	+1.6
$t=0.10$	0.3069	0.3021	0.0048	1.6

unwanted finite oscillations into the numerical solution. Such oscillations die away only very slowly with increasing j, and usually occur in the x-neighbourhood of points of discontinuity in the initial values or between initial values and boundary values.

Solution of the equations by Gauss's elimination method (without pivoting)

When there are $N-1$ internal mesh points along each time row the Crank–Nicolson equations (2.10) can be written very generally as

$$\begin{aligned}
&+b_1u_1-c_1u_2 &&= d_1,\\
&-a_2u_1+b_2u_2-c_2u_3 &&= d_2,\\
&\quad\cdot\qquad\cdot\qquad\cdot &&\\
&\qquad\quad\cdot\qquad\cdot\qquad\cdot &&\\
&\qquad\qquad\qquad -a_iu_{i-1}+b_iu_i-c_iu_{i+1} &&= d_i,\\
&\qquad\qquad\qquad\qquad\cdot\qquad\cdot\qquad\cdot &&\\
&\qquad\qquad\qquad\qquad -a_{N-1}u_{N-2}+b_{N-1}u_{N-1} &&= d_{N-1},
\end{aligned}$$

where the a's, b's, c's, and d's are known. The first equation can be used to eliminate u_1 from the second equation, the new second equation used to eliminate u_2 from the third equation and so on, until finally, the new last but one equation can be used to eliminate u_{N-2} from the last equation, giving one equation with only one unknown, u_{N-1}. The unknowns $u_{N-2}, u_{N-3}, \ldots u_2, u_1$ can then be found in turn by back-substitution. Noting that the coefficient c in each new equation is the same as in the corresponding old equation, assume that the following stage of the eliminations has been reached,

$$\begin{aligned}
\alpha_{i-1}u_{i-1}-c_{i-1}u_i &= S_{i-1},\\
-a_iu_{i-1}+b_iu_i-c_iu_{i+1} &= d_i,
\end{aligned}$$

where $\alpha_1=b_1$, $S_1=d_1$.

Eliminating u_{i-1} leads to

$$\left(b_i-\frac{a_ic_{i-1}}{\alpha_{i-1}}\right)u_i-c_iu_{i+1}=d_i+\frac{a_iS_{i-1}}{\alpha_{i-1}},$$

i.e.

$$\alpha_iu_i-c_iu_{i+1}=S_i, \tag{2.12}$$

where

$$\alpha_i = b_i - \frac{a_i c_{i-1}}{\alpha_{i-1}} \quad \text{and} \quad S_i = d_i + \frac{a_i S_{i-1}}{\alpha_{i-1}} \quad (i = 2, 3, \ldots).$$

The last pair of simultaneous equations are

$$\alpha_{N-2} u_{N-2} - c_{N-2} u_{N-1} = S_{N-2}$$

and

$$-a_{N-1} u_{N-2} + b_{N-1} u_{N-1} = d_{N-1}.$$

Elimination of u_{N-2} gives

$$\left(b_{N-1} - \frac{a_{N-1} c_{N-2}}{\alpha_{N-2}}\right) u_{N-1} = d_{N-1} + \frac{a_{N-1} S_{N-2}}{\alpha_{N-2}},$$

i.e.

$$\alpha_{N-1} u_{N-1} = S_{N-1}, \tag{2.13}$$

Equations (2.12) and (2.13) show that the solution can be calculated from

$$u_{N-1} = \frac{S_{N-1}}{\alpha_{N-1}},$$

$$u_i = \frac{1}{\alpha_i}(S_i + c_i u_{i+1}) \quad (i = N-2, N-3, \ldots, 1),$$

where the α's and S's are given recursively by

$$\alpha_1 = b_1; \quad \alpha_i = b_i - \frac{a_i}{\alpha_{i-1}} c_{i-1},$$

$$S_1 = d_1; \quad S_i = d_i + \frac{a_i}{\alpha_{i-1}} S_{i-1} \quad (i = 2, 3, \ldots, N-1).$$

In many problems α_i and a_i/α_{i-1} are independent of time and need only be calculated once, irrespective of the number of time-steps.

As an illustration consider the last worked example for which the equations were

$$4u_1 - u_2 = 0.4,$$

$$-u_1 + 4u_2 - u_3 = 0.8,$$

$$-u_2 + 4u_3 - u_4 = 1.2,$$

$$-u_3 + 4u_4 - u_5 = 1.6,$$

$$-2u_4 + 4u_5 = 1.6.$$

Hence

$$a_2=a_3=a_4=1,\ a_5=2,\ b_1=b_2=b_3=b_4=b_5=4;$$

$$c_1=c_2=c_3=c_4=1;\ d_1=0.4,\ d_2=0.8,\ d_3=1.2,\ d_4=d_5=1.6,$$

so

$$\alpha_1=b_1=4;\quad \alpha_i=b_i-\frac{a_i}{\alpha_{i-1}}c_{i-1}=4-\frac{a_i}{\alpha_{i-1}}\quad (i=2,3,4,5),$$

giving the following coefficients which are invariant for every time-step.

$$\alpha_1=4,$$

$$\frac{a_2}{\alpha_1}=\frac{1}{4}=0.25,\qquad \alpha_2=4-\frac{a_2}{\alpha_1}=3.75,$$

$$\frac{a_3}{\alpha_2}=\frac{1}{3.75}=0.2667,\qquad \alpha_3=4-\frac{a_3}{\alpha_2}=3.7333,$$

$$\frac{a_4}{\alpha_3}=\frac{1}{3.7333}=0.2679,\qquad \alpha_4=4-\frac{a_4}{\alpha_3}=3.7321,$$

$$\frac{a_5}{\alpha_4}=\frac{2}{3.7321}=0.5359,\qquad \alpha_5=4-\frac{a_5}{\alpha_4}=3.4641.$$

As

$$S_1=d_1=0.4\quad\text{and}\quad S_i=d_i+\frac{a_i}{\alpha_{i-1}}S_{i-1}\quad (i=2,3,4,5),$$

$$S_1=0.4,$$

$$S_2=0.8+\frac{a_2}{\alpha_1}S_1=0.8+(0.25)(0.4)=0.9,$$

$$S_3=1.2+\frac{a_3}{\alpha_2}S_2=1.4400,$$

$$S_4=1.6+\frac{a_4}{\alpha_3}S_3=1.9858,$$

$$S_5=1.6+\frac{a_5}{\alpha_4}S_4=2.6642,$$

and the solution for the first time-step is

$$u_5=\frac{S_5}{\alpha_5}=0.7691,$$

$$u_4=\frac{1}{\alpha_4}(S_4+c_4u_5)=0.7381,$$

$$u_3=\frac{1}{\alpha_3}(S_3+c_3u_4)=0.5834,$$

$$u_2=\frac{1}{\alpha_2}(S_2+c_2u_3)=0.3956,$$

$$u_1=\frac{1}{\alpha_1}(S_1+c_1u_2)=0.1989.$$

A comment on the stability of the elimination method

The non-pivoting elimination method previously described for solving the set of linear equations $\mathbf{Au}=\mathbf{d}$, with a tridiagonal matrix $\mathbf{A}$, is always stable, that is, with no growth of rounding errors, if

(i) $a_i>0$, $b_i>0$ and $c_i>0$,
(ii) $b_i>a_{i+1}+c_{i-1}$ for $i=1,2,\ldots,N-1$, defining $c_0=a_N=0$, and
(iii) $b_i>a_i+c_i$ for $i=1,2,\ldots,N-1$, defining $a_1=c_{N-1}=0$.

Conditions (i) and (ii), which ensure that the forward elimination is stable, state that the diagonal element must exceed the sum of the moduli of the other elements in the same column of the matrix $\mathbf{A}$ of coefficients. Conditions (i) and (iii), which ensure that the back substitution is stable, state that the diagonal element must exceed the sum of the moduli of the other elements in the same row. When these conditions are satisfied the algorithm is a very efficient one for programming on a digital computer, using a minimum of storage space.

Proof

To prove that the forward elimination procedure is stable it is necessary to show that the moduli of the multipliers $m_i=a_i/\alpha_{i-1}$

used to eliminate $u_1, u_2, \ldots,$ are $\leqslant 1$. By p. 25.

$$\alpha_i = b_i - \frac{a_i c_{i-1}}{\alpha_{i-1}} = b_i - m_i c_{i-1},$$

Therefore,

$$m_{i+1} = \frac{a_{i+1}}{\alpha_i} = \frac{a_{i+1}}{b_i - m_i c_{i-1}}.$$

Hence,

$$0 < m_2 < \frac{a_2}{b_1} < 1 \quad \text{since} \quad b_1 > a_2 > 0 = c_0.$$

Similarly,

$$\begin{aligned} 0 < m_3 &= \frac{a_3}{b_2 - m_2 c_1} \quad \text{since} \quad a_3 > 0, b_2 > c_1 \text{ and } 0 < m_2 < 1, \\ &< \frac{a_3}{b_2 - c_1} \quad \text{since} \quad c_1 > 0, \\ &< \frac{a_3}{(a_3 + c_1) - c_1} = 1 \quad \text{since} \quad b_2 > a_3 + c_1. \end{aligned}$$

In this way, $0 < m_4, m_5, \ldots, m_{N-1} < 1$. The stability of the back substitution is proved in Exercise 4, Chapter 2.

A weighted average approximation

A more general finite-difference approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$ than those considered is given by

$$\frac{u_{i,j+1} - u_{i,j}}{\delta t} = \frac{1}{(\delta x)^2}\{\theta(u_{i+1,j+1} - 2u_{i,j+1} + u_{i-1,j+1}) + (1-\theta) \times (u_{i+1,j} - 2u_{i,j} + u_{i-1,j})\},$$

where, in practice, $0 \leqslant \theta \leqslant 1$. For readers familiar with finite-difference notation this replacement approximates the partial differential equation at the point $\{i\delta x, (j+\frac{1}{2})\delta t\}$ by the difference equation

$$\frac{1}{\delta t}\delta_t u_{i,j+\frac{1}{2}} = \frac{1}{(\delta x)^2}\{\theta\delta_x^2 u_{i,j+1} + (1-\theta)\delta_x^2 u_{i,j}\},$$

where the subscripts t and x denote differencing in the t- and x-directions respectively. $\theta=0$ gives the explicit scheme, $\theta=\frac{1}{2}$ the Crank–Nicolson, and $\theta=1$ a fully implicit backward time-difference method. The equations are unconditionally valid, i.e. stable and convergent for $\frac{1}{2}\leqslant\theta\leqslant 1$, but for $0\leqslant\theta<\frac{1}{2}$ we must have

$$r=\frac{\delta t}{(\delta x)^2}\leqslant\frac{1}{2(1-2\theta)}. \quad \text{(See Exercise 20, Chapter 2.)}$$

Derivative boundary conditions

Boundary conditions expressed in terms of derivatives occur very frequently in practice. When, for example, the surface of a heat-conducting material is thermally insulated, there is no heat flow normal to the surface and the corresponding boundary condition is $\partial U/\partial n=0$ at every point of the insulated surface, where the differentiation of the temperature U is in the direction of the normal to the surface. Similarly, the rate at which heat is transferred by radiation from an external surface at temperature U into a surrounding medium at temperature v is often assumed to be proportional to $(U-v)$. As the fundamental assumption of heat-conduction theory is that the rate of flow across any surface is equal to $-K\partial U/\partial n$ units of heat per unit time in the direction of the outward normal, the corresponding boundary condition for surface radiation is

$$-K\frac{\partial U}{\partial n}=H(U-v).$$

The constant K is the thermal conductivity of the material and the constant H its coefficient of surface heat transfer. The negative sign indicates that heat is assumed to flow in the opposite direction to that in which U increases algebraically. This equation can be written as

$$\frac{\partial U}{\partial n}=-h(U-v),$$

where h is a positive constant.

Consider a thin rod that is thermally insulated along its length and which radiates heat from the end $x=0$. The temperature at this end at time t is now unknown and its determination requires an extra equation. This equation can be the boundary condition itself when a forward difference is used for $\partial U/\partial x$, because the boundary condition at $x=0$, the left-hand end, namely,

$$-\frac{\partial U}{\partial x}=-h(U-v),$$

will be represented by

$$\frac{u_{1,j}-u_{0,j}}{\delta x}=h(u_{0,j}-v),$$

giving one extra equation for the temperature $u_{0,j}$. A negative sign must be associated with $\partial U/\partial x$ because the outward normal to the rod at this end is in the negative direction of the x-axis. Alternatively, in the heat-flow law, $-K\partial U/\partial n$ implies that when the positive direction of the x-axis (and of U) is to the *right*, then the quantity of heat flowing from *right to left* across unit area per unit time is $+K\partial U/\partial x$, and this is proportional to the excess temperature at $x=0$.

If we wish to represent $\partial U/\partial x$ more accurately at $x=0$ by a central difference formula it is necessary to introduce the 'fictitious' temperature $u_{-1,j}$ at the external mesh point $(-\delta x, j\delta t)$ (Fig. 2.7), by imagining the rod to be extended a distance δx at this end. The boundary condition can then be represented by

$$\frac{u_{1,j}-u_{-1,j}}{2\delta x}=h(u_{0,j}-v).$$

The temperature $u_{-1,j}$ is unknown and necessitates another equation. This is obtained by assuming that the heat conduction

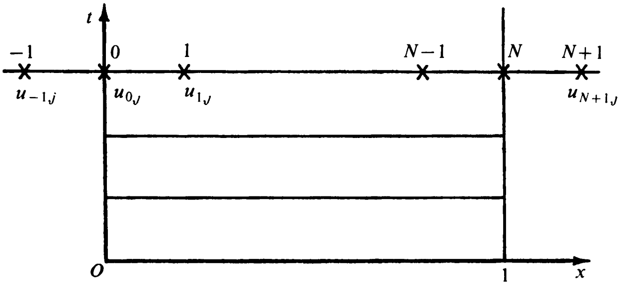


Fig. 2.7

equation is satisfied at the end $x = 0$ of the rod. The unknown $u_{-1,j}$ can then be eliminated between these equations. Similar equations can be written down for radiation from the other end of the rod.

These methods are applied below to the problem of the cooling of a homogeneous rod by radiation from its ends into air at a constant temperature, the rod being at a different constant temperature initially.

Example 2.3

Solve the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \tag{2.14}$$

satisfying the initial condition,

$$U = 1 \text{ for } 0 \leqslant x \leqslant 1 \text{ when } t = 0,$$

and the boundary conditions,

$$\frac{\partial U}{\partial x} = U \text{ at } x = 0, \text{ for all } t,$$

$$\frac{\partial U}{\partial x} = -U \text{ at } x = 1, \text{ for all } t,$$

using an explicit method and employing central-differences for the boundary conditions.

One explicit finite-difference representation of eqn (2.14) is

$$\frac{u_{i,j+1} - u_{i,j}}{\delta t} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{(\delta x)^2},$$

i.e.

$$u_{i,j+1} = u_{i,j} + r(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}), \tag{2.15}$$

where $r = \delta t/(\delta x)^2$.

At $x = 0$,

$$u_{0,j+1} = u_{0,j} + r(u_{-1,j} - 2u_{0,j} + u_{1,j}). \tag{2.16}$$

The boundary condition at $x = 0$, in terms of central-differences, can be written as

$$\frac{u_{1,j} - u_{-1,j}}{2\delta x} = u_{0,j}. \tag{2.17}$$

Eliminating $u_{-1,j}$ between (2.16) and (2.17) gives

$$u_{0,j+1}=u_{0,j}+2r\{u_{1,j}-(1+\delta x)u_{0,j}\}. \tag{2.18}$$

Let $\delta x=0.1$. Then at $x=1$, eqn (2.15) becomes

$$u_{10,j+1}=u_{10,j}+r(u_{9,j}-2u_{10,j}+u_{11,j}), \tag{2.19}$$

and the boundary condition is

$$\frac{u_{11,j}-u_{9,j}}{2\delta x}=-u_{10,j}. \tag{2.20}$$

Elimination of the 'fictitious' value $u_{11,j}$ between (2.19) and (2.20) yields

$$u_{10,j+1}=u_{10,j}+2r\{u_{9,j}-(1+\delta x)u_{10,j}\}. \tag{2.21}$$

This result could have been deduced from the corresponding equation at $x=0$ because of the symmetry with respect to $x=\frac{1}{2}$.

Later in this chapter this scheme is proved to be valid for $r\leqslant 1/(2+h\delta x)$, i.e. $r\leqslant 1/2.1$ in this example.

Choose $r=\frac{1}{4}$. The difference equations (2.18), (2.15), then become

$$u_{0,j+1}=\tfrac{1}{2}(0.9u_{0,j}+u_{1,j}),$$

$$u_{i,j+1}=\tfrac{1}{4}(u_{i-1,j}+2u_{i,j}+u_{i+1,j}) \quad (i=1,2,3,4),$$

and the use of symmetry rather than eqn (2.21) gives

$$u_{5,j+1}=\tfrac{1}{4}(2u_{4,j}+2u_{5,j}).$$

As the initial temperature is $u=1$, the values of u at the end of the first time-step when $t=r(\delta x)^2=\frac{1}{400}$, are

$$u_{0,1}=\tfrac{1}{2}(0.9+1)=0.95,$$

$$u_{1,1}=\tfrac{1}{4}(1+2+1)=1=u_{2,1}=u_{3,1}=u_{4,1}=u_{5,1},$$

and the values at the end of the second time-step are

$$u_{0,2}=\tfrac{1}{2}(0.9\times 0.95+1)=0.9275,$$

$$u_{1,2}=\tfrac{1}{4}(0.95+2+1)=0.9875,$$

$$u_{2,2}=\tfrac{1}{4}(1+2+1)=1=u_{3,2}=u_{4,2}=u_{5,2}.$$

Similarly for subsequent time-steps. The values for several steps are recorded in Table 2.10.

The analytical solution of the partial differential equation satis-

TABLE 2.10

	$i=0$	1	2	3	4	5
	$x=0$	0.1	0.2	0.3	0.4	0.5
$t=0.0000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.0025	0.9500	1.0000	1.0000	1.0000	1.0000	1.0000
$t=0.005$	0.9275	0.9875	1.0000	1.0000	1.0000	1.0000
0.0075	0.9111	0.9756	0.9969	1.0000	1.0000	1.0000
0.0100	0.8978	0.9648	0.9923	0.9992	1.0000	1.0000
0.0125	0.8864	0.9549	0.9872	0.9977	0.9998	1.0000
0.0150	0.8764	0.9459	0.9818	0.9956	0.9993	0.9999
0.0175	0.8673	0.9375	0.9762	0.9931	0.9985	0.9996
0.0200	0.8590	0.9296	0.9708	0.9902	0.9974	0.9991
....						
....						
0.1000	0.7175	0.7829	0.8345	0.8718	0.8942	0.9017
0.2500	0.5542	0.6048	0.6452	0.6745	0.6923	0.6983
0.5000	0.3612	0.3942	0.4205	0.4396	0.4512	0.4551
1.0000	0.1534	0.1674	0.1786	0.1867	0.1917	0.1933

fying these boundary and initial conditions is

$$U=4\sum_{n=1}^{\infty}\left\{\frac{\sec\alpha_n}{(3+4\alpha_n^2)}e^{-4\alpha_n{}^2t}\cos 2\alpha_n(x-\tfrac{1}{2})\right\}\quad(0<x<1),$$

where α_n are the positive roots of

$$\alpha\tan\alpha=\tfrac{1}{2}.$$

Values of U calculated from this analytical solution are recorded in Table 2.11.

The two solutions are compared at $x=0{\cdot}2$ in Table 2.12.

The finite-difference solution is clearly very accurate for this small value of r.

Because of the symmetry with respect to $x=\frac{1}{2}$ the solution above is the same for a rod of length $\frac{1}{2}$, thermally insulated along its length and at $x=\frac{1}{2}$, and which cools by radiation from $x=0$ into a medium at zero temperature.

Example 2.4

Re-solve the Worked Example 2.3 using an explicit method and employing a forward-difference for the boundary condition at $x=0$.

TABLE 2.11

$x =$	0	0.1	0.2	0.3	0.4	0.5
t						
0.0025	0.9460	0.9951	0.9999	1.0000	1.0000	1.0000
0.0050	0.9250	0.9841	0.9984	0.9999	1.0000	1.0000
0.0075	0.9093	0.9730	0.9950	0.9994	1.0000	1.0000
0.0100	0.8965	0.9627	0.9905	0.9984	0.9998	1.0000
0.0125	0.8854	0.9532	0.9855	0.9967	0.9994	0.9999
0.0150	0.8755	0.9444	0.9802	0.9945	0.9988	0.9996
0.0175	0.8666	0.9362	0.9748	0.9919	0.9979	0.9992
0.0200	0.8585	0.9286	0.9695	0.9891	0.9967	0.9985
....						
....						
0.1000	0.7176	0.7828	0.8342	0.8713	0.8936	0.9010
0.2500	0.5546	0.6052	0.6454	0.6747	0.6924	0.6984
0.5000	0.3619	0.3949	0.4212	0.4403	0.4519	0.4558
1.0000	0.1542	0.1682	0.1794	0.1875	0.1925	0.1941

TABLE 2.12

	Finite-difference solution ($x = 0.2$)	Analytical solution ($x = 0.2$)	Percentage error
$t = 0.005$	1.0000	0.9984	0.16
0.050	0.9126	0.9120	0.07
0.100	0.8345	0.8342	0.04
0.250	0.6452	0.6454	−0.03
0.500	0.4205	0.4212	−0.16
1.000	0.1786	0.1794	−0.45

By eqn (2.15), one explicit finite-difference representation of the partial differential equation is

$$u_{i,j+1} = u_{i,j} + r(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}).$$

Hence, for $i = 1$,

$$u_{1,j+1} = u_{1,j} + r(u_{0,j} - 2u_{1,j} + u_{2,j}). \tag{2.22}$$

The boundary condition at $x = 0$, namely $\partial U/\partial x = U$, in terms of a forward difference is

$$\frac{u_{1,j} - u_{0,j}}{\delta x} = u_{0,j},$$

so

$$u_{0,j}=\frac{u_{1,j}}{1+\delta x}. \tag{2.23}$$

Eliminating $u_{0,j}$ between (2.22) and (2.23) gives

$$u_{1,j+1}=\left(1-2r+\frac{r}{1+\delta x}\right)u_{1,j}+ru_{2,j}. \tag{2.24}$$

This scheme is valid for $0<r\leqslant\frac{1}{2}$ (see Exercise 24, Chapter 2), but in order to compare its solution with the previous one, put $r=\frac{1}{4}$ and $\delta x=0.1$. The relevant equations are then

$$u_{1,j+1}=\tfrac{8}{11}u_{1,j}+\tfrac{1}{4}u_{2,j},$$

$$u_{0,j+1}=\tfrac{10}{11}u_{1,j+1},$$

$$u_{i,j+1}=\tfrac{1}{4}(u_{i-1,j}+2u_{i,j}+u_{i+1,j})\quad(i=2,3,4),$$

and

$$u_{5,j+1}=\tfrac{1}{4}(2u_{4,j}+2u_{5,j}),\text{ because of symmetry.}$$

The solution of these equations for an initial value of $U=1$ is shown in Table 2.13. A comparison with the analytical solution at $x=0.2$ is given in Table 2.14.

TABLE 2.13

$x=$	0	0.1	0.2	0.3	0.4	0.5
$t=0.0000$	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
0.0025	0.8884	0.9773	1.0000	1.0000	1.0000	1.0000
0.0050	0.8734	0.9607	0.9943	1.0000	1.0000	1.0000
0.0075	0.8612	0.9473	0.9873	0.9986	1.0000	1.0000
0.0100	0.8507	0.9358	0.9801	0.9961	0.9996	1.0000
0.0125	0.8415	0.9256	0.9730	0.9930	0.9989	0.9998
0.0150	0.8331	0.9164	0.9662	0.9895	0.9976	0.9993
0.0175	0.8255	0.9080	0.9596	0.9857	0.9960	0.9985
0.0200	0.8184	0.9003	0.9532	0.9817	0.9941	0.9973
....						
....						
0.1000	0.6869	0.7556	0.8102	0.8498	0.8738	0.8818
0.2500	0.5206	0.5727	0.6142	0.6444	0.6628	0.6689
0.5000	0.3283	0.3611	0.3873	0.4063	0.4179	0.4218
1.0000	0.1305	0.1435	0.1540	0.1615	0.1661	0.1677

TABLE 2.14

	Finite-difference solution ($x=0.2$)	Analytical solution ($x=0.2$)	Percentage error
$t=0.005$	0.9943	0.9984	−0.4
0.050	0.8912	0.9120	−2.3
0.100	0.8102	0.8342	−2.9
0.250	0.6142	0.6454	−4.8
0.500	0.3873	0.4212	−8.0
1.000	0.1540	0.1794	−14.2

Although this solution is not as good as the previous one it is still sufficiently accurate for many practical purposes.

Example 2.5

Solve Example 2.3 by the Crank–Nicolson method.

This method represents $\partial U/\partial t = \partial^2 U/\partial x^2$ by

$$\frac{u_{i,j+1}-u_{i,j}}{\delta t}=\frac{1}{2}\left\{\frac{u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}}{(\delta x)^2}+\frac{u_{i+1,j}-2u_{i,j}+u_{i-1,j}}{(\delta x)^2}\right\},$$

which can be written as

$$-ru_{i-1,j+1}+(2+2r)u_{i,j+1}-ru_{i+1,j+1}=ru_{i-1,j}+(2-2r)u_{i,j}+ru_{i+1,j}. \tag{2.25}$$

The central difference representation of the boundary condition at $x=0$ is

$$\frac{u_{1,j}-u_{-1,j}}{2\delta x}=u_{0,j},$$

from which it follows that

$$u_{-1,j}=u_{1,j}-2\delta x u_{0,j}$$

and

$$u_{-1,j+1}=u_{1,j+1}-2\delta x u_{0,j+1}.$$

The last two equations enable us to eliminate $u_{-1,j}$ and $u_{-1,j+1}$ from the equation obtained by putting $i=0$ in (2.25).

The boundary condition at $x=1$ can be dealt with in the same way although in this problem it is easier to make use of the symmetry with respect to $x=\frac{1}{2}$, namely, $u_{6,j}=u_{4,j}$.

This scheme is formally valid for all finite values of r but we must keep it reasonably small if we want a close approximation to the solution of the partial differential equation. Choose $r=1$ and $\delta x=0.1$. A small amount of algebra soon shows that the equations for the unknown pivotal values $u_{0,j+1}, u_{1,j+1}, \ldots, u_{5,j+1}$ are

$$2.1u_{0,j+1}-u_{1,j+1}=-0.1u_{0,j}+u_{1,j},$$

$$-u_{i-1,j+1}+4u_{i,j+1}-u_{i+1,j+1}=u_{i-1,j}+u_{i+1,j} \quad (i=1,2,3,4),$$

$$-u_{4,j+1}+2u_{5,j+1}=u_{4,j}.$$

For the first time-step these give

$$\begin{aligned} 2.1u_0-u_1&=0.9,\\ -u_0+4u_1-u_2&=2.0,\\ -u_1+4u_2-u_3&=2.0,\\ -u_2+4u_3-u_4&=2.0,\\ -u_3+4u_4-u_5&=2.0,\\ -u_4+2u_5&=1.0. \end{aligned}$$

These can be solved by the direct elimination method previously described. Table 2.15 records the solution for several time-steps and Table 2.16 compares it with the analytical solution at $x=0.2$.

TABLE 2.15

$i=$	0	1	2	3	4	5
$t=0.00$	1.0	1.0	1.0	1.0	1.0	1.0
0.01	0.8908	0.9707	0.9922	0.9979	0.9994	0.9997
0.02	0.8624	0.9293	0.9720	0.9900	0.9964	0.9979
0.10	0.7179	0.7834	0.8349	0.8720	0.8944	0.9018
0.25	0.5547	0.6054	0.6458	0.6751	0.6929	0.6989
0.50	0.3618	0.3949	0.4212	0.4404	0.4520	0.4559
1.00	0.1540	0.1680	0.1793	0.1874	0.1923	0.1940

TABLE 2.16

	Finite-difference solution ($x = 0.2$)	Analytical solution ($x = 0.2$)	Percentage error
$t = 0.01$	0.9922	0.9905	0.17
0.05	0.9131	0.9120	0.12
0.10	0.8349	0.8342	0.08
0.25	0.6458	0.6454	0.06
0.50	0.4212	0.4212	0.00
1.00	0.1793	0.1794	−0.06

The local truncation error and consistency

The local truncation error

Let $F_{i,j}(u) = 0$ represent the difference equation approximating the partial differential equation at the (i, j)th mesh point, with exact solution u. If u is replaced by U at the mesh points of the difference equation, where U is the exact solution of the partial differential equation, the value of $F_{i,j}(U)$ is called *the local truncation error* $T_{i,j}$ at the (i, j) mesh point. $F_{i,j}(U)$ clearly measures the amount by which the exact solution values of the partial differential equation at the mesh points of the difference equation do *not* satisfy the difference equation at the point (ih, jk).

Using Taylor expansions, it is easy to express $T_{i,j}$ in terms of powers of h and k and partial derivatives of U at (ih, jk). Although U and its derivatives are generally unknown, the analysis is worthwhile because it provides a method for comparing the local accuracies of different difference schemes approximating the partial differential equation.

Example 2.6

Calculate the order of the local truncation error of the classical explicit difference approximation to

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$$

at the point (ih, jk).

$$F_{i,j}(u)=\frac{u_{i,j+1}-u_{i,j}}{k}-\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}=0$$

Therefore,

$$T_{i,j}=F_{i,j}(U)=\frac{U_{i,j+1}-U_{i,j}}{k}-\frac{U_{i-1,j}-2U_{i,j}+U_{i+1,j}}{h^2}.$$

By Taylor's expansion

$$\begin{aligned}U_{i+1,j}&=U\{(i+1)h,jk\}=U(x_i+h,\, t_j)\\&=U_{i,j}+h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\tfrac{1}{2}h^2\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j}+\tfrac{1}{6}h^3\left(\frac{\partial^3 U}{\partial x^3}\right)_{i,j}+\dots\end{aligned}$$

$$\begin{aligned}U_{i-1,j}&=U\{(i-1)h,\, jk\}=U(x_i-h,\, t_j)\\&=U_{i,j}-h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\tfrac{1}{2}h^2\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j}-\tfrac{1}{6}h^3\left(\frac{\partial^3 U}{\partial x^3}\right)_{i,j}+\dots\end{aligned}$$

$$\begin{aligned}U_{i,j+1}&=U(x_i,\, t_j+k)\\&=U_{i,j}+k\left(\frac{\partial U}{\partial t}\right)_{i,j}+\tfrac{1}{2}k^2\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j}+\tfrac{1}{6}k^3\left(\frac{\partial^3 U}{\partial t^3}\right)_{i,j}+\dots.\end{aligned}$$

Substitution into the expression for $T_{i,j}$ then gives

$$\begin{aligned}T_{i,j}=\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{i,j}+\tfrac{1}{2}k\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j}-\tfrac{1}{12}h^2\left(\frac{\partial^4 U}{\partial x^4}\right)_{i,j}\\+\tfrac{1}{6}k^2\left(\frac{\partial^3 U}{\partial t^3}\right)_{i,j}-\tfrac{1}{360}h^4\frac{\partial^6 U}{\partial x^6}+\dots.\end{aligned}$$

But U is the solution of the differential equation so

$$\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{i,j}=0.$$

Therefore the principal part of the local truncation error is

$$\left(\tfrac{1}{2}k\frac{\partial^2 U}{\partial t^2}-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}\right)_{i,j}.$$

Hence

$$T_{i,j}=O(k)+O(h^2).$$

When $k=rh^2$, $0<r\leqslant\frac{1}{2}$, $T_{i,j}$ is $O(k)$ or $O(h^2)$, as one would expect by eqns (1.8) and (1.10).

This error may be further reduced by choosing a special value for k/h^2 because the equation for $T_{i,j}$ can be written as

$$T_{i,j}=\tfrac{1}{12}h^2\left(6\frac{k}{h^2}\frac{\partial^2 U}{\partial t^2}-\frac{\partial^4 U}{\partial x^4}\right)_{i,j}+O(k^2)+O(h^4).$$

By the differential equation,

$$\frac{\partial}{\partial t}=\frac{\partial^2}{\partial x^2},$$

so

$$\frac{\partial}{\partial t}\left(\frac{\partial U}{\partial t}\right)=\frac{\partial^2}{\partial x^2}\left(\frac{\partial^2 U}{\partial x^2}\right),$$

assuming that these derivatives exist. If we put $6k/h^2=1$, the expression in the brackets is then zero and $T_{i,j}$ is $O(k^2)+O(h^4)$. This is of little use in practice because $k=\frac{1}{6}h^2$ is very small for small h so the volume of arithmetic needed to advance the solution to a large time-level is substantial.

Consistency or compatibility

It is sometimes possible to approximate a parabolic or hyperbolic equation by a finite-difference scheme that is stable, (i.e. limits the amplification of all the components of the initial conditions), but which has a solution that converges to the solution of a different differential equation as the mesh lengths tend to zero. Such a difference scheme is said to be *inconsistent* or *incompatible* with the partial differential equation and an example is given in Worked Example 2.7.

The real importance of the concept of consistency lies in a theorem by Lax (reference 25), which states that if a linear finite-difference equation is consistent with a properly posed linear initial-value problem then stability guarantees convergence of u to U as the mesh lengths tend to zero. Consistency can be defined in either of two equivalent but slightly different ways.

The more general definition is as follows. Let $L(U)=0$ represent the partial differential equation in the independent variables x and t, with exact solution U.

Let $F(u)=0$ represent the approximating finite-difference equation with exact solution u.

Let v be a continuous function of x and t with a sufficient number of continuous derivatives to enable $L(v)$ to be evaluated at the point (ih, jk).

Then the truncation error $T_{i,j}(v)$ at the point (ih, jk) is defined by

$$T_{i,j}(v) = F_{i,j}(v) - L(v_{i,j}).$$

If $T_{i,j}(v) \to 0$ as $h \to 0$, $k \to 0$, the difference equation is said to be consistent or compatible with the partial differential equation. With this definition $T_{i,j}$ gives an indication of the error resulting from the replacement of $L(v_{i,j})$ by $F_{i,j}(v)$.

Most authors put $v = U$ because $L(U) = 0$. It then follows that

$$T_{i,j}(U) = F_{i,j}(U),$$

and the truncation error coincides with the local truncation error. The difference equation is then consistent if the limiting value of the local truncation error is zero as $h \to 0$, $k \to 0$. This is the definition that will be adopted in this book. By the Worked Example 2.6 it follows that the classical explicit approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$ is consistent with the differential equation.

Example 2.7

The equation

$$\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2} = 0$$

is approximated at the point (ih, jk) by the difference equation

$$\frac{u_{i,j+1} - u_{i,j-1}}{2k} - \frac{u_{i+1,j} - 2\{\theta u_{i,j+1} + (1-\theta)u_{i,j-1}\} + u_{i-1,j}}{h^2} = 0.$$

Show that the local truncation error at this point is

$$\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3} - \frac{h^2}{12}\frac{\partial^4 U}{\partial x^4} + (2\theta - 1)\frac{2k}{h^2}\frac{\partial U}{\partial t} + \frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2} + O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Discuss the consistency of this scheme with the partial differential equation when:

(i) $k = rh$ and (ii) $k = rh^2$,

where r is a positive constant and θ a variable parameter.

Expansion of the terms $U_{i,j+1}$, $U_{i,j-1}$, $U_{i+1,j}$, and $U_{i-1,j}$ about the point (ih, jk) by Taylor's series, as in Example 2.6, and substitution into

$$T_{i,j}=\frac{U_{i,j+1}-U_{i,j-1}}{2k}-\frac{U_{i+1,j}-2\{\theta U_{i,j+1}+(1-\theta)U_{i,j-1}\}+U_{i-1,j}}{h^2}$$

leads to

$$T_{i,j}=\left(\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right)_{i,j}+\left\{\frac{k^2}{6}\frac{\partial^3 U}{\partial t^3}-\frac{h^2}{12}\frac{\partial^4 U}{\partial x^4}\right.$$
$$\left.+(2\theta-1)\frac{2k}{h^2}\frac{\partial U}{\partial t}+\frac{k^2}{h^2}\frac{\partial^2 U}{\partial t^2}\right\}+O\left(\frac{k^3}{h^2}, h^4, k^4\right).$$

Hence the result since $\partial U/\partial t-\partial^2 U/\partial x^2=0$.

Case (i) $k=rh$

As $h\rightarrow 0$,

$$T_{i,j}=F_{i,j}(U)\rightarrow\left\{\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+(2\theta-1)\frac{2r}{h}\frac{\partial U}{\partial t}+r^2\frac{\partial^2 U}{\partial t^2}\right\}_{i,j}.$$

When $\theta\neq\frac{1}{2}$ the third term tends to infinity. When $\theta=\frac{1}{2}$ the limiting value of $T_{i,j}$ is

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+r^2\frac{\partial^2 U}{\partial t^2}.$$

In this case the finite-difference equation is consistent with the *hyperbolic* equation

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+r^2\frac{\partial^2 U}{\partial t^2}=0.$$

Hence the difference equation is always inconsistent with $\partial U/\partial t-\partial^2 U/\partial x^2=0$ when $k=rh$.

Case (ii) $k=rh^2$

As $h\rightarrow 0$,

$$T_{i,j}\rightarrow\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}+2(2\theta-1)r\frac{\partial U}{\partial t}.$$

When $\theta\neq\frac{1}{2}$ the difference scheme is consistent with the parabolic equation

$$\{1+2(2\theta-1)r\}\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0.$$

It is only when $\theta=\frac{1}{2}$ that the difference scheme is consistent with the given differential equation. This is then the well-known Du Fort and Frankel three-level explicit scheme which is also stable for all $r>0$. (See Worked Example 3.2). It was devised to overcome the unconditional instability of the early Richardson explicit scheme

$$\frac{u_{i,j+1}-u_{i,j-1}}{2k}-\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}=0,$$

but to retain the advantage of the central-difference approximation to the time-derivative which gives a local truncation error of $O(k^2)+O(h^2)$ as opposed to $O(k)+O(h^2)$ for the classical explicit approximation to $\partial U/\partial t-\partial^2 U/\partial x^2=0$.

Convergence and stability

The following sections are concerned with the conditions that must be satisfied if the solution of the finite-difference equations is to be a reasonably accurate approximation to the solution of the corresponding parabolic or hyperbolic partial differential equation.

These conditions are associated with two different but interrelated problems. The first concerns the convergence of the exact solution of the approximating difference equations to the solution of the differential equation; the second concerns the unbounded growth, or controlled decay or boundedness of the exact solution of the finite-difference equations, and therefore of all rounding errors introduced during the computation because the errors and exact solution are processed by the same arithmetic operations. (The stability problem.)

Descriptive treatment of convergence

Let U represent the *exact* solution of a partial differential equation with independent variables x and t, and u the *exact* solution of the difference equations used to approximate the partial differential equation. Then the finite-difference equation is said to be convergent when u tends to U at a fixed point or along a fixed t-level as δx and δt both tend to zero.

Although the conditions under which u converges to U have been established for linear elliptic, parabolic and hyperbolic second-order partial differential equations with solutions satisfying fairly general boundary and initial conditions, they are not yet known for non-linear equations except in a few particular cases. (The equation,

$$a\frac{\partial^2 U}{\partial x^2}+b\frac{\partial^2 U}{\partial x\partial t}+c\frac{\partial^2 U}{\partial t^2}+d\frac{\partial U}{\partial x}+e\frac{\partial U}{\partial t}+fU+g=0,$$

is linear when the coefficients $a, b, \ldots, g$, are constants or functions of x and t only. Otherwise it is non-linear. If the coefficients of the second-order derivatives are functions of x, t, U, $\partial U/\partial x$ and $\partial U/\partial t$ but not of second-order derivatives the equation is described as quasi-linear even though it is non-linear. The important feature of linear equations is that the sum of separate solutions is also a solution.)

The difference $(U-u)$ is called the *discretization error.* Some texts call it the truncation error but in this book the latter term will be reserved for the difference between the differential equation and its approximating difference equation. The magnitude of the discretization error at any mesh point depends on the finite-sizes of the mesh lengths, δx and δt, i.e. on the distances between consecutive, discrete grid-points, and on the number of terms in the truncated series of differences used to approximate the derivatives. Readers familiar with the calculus of finite-differences will have recognized the approximation used earlier for $\partial U/\partial t$ as the first term in either the series

$$(\delta t)\left(\frac{\partial U}{\partial t}\right)_{i,j}=(\Delta_t-\tfrac{1}{2}\Delta_t^2+\tfrac{1}{3}\Delta_t^3-\ldots)U_{i,j}$$

or the series

$$(\delta t)\left(\frac{\partial U}{\partial t}\right)_{i,j+\frac{1}{2}}=(\delta_t-\tfrac{1}{24}\delta_t^3+\tfrac{3}{640}\delta_t^5+\ldots)U_{i,j+\frac{1}{2}},$$

and the approximation for $\partial^2 U/\partial x^2$ as the first term in the series

$$(\delta x)^2\frac{\partial^2 U_{i,j}}{\partial x^2}=(\delta_x^2-\tfrac{1}{12}\delta_x^4+\tfrac{1}{90}\delta_x^6-\ldots)U_{i,j},$$

where the subscripts t and x denote the directions in which the differences are calculated. The symbols Δ and δ are the forward and central difference operators defined by $\Delta_t u_{i,j} = u_{i,j+1} - u_{i,j}$ and $\delta_t u_{i,j+\frac{1}{2}} = u_{i,j+1} - u_{i,j}$, so that $\delta_x^2 u_{i,j} = \delta_x(\delta_x u_{i,j}) = \delta_x(u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j}) = u_{i+1,j} - 2u_{i,j} + u_{i-1,j}$. Better approximations can be obtained by truncating the series after two or more terms but have the disadvantage of involving more pivotal values of u. It will be shown later that the discretization error can be analysed in terms of preceding local truncation errors. (See p. 73.)

The discretization error can usually be diminished by decreasing δx and δt, subject invariably to some relationship between them, but as this leads to an increase in the number of equations to be solved, this method of improvement is limited by such factors as cost of computation and computer storage requirements, etc.

In general, the problem of convergence is a difficult one to investigate usefully because the final expression for the discretization error is usually in terms of unknown derivatives for which no bounds can be estimated. Fortunately, however, the convergence of difference equations approximating *linear* parabolic and hyperbolic differential equations can be investigated in terms of stability and consistency, which are easier to deal with. (Reference Lax's equivalence theorem, p. 72).

Analytical treatment of convergence *(A direct method)*

The convergence of the solution of an approximating set of *linear*-difference equations to the solution of a *linear* partial differential equation is dealt with most easily via Lax's equivalence theorem. Explicit difference schemes however, can be investigated directly by deriving a difference equation for the discretization error e. Denote the exact solution of the partial differential equation by U and the exact solution of the finite-difference equation by u. Then $e = U - u$.

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1, t > 0, \tag{2.26}$$

where U is known for $0 \leqslant x \leqslant 1$ when $t = 0$, and at $x = 0$ and 1 when $t > 0$.

The simplest explicit finite-difference approximation to (2.26) is

$$\frac{u_{i,j+1}-u_{i,j}}{k}=\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2}. \tag{2.27}$$

At the mesh points,

$$u_{i,j}=U_{i,j}-e_{i,j},\quad u_{i,j+1}=U_{i,j+1}-e_{i,j+1},\text{ etc.}$$

Substitution into (2.27) leads to

$$\begin{aligned}e_{i,j+1}=re_{i-1,j}+(1-2r)e_{i,j}+re_{i+1,j}+U_{i,j+1}-U_{i,j}&\\+r(2U_{i,j}-U_{i-1,j}-U_{i+1,j}).&\end{aligned} \tag{2.28}$$

By Taylor's theorem,

$$U_{i+1,j}=U(x_i+h,t_j)=U_{i,j}+h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\frac{h^2}{2!}\frac{\partial^2 U}{\partial x^2}(x_i+\theta_1 h,t_j),$$

$$U_{i-1,j}=U(x_i-h,t_j)=U_{i,j}-h\left(\frac{\partial U}{\partial x}\right)_{i,j}+\frac{h^2}{2!}\frac{\partial^2 U}{\partial x^2}(x_i-\theta_2 h,t_j),$$

$$U_{i,j+1}=U(x_i,t_j+k)=U_{i,j}+k\frac{\partial U}{\partial t}(x_i,t_j+\theta_3 k),$$

where $0<\theta_1<1$, $0<\theta_2<1$ and $0<\theta_3<1$. Substitution into eqn (2.28) gives

$$\begin{aligned}e_{i,j+1}=re_{i-1,j}+(1-2r)e_{i,j}+re_{i+1,j}&\\+k\left\{\frac{\partial U}{\partial t}(x_i,t_j+\theta_3 k)-\frac{\partial^2 U}{\partial x^2}(x_i+\theta_4 h,t_j)\right\},&\end{aligned} \tag{2.29}$$

where $-1<\theta_4<1$.

This is a difference equation for $e_{i,j}$ which fortunately we need not solve.

Let E_j denote the maximum value of $|e_{i,j}|$ along the jth time-row and M the maximum modulus of the expression in the braces for all i and j. When $r\leqslant\frac{1}{2}$, all the coefficients of e in eqn (2.29) are positive or zero, so

$$\begin{aligned}|e_{i,j+1}|&\leqslant r|e_{i-1,j}|+(1-2r)|e_{i,j}|+r|e_{i+1,j}|+kM\\&\leqslant rE_j+(1-2r)E_j+rE_j+kM\\&=E_j+kM.\end{aligned}$$

As this is true for all values of i it is true for $\max|e_{i,j+1}|$. Hence

$$E_{j+1}\leqslant E_j+kM\leqslant(E_{j-1}+kM)+kM=E_{j-1}+2kM,$$

etc., from which it follows that

$$E_j \leqslant E_0 + jkM = tM,$$

because the initial values for u and U are the same, i.e. $E_0 = 0$. When h tends to zero, $k = rh^2$ also tends to zero and M tends to

$$\left(\frac{\partial U}{\partial t} - \frac{\partial^2 U}{\partial x^2}\right)_{i,j}.$$

Since U is a solution of eqn (2.26) the limiting value of M and therefore of E_j is zero. As $|U_{i,j} - u_{i,j}| \leqslant E_j$, this proves that u converges to U as h tends to zero when $r \leqslant \frac{1}{2}$ and t is finite.

When $r > \frac{1}{2}$ it can be shown that the complementary function of the difference equation (2.29) tends to infinity as h tends to zero. There is no need, however, to do this when our main purpose is to find the conditions necessary for a useful numerical solution because we shall prove later that this finite-difference scheme is stable for $r \leqslant \frac{1}{2}$ but unstable for $r > \frac{1}{2}$.

The proof above implies that $\partial U/\partial t$ and $\partial^2 U/\partial x^2$ are uniformly continuous and bounded throughout the solution domain. This was so in the Worked Example 2.1 in which $\partial^2 U/\partial x^2$ was initially zero in spite of the discontinuity in $\partial U/\partial x$. If it is assumed that U possesses continuous bounded derivatives up to order three in t and order six in x, Exercise 13 shows that the discretization error is of order h^2, except when $r = \frac{1}{6}$ in which case it is of order h^4.

Descriptive treatment of stability

The equations that are actually solved are, of course, the finite-difference equations, and their application and solution to successive time-rows advances the finite-difference solution from the initial line, on which initial values are known, to time-levels k, $2k, \ldots Jk = T$, say, where T is finite. If no rounding errors were introduced into this numerical process then the exact solution $u_{i,j}$ of the finite-difference equations would be obtained at each mesh point (i, j), $i = 0(1)N$, $0 < j \leqslant J$.

The essential idea defining stability is that this numerical process, applied exactly, should limit the amplification of all components of the initial conditions.

For linear initial-value boundary-value problems, Lax and

Richtmyer have related stability to convergence via Lax's Equivalence Theorem (p. 72) by defining stability, in effect, in terms of the boundedness of the solution of the finite-difference equations at a fixed time-level T as $k \rightarrow 0$, i.e. as $J \rightarrow \infty$, it being assumed that $\delta x = h$ is related to k in such a way that $h \rightarrow 0$ as $k \rightarrow 0$.

Assume that the vector of solution values $\boldsymbol{u}_{j+1} = [u_{1,j+1}, u_{2,j+1}, \ldots, u_{N-1,j+1}]^T$ of the finite-difference equations at the $(j+1)$th time-level is related to the vector of solution values at the jth time-level by the equation

$$\boldsymbol{u}_{j+1} = \mathbf{A}\mathbf{u}_j + \mathbf{b}_j,$$

where $\mathbf{b}_j$ is a column vector of known boundary-values and zeros, and matrix $\mathbf{A}$ an $(N-1)\times(N-1)$ matrix of known elements. Then it will be shown that the practical consequence of this definition of stability is that a norm of matrix $\mathbf{A}$ compatible with a norm of $\mathbf{u}$ must satisfy

$$\|\mathbf{A}\| \leqslant 1$$

when the solution of the partial differential equation does not increase as t increases, or

$$\|\mathbf{A}\| \leqslant 1 + O(k)$$

when the solution of the partial differential increases as t increases.

These conditions also ensure the boundedness of all rounding errors because they are subject to the same arithmetic operations as the finite-difference solution.

In an actual computation, however, k and h are normally kept constant as the solution is propagated forward time-level by time-level from $t = 0$ to $t_j = jk$, and in many textbooks and papers stability is defined in terms of the boundedness of this numerical solution as $j \rightarrow \infty$, k fixed. In this process the order $(N-1)$ of matrix $\mathbf{A}$ remains constant, unlike the matrix $\mathbf{A}$ associated with Lax and Richtmyer's definition. The matrix method of analysis then shows that the equations are stable if the largest of the moduli of the eigenvalues of matrix $\mathbf{A}$, i.e. the spectral radius $\rho(\mathbf{A})$ of $\mathbf{A}$, satisfies

$$\rho(\mathbf{A}) \leqslant 1,$$

when the solution of the differential equation does not increase with increasing t.

Although this condition ensures the boundedness of the computed solution it does not guarantee convergence unless the eigenvalues of **A** are restricted to satisfy $\rho(\mathbf{A}) \leqslant \|\mathbf{A}\| \leqslant 1$, as $N \rightarrow \infty$. In practice, assuming that the difference equations are consistent, it is usually only in the immediate neighbourhood of $\rho(\mathbf{A}) = 1$ that non-convergence might occur. An illuminating discussion of these points is given in reference 19. (If the solution of the partial differential equation increases as $t \rightarrow \infty$, the condition for stability with fixed h and k is then $\rho(\mathbf{A}) \leqslant 1 + O(k)$. See p. 66.)

Vector and matrix norms

This section is needed for the Lax–Richtmyer definition of stability.

Vector norms

The norm of vector **x** is a real positive number giving a measure of the 'size' of the vector and is denoted by $\|\mathbf{x}\|$. It must satisfy the following axioms.

(i) $\|\mathbf{x}\| > 0$ if $\mathbf{x} \neq \mathbf{0}$ and $\|\mathbf{x}\| = 0$ if $\mathbf{x} = \mathbf{0}$.
(ii) $\|c\mathbf{x}\| = |c|\,\|\mathbf{x}\|$ for a real or complex scalar c.
(iii) $\|\mathbf{x} + \mathbf{y}\| \leqslant \|\mathbf{x}\| + \|\mathbf{y}\|$.

If the $n \times 1$ vector **x** has components $x_1, x_2, \ldots, x_n$, then the three most commonly used norms are defined as follows.

The 1-norm of **x** is the sum of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_1 = |x_1| + |x_2| + \ldots + |x_n| = \sum_{i=1}^{n} |x_i|.$$

The infinity norm of **x** is the maximum of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_\infty = \max_i |x_i|.$$

The 2-norm of **x** is the square root of the sum of the squares of the moduli of the components of **x**, i.e.

$$\|\mathbf{x}\|_2 = (|x_1|^2 + |x_2|^2 + \cdots |x_n|^2)^{\frac{1}{2}} = \left[\sum_{i=1}^{n} |x_i|^2\right]^{\frac{1}{2}}.$$

The 2-norm gives the 'length' of the vector.

Matrix norms

The norm of matrix $\mathbf{A}$ is a real positive number giving a measure of the 'size' of the matrix and must satisfy the following axioms.

(i) $\|\mathbf{A}\|>0$ if $\mathbf{A}\neq\mathbf{0}$ and $\|\mathbf{A}\|=0$ if $\mathbf{A}=\mathbf{0}$.
(ii) $\|c\mathbf{A}\|=|c|\,\|\mathbf{A}\|$ for a real or complex scalar c.
(iii) $\|\mathbf{A}+\mathbf{B}\|\leqslant\|\mathbf{A}\|+\|\mathbf{B}\|$.
(iv) $\|\mathbf{AB}\|\leqslant\|\mathbf{A}\|\,\|\mathbf{B}\|$.

Compatible or consistent norms

Vectors and matrices occur together so it is essential that they satisfy a condition equivalent to (iv). As a consequence, matrix and vector norms are said to be compatible or consistent if

$$\|\mathbf{Ax}\|\leqslant\|\mathbf{A}\|\,\|\mathbf{x}\|,\quad \mathbf{x}\neq\mathbf{0}.$$

Subordinate matrix norms

Let $\mathbf{A}$ be an $n\times n$ matrix and $\mathbf{x}$ a member of the set S of $n\times 1$ vectors whose norms are unity, i.e. $\mathbf{x}\in S$ if $\|\mathbf{x}\|=1$. In general, the norm of the vector $\mathbf{Ax}$ will vary as $\mathbf{x}$ varies, $\mathbf{x}\in S$. Let $\mathbf{x}_0$ be a member of S that makes $\|\mathbf{Ax}\|$ attain its maximum value. Then the norm of matrix $\mathbf{A}$ is defined by

$$\|\mathbf{A}\|=\|\mathbf{Ax}_0\|=\max_{\|\mathbf{x}\|=1}\|\mathbf{Ax}\|.$$

This matrix norm is said to be subordinate to the vector norm and automatically satisfies the compatibility condition, because, if $\mathbf{x}=\mathbf{x}_1$ is any other member of S, $\|\mathbf{Ax}_1\|\leqslant\|\mathbf{Ax}_0\|=\|\mathbf{A}\|=\|\mathbf{A}\|\,\|\mathbf{x}_1\|$, since $\|\mathbf{x}_1\|=1$. It also follows that for all subordinate matrix norms,

$$\|\mathbf{I}\|=\max_{\|\mathbf{x}\|=1}\|\mathbf{Ix}\|=\max_{\|\mathbf{x}\|=1}\|\mathbf{x}\|=1$$

where $\mathbf{I}$ is the unit matrix. The definitions of the 1, 2, and ∞ norms with $\|\mathbf{x}\|=1$ lead to the following results which are proved in most linear algebra books. (See Exercise 15.)

The 1-norm of matrix $\mathbf{A}$ is the maximum column sum of the moduli of the elements of $\mathbf{A}$.

The infinity norm of matrix $\mathbf{A}$ is the maximum row sum of the moduli of the elements of $\mathbf{A}$.

The 2-norm of matrix $\mathbf{A}$ is the square root of the spectral radius of $\mathbf{A}^H\mathbf{A}$, where $\mathbf{A}^H=(\bar{\mathbf{A}})^T$, the transpose of the conjugate complex of $\mathbf{A}$. For example, if

$$\mathbf{A}=\begin{bmatrix}-1 & 1\\ 3 & -2\end{bmatrix}, \text{ then } \mathbf{A}^H\mathbf{A}=\mathbf{A}^T\mathbf{A}=\begin{bmatrix}10 & -7\\ -7 & 5\end{bmatrix}$$

with eigenvalues 14.93 and 0.067. Hence $\|\mathbf{A}\|_1=1+3=4$, $\|\mathbf{A}\|_\infty=3+2=5$ and $\|\mathbf{A}\|_2=\sqrt{14.93}=3.86$.

When matrix $\mathbf{A}$ is real and symmetric, $\mathbf{A}^H=\mathbf{A}$, and

$$\|\mathbf{A}\|_2=[\rho(\mathbf{A}^2)]^{\frac{1}{2}}=[\rho^2(\mathbf{A})]^{\frac{1}{2}}=\rho(\mathbf{A})=\max_i|\lambda_i|.$$

A bound for the spectral radius

Let λ_i be an eigenvalue of the $n\times n$ matrix $\mathbf{A}$ and $\mathbf{x}_i$ the corresponding eigenvector. Hence

$$\mathbf{A}\mathbf{x}_i=\lambda_i\mathbf{x}_i$$

and

$$\|\mathbf{A}\mathbf{x}_i\|=\|\lambda_i\mathbf{x}_i\|=|\lambda_i|\,\|\mathbf{x}_i\|.$$

For all compatible matrix and vector norms it follows that

$$|\lambda_i|\,\|\mathbf{x}_i\|=\|\mathbf{A}\mathbf{x}_i\|\leqslant\|\mathbf{A}\|\,\|\mathbf{x}_i\|.$$

Therefore,

$$|\lambda_i|\leqslant\|\mathbf{A}\|,\quad i=1(1)n.$$

Hence,

$$\rho(\mathbf{A})\leqslant\|\mathbf{A}\|.$$

A necessary and sufficient condition for stability. *(Constants coefficients)*

Let the solution domain of the partial differential equation be the finite rectangle $0\leqslant x\leqslant 1$, $0\leqslant t\leqslant T$, and subdivide it into uniform rectangular meshes by the lines $x_i=ih$, $i=0(1)N$, where $Nh=1$, and the lines $t_j=jk$, $j=0(1)J$ where $Jk=T$. It will be assumed that h is related to k by some relationship such as $k=rh$ or $k=rh^2$, $r>0$ and finite, so that $h\to 0$ as $k\to 0$.

Assume that the finite-difference equation relating the mesh-point values along the $(j+1)$th and jth time-rows is

$$b_{i-1}u_{i-1,j+1}+b_iu_{i,j+1}+b_{i+1}u_{i+1,j+1}=c_{i-1}u_{i-1,j}+c_iu_{i,j}+c_{i+1}u_{i+1,j}$$

where the coefficients are constants.

If the boundary values at $i=0$ and N, $j>0$, are known, these $(N-1)$ equations for $i=1(1)N-1$ can be written in matrix form as

$$\begin{bmatrix} b_1 & b_2 & & & & \\ b_1 & b_2 & b_3 & & & \\ \cdot & \cdot & \cdot & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & b_{N-3} & b_{N-2} & b_{N-1} \\ & & & & b_{N-2} & b_{N-1} \end{bmatrix} \begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix}$$

$$= \begin{bmatrix} c_1 & c_2 & & & & \\ c_1 & c_2 & c_3 & & & \\ \cdot & \cdot & \cdot & & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & \cdot & \cdot & \\ & & & c_{N-3} & c_{N-2} & c_{N-1} \\ & & & & c_{N-2} & c_{N-1} \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix} + \begin{bmatrix} c_0u_{0,j}-b_0u_{0,j+1} \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ c_Nu_{N,j}-b_Nu_{N,j+1} \end{bmatrix}$$

i.e. as $\mathbf{B}\mathbf{u}_{j+1}=\mathbf{C}\mathbf{u}_j+\mathbf{d}_j$, where the matrices **B** and **C** of order $(N-1)$ are as shown, $\mathbf{u}_{j+1}$ denotes the column vector with components $u_{1,j+1}, u_{2,j+1}, \ldots, u_{N-1,j+1}$ and $\mathbf{d}_j$ denotes the column vector of known boundary values and zeros.

Hence,

$$\mathbf{u}_{j+1}=\mathbf{B}^{-1}\mathbf{C}\mathbf{u}_j+\mathbf{B}^{-1}\mathbf{d}_j,$$

which may be expressed more conveniently as

$$\mathbf{u}_{j+1}=\mathbf{A}\mathbf{u}_j+\mathbf{f}_j,$$

where $\mathbf{A}=\mathbf{B}^{-1}\mathbf{C}$ and $\mathbf{f}_j=\mathbf{B}^{-1}\mathbf{d}_j$. Applied recursively, this leads to

$$\begin{aligned} \mathbf{u}_j &= \mathbf{A}\mathbf{u}_{j-1}+\mathbf{f}_{j-1}=\mathbf{A}(\mathbf{A}\mathbf{u}_{j-2}+\mathbf{f}_{j-2})+\mathbf{f}_{j-1} \\ &= \mathbf{A}^2\mathbf{u}_{j-2}+\mathbf{A}\mathbf{f}_{j-2}+\mathbf{f}_{j-1} \\ &= \ldots \\ &= \mathbf{A}^j\mathbf{u}_0+\mathbf{A}^{j-1}\mathbf{f}_0+\mathbf{A}^{j-2}\mathbf{f}_1+\ldots+\mathbf{f}_{j-1}, \qquad (2.30) \end{aligned}$$

where $\mathbf{u}_0$ is the vector of initial values and $\mathbf{f}_0, \mathbf{f}_1, \ldots, \mathbf{f}_{j-1}$ are vectors of known boundary-values. When we are concerned more with a property of the equations, such as stability, than with a numerical solution, the constant vectors can be eliminated by investigating the propagation of a perturbation.

Perturb the vector of initial values $\mathbf{u}_0$ to $\mathbf{u}_0^*$. The exact solution at the jth time-row will then be

$$\mathbf{u}_j^* = \mathbf{A}^j\mathbf{u}_0^* + \mathbf{A}^{j-1}\mathbf{f}_0 + \mathbf{A}^{j-2}\mathbf{f}_1 + \ldots + \mathbf{f}_{j-1}. \tag{2.31}$$

If the perturbation or 'error' vector $\mathbf{e}$ is defined by

$$\mathbf{e} = \mathbf{u}^* - \mathbf{u},$$

it follows by eqns (2.30) and (2.31) that

$$\mathbf{e}_j = \mathbf{u}_j^* - \mathbf{u}_j = \mathbf{A}^j(\mathbf{u}_0^* - \mathbf{u}_0) = \mathbf{A}^j\mathbf{e}_0,\ j = 1(1)J. \tag{2.32}$$

In other words, a perturbation $\mathbf{e}_0$ of the initial values will be propagated according to the equation

$$\mathbf{e}_j = \mathbf{A}\mathbf{e}_{j-1} = \mathbf{A}^2\mathbf{e}_{j-2} = \ldots = \mathbf{A}^j\mathbf{e}_0, \quad j = 1(1)J.$$

Hence, for compatible matrix and vector norms,

$$\|\mathbf{e}_j\| \leqslant \|\mathbf{A}^j\|\,\|\mathbf{e}_0\|.$$

Lax and Richtmyer define the difference scheme to be stable when there exists a positive number M, independent of j, h, and k, such that

$$\|\mathbf{A}^j\| \leqslant M, \quad j = 1(1)J.$$

This clearly limits the amplification of any initial perturbation, and therefore of any arbitrary initial rounding errors, because

$$\|\mathbf{e}_j\| \leqslant M\,\|\mathbf{e}_0\|.$$

In reference 25, this definition of stability is related to convergence via Lax's equivalence theorem (p. 72).

Since

$$\|\mathbf{A}^j\| = \|\mathbf{A}\mathbf{A}^{j-1}\| \leqslant \|\mathbf{A}\|\,\|\mathbf{A}^{j-1}\| \leqslant \ldots \leqslant \|\mathbf{A}\|^j,$$

it follows that the Lax–Richtmyer definition of stability is satisfied by

$$\|\mathbf{A}\| \leqslant 1.$$

This is the necessary and sufficient condition for the difference equations to be stable when the solution of the partial differential equation does not increase as t increases.

When this condition is satisfied it follows automatically that the spectral radius $\rho(\mathbf{A}) \leqslant 1$ since $\rho(\mathbf{A}) \leqslant \|\mathbf{A}\|$.

If, however, $\rho(\mathbf{A}) \leqslant 1$, it does not follow that $\|\mathbf{A}\| \leqslant 1$. This is demonstrated by the simple example

$$\mathbf{A} = \begin{bmatrix} -0.8 & 0 \\ 0.4 & 0.7 \end{bmatrix},$$

for which $\lambda_1 = -0.8$, $\lambda_2 = 0.7$, $\rho(\mathbf{A}) = 0.8$, $\|\mathbf{A}\|_1 = 1.2$ and $\|\mathbf{A}\|_\infty = 1.1$. If, however, $\mathbf{A}$ is real and symmetric, then

$$\|\mathbf{A}\|_2 = \surd\rho(\mathbf{A}^T\mathbf{A}) = \surd\rho(\mathbf{A}^2) = \surd\rho^2(\mathbf{A}) = \rho(\mathbf{A}).$$

Example 2.8

Consider the stability of the classical explicit equations

$$u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{i,j} + ru_{i+1,j}, \quad i = 1(1)N-1,$$

for which the $(N-1)\times(N-1)$ matrix $\mathbf{A}$ is

$$\begin{bmatrix} (1-2r) & r & & \\ r & (1-2r)r & & \\ & r & (1-2r) & r \\ & & r & (1-2r), \end{bmatrix}$$

where $r = k/h^2 > 0$, and it is assumed that the boundary values $u_{0,j}$ and $u_{N,j}$ are known for $j = 1, 2, \ldots$.
When $1-2r \geqslant 0$, then $0 < r \leqslant \frac{1}{2}$
and

$$\|\mathbf{A}\|_\infty = r + (1-2r) + r = 1.$$

When $1-2r < 0$, $r > \frac{1}{2}$, $|1-2r| = 2r-1$
and

$$\|\mathbf{A}\|_\infty = r + 2r - 1 + r = 4r - 1 > 1.$$

Therefore the scheme is stable for $0 < r \leqslant \frac{1}{2}$.
Alternatively, since matrix $\mathbf{A}$ is real and symmetric,

$$\|\mathbf{A}\|_2 = \rho(\mathbf{A}) = \max_s |\mu_s|,$$

where μ_s is the sth eigenvalue of $\mathbf{A}$. Now $\mathbf{A}$ can be written as

$$\begin{bmatrix} 1 & & & & \\ & 1 & & & \\ & & \ddots & & \\ & & & 1 & \\ & & & & 1 \end{bmatrix} + r\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & & \ddots & & \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix} = \mathbf{I}_{N-1} + r\mathbf{T}_{N-1},$$

where $\mathbf{I}_{N-1}$ is the unit matrix of order $(N-1)$ and $\mathbf{T}_{N-1}$ an $(N-1)\times(N-1)$ matrix whose eigenvalues λ_s are given by

$$\lambda_s = -4\sin^2 s\pi/2N, \quad s = 1(1)N-1. \text{ (See p. 59.)}$$

Hence the eigenvalues of $\mathbf{A}$, as shown later in 'A note on eigenvalues and eigenvectors' are $\mu_s = 1-4r\sin^2 s\pi/2N$.

Therefore the equations will be stable when

$$\|\mathbf{A}\|_2 = \max_s |1-4r\sin^2 s\pi/2N| \leqslant 1,$$

i.e.,

$$-1 \leqslant 1-4r\sin^2 s\pi/2N \leqslant 1, \quad s = 1(1)N-1.$$

The left-hand inequality gives that

$$r \leqslant 1/2\sin^2(N-1)\pi/2N.$$

As $h \to 0$, $N \to \infty$ and $\sin^2(N-1)\pi/2N \to 1$.
Hence $r \leqslant \frac{1}{2}$.

It has been shown that these equations are also consistent. Hence by Lax's equivalence theorem they are also convergent for $0 < r \leqslant \frac{1}{2}$.

Example 2.9

The Crank–Nicolson equations (2.10) are

$$-ru_{i-1,j+1} + (2+2r)u_{i,j+1} - ru_{i+1,j+1} = ru_{i-1,j} + (2-2r)u_{i,j} + ru_{i+1,j}, \; i = 1(1)N-1.$$

In matrix form, for known boundary values, these give

$$\begin{bmatrix} (2+2r) & -r & & & \\ -r & (2+2r) & -r & & \\ & \cdot & & & \\ & & \cdot & & \\ & & \cdot & & \\ & & -r & (2+2r) & -r \\ & & & -r & (2+2r) \end{bmatrix} \begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix}$$

$$= \begin{bmatrix} (2-2r) & r & & & \\ r & (2-2r) & r & & \\ & \cdot & & & \\ & & \cdot & & \\ & & \cdot & & \\ & & r & (2-2r) & r \\ & & & r & (2-2r) \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix} + \mathbf{b}_j,$$

where $\mathbf{b}_j$ is a vector of known boundary values and zeros. This can be written as

$$(2\mathbf{I}_{N-1} - r\mathbf{T}_{N-1})u_{j+1} = (2\mathbf{I}_{N-1} + r\mathbf{T}_{N-1})\mathbf{u}_j + \mathbf{b}_j$$

from which it follows that matrix $\mathbf{A}$ of eqn (2.30) is

$$\mathbf{A} = (2\mathbf{I}_{N-1} - r\mathbf{T}_{N-1})^{-1}(2\mathbf{I}_{N-1} + r\mathbf{T}_{N-1}).$$

In Exercise 14 it is proved that if the $n \times n$ symmetric matrices $\mathbf{B}$ and $\mathbf{C}$ commute then $\mathbf{B}^{-1}\mathbf{C}$, $\mathbf{B}\mathbf{C}^{-1}$ and $\mathbf{B}^{-1}\mathbf{C}^{-1}$ are symmetric. Matrix $\mathbf{T}_{N-1}$ is symmetric so $2\mathbf{I}_{N-1} - r\mathbf{T}_{N-1}$ and $2\mathbf{I}_{N-1} + r\mathbf{T}_{N-1}$ are symmetric. They also commute as their multiplication immediately shows. Hence matrix $\mathbf{A}$ is symmetric. Since the eigenvalues of $\mathbf{T}_{N-1}$ are $\lambda_s = -4\sin^2 s\pi/2N$, $s = 1(1)N-1$, it follows that the eigenvalues of $\mathbf{A}$ are $(2+4r\sin^2 s\pi/2N)^{-1}(2-4r\sin^2 s\pi/2N)$. (See p. 59.)
Therefore

$$\|\mathbf{A}\|_2 = \rho(\mathbf{A}) = \max_s \left|\frac{1-2r\sin^2 s\pi/2N}{1+2r\sin^2 s\pi/2N}\right| < 1 \text{ for all } r > 0,$$

proving that the Crank–Nicolson equations are unconditionally stable. They are also consistent, reference Exercise 11, so they are also convergent, although it will be shown under A_0-stability that r must be restricted in order to avoid the possibility of finite oscillations near points of discontinuity.

Matrix method of analysis, fixed mesh lengths

The following method of analysis will establish the conditions necessary for the boundedness of the analytical solution of the finite-difference equations as $t_j = jk$ tends to infinity for fixed mesh lengths h and k. These conditions may not, however, be sufficient to ensure convergence when the equations are also consistent and large finite errors can occur near the end points of the range of values of some parameter, such as $r = k/h^2$, for which the equations are bounded.

Consider the classical explicit finite-difference equations incorporating known boundary values, namely,

$$\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix} = \begin{bmatrix} (1-2r) & r & & & \\ r & (1-2r) & r & & \\ & \cdot & \cdot & \cdot & \\ & & \cdot & \cdot & \cdot \\ & & \cdot & \cdot & \cdot \\ & & r & (1-2r) & r \\ & & & r & (1-2r) \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \cdot \\ \cdot \\ \cdot \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix} + \begin{bmatrix} ru_{0,j} \\ 0 \\ \cdot \\ \cdot \\ \cdot \\ 0 \\ ru_{N,j} \end{bmatrix},$$

i.e.

$$\mathbf{u}_{j+1} = \mathbf{A}\mathbf{u}_j + \mathbf{b}_j.$$

As shown by eqn (2.32), if the vector of initial values $\mathbf{u}_0$ is perturbed to $\mathbf{u}_0^*$ and no further perturbations or errors are introduced into the subsequent calculations, the perturbation vector $\mathbf{e} = \mathbf{u}^* - \mathbf{u}$ will be propagated forward in time, according to the equation

$$\mathbf{e}_j = \mathbf{A}^j \mathbf{e}_0,$$

a procedure that eliminates the boundary values.

For fixed mesh lengths h and k the difference equations will be stable if $\mathbf{e}_j$ remains bounded as j increases indefinitely. This can always be investigated by expressing the initial perturbation vector in terms of the eigenvectors of $\mathbf{A}$, which remain fixed as j increases.

Assume that matrix $\mathbf{A}$ is non-deficient, i.e. has $(N-1)$ linearly independent eigenvectors $\mathbf{v}_s$, which will be so if the eigenvalues λ_s of $\mathbf{A}$ are all distinct or $\mathbf{A}$ is real and symmetric. Then these eigenvectors can be used as a basis for our $(N-1)$ dimensional vector space and the perturbation vector $\mathbf{e}_0$, with its $(N-1)$ components, can be expressed uniquely as a linear combination of them, namely,

$$\mathbf{e}_0 = \sum_{s=1}^{N} c_s \mathbf{v}_s,$$

where the c_s, $s = 1(1)N-1$, are known scalars.

The perturbations along time-level $t = k$, resulting from the initial perturbation $\mathbf{e}_0$ will be

$$\mathbf{e}_1 = \mathbf{A}\mathbf{e}_0 = \mathbf{A} \sum_{s=1}^{N-1} c_s \mathbf{v}_s = \sum c_s \mathbf{A}\mathbf{v}_s.$$

But $\mathbf{A}\mathbf{v}_s = \lambda_s \mathbf{v}_s$ by the definition of an eigenvalue. Therefore,

$$\mathbf{e}_1 = \sum c_s \lambda_s \mathbf{v}_s.$$

Similarly,

$$\mathbf{e}_j = \sum_{s=1}^{N-1} c_s \lambda_s^j \mathbf{v}_s.$$

This shows that the perturbations will not increase exponentially with j provided

$$\max_s |\lambda_s| \leqslant 1, \quad s = 1(1)N-1.$$

By p. 59, $\lambda_s = 1 - 4r \sin^2 s\pi/2N$.
Therefore any perturbation, rounding errors and $u_{i,j}$ will be bounded as j increases if

$$-1 \leqslant 1 - 4r \sin^2 s\pi/2N \leqslant 1, \quad \text{where } r > 0.$$

This is satisfied by $r \leqslant \frac{1}{2}$.

A note on eigenvalues and eigenvectors

Let $\mathbf{x}$ be an eigenvector of the matrix $\mathbf{A}$ corresponding to the eigenvalue λ. Then $\mathbf{A}\mathbf{x} = \lambda\mathbf{x}$. Hence $\mathbf{A}(\mathbf{A}\mathbf{x}) = \mathbf{A}^2\mathbf{x} = \lambda\mathbf{A}\mathbf{x} = \lambda^2\mathbf{x}$, showing that the matrix $\mathbf{A}^2$ has an eigenvalue λ^2 corresponding to the eigenvector $\mathbf{x}$. Similarly $\mathbf{A}^p\mathbf{x} = \lambda^p\mathbf{x}$, $p = 3, 4, \ldots$.

(i) If $f(\mathbf{A}) = a_p\mathbf{A}^p + a_{p-1}\mathbf{A}^{p-1} + \ldots + a_0\mathbf{I}$ is a polynomial in $\mathbf{A}$ with scalar coefficients $a_p, \ldots, a_0$, then $f(\mathbf{A})\mathbf{x} = (a_p\lambda^p + \ldots + a_0)\mathbf{x} = f(\lambda)\mathbf{x}$, showing that $f(\mathbf{A})$ has an eigenvalue $f(\lambda)$ corresponding to the eigenvector $\mathbf{x}$.

(ii) The eigenvalue of $[f_1(\mathbf{A})]^{-1}f_2(\mathbf{A})$ corresponding to the eigenvector $\mathbf{x}$ is $f_2(\lambda)/f_1(\lambda)$, where $f_1(\mathbf{A})$ and $f_2(\mathbf{A})$ are polynomials in $\mathbf{A}$. The proof is as follows. By (i),

$$f_1(\mathbf{A})\mathbf{x}=f_1(\lambda)\mathbf{x} \quad \text{and} \quad f_2(\mathbf{A})\mathbf{x}=f_2(\lambda)\mathbf{x}.$$

Premultiply both equations by $[f_1(\mathbf{A})]^{-1}$ and write as

$$[f_1(\mathbf{A})]^{-1}\mathbf{x}=\mathbf{x}/f_1(\lambda) \quad \text{and} \quad [f_1(\mathbf{A})]^{-1}f_2(\mathbf{A})\mathbf{x}=f_2(\lambda)[f_1(\mathbf{A})]^{-1}\mathbf{x}.$$

Then the elimination of $[f_1(\mathbf{A})]^{-1}\mathbf{x}$ between these two equations shows that

$$[f_1(\mathbf{A})]^{-1}f_2(\mathbf{A})\mathbf{x}=\{f_2(\lambda)/f_1(\lambda)\}\mathbf{x},$$

which states, by the definition of an eigenvalue, that $f_2(\lambda)/f_1(\lambda)$ is an eigenvalue of $[f_1(\mathbf{A})]^{-1}f_2(\mathbf{A})$ corresponding to the eigenvector $\mathbf{x}$. In a similar manner the eigenvalue of $f_2(\mathbf{A})[f_1(\mathbf{A})]^{-1}$ corresponding to the eigenvector $\mathbf{x}$ is $f_2(\lambda)/f_1(\lambda)$.

In particular, the eigenvalue of $[f_1(\mathbf{A})]^{-1}$ corresponding to the eigenvector $\mathbf{x}$ is $1/f_1(\lambda)$.

The eigenvalues of a common tridiagonal matrix

The eigenvalue of the $N\times N$ matrix

$$\begin{bmatrix} a & b & & & & \\ c & a & b & & & \\ & c & a & b & & \\ & & & \cdots & & \\ & & & c & a & b \\ & & & & c & a \end{bmatrix}$$

are

$$\lambda_s = a+2\{\sqrt{(bc)}\}\cos\frac{s\pi}{N+1}, \quad s=1(1)N,$$

where a, b, and c may be real or complex. A proof is given on p. 154.

Another useful result is the following. If a real tridiagonal matrix has either all its off-diagonal elements positive or all its off-diagonal elements negative, then all its eigenvalues are real. A proof is given in Exercise 16.

Useful theorems on bounds for eigenvalues

Gerschgorin's first theorem

The largest of the moduli of the eigenvalues of the square matrix **A** cannot exceed the largest sum of the moduli of the elements along any row or any column. In other words

$$\rho(A) \leqslant \|\mathbf{A}\|_1 \quad \text{or} \quad \|\mathbf{A}\|_\infty.$$

Proof

Let λ_i be an eigenvalue of the $N \times N$ matrix **A**, and $\mathbf{x}_i$ the corresponding eigenvector with components $v_1, v_2 \ldots v_n$. Then the equation

$$\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$$

in detail, is

$$\begin{aligned} a_{1,1}v_1 + a_{1,2}v_2 + \ldots + a_{1,n}v_n &= \lambda_i v_1, \\ a_{2,1}v_1 + a_{2,2}v_2 + \ldots + a_{2,n}v_n &= \lambda_i v_2, \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad & \quad \cdot \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad & \quad \cdot \\ a_{s,1}v_1 + a_{s,2}v_2 + \ldots + a_{s,n}\,v_n &= \lambda_i v_s, \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad & \quad \cdot \\ \cdot \qquad \cdot \qquad\qquad \cdot \quad & \quad \cdot \end{aligned}$$

Let v_s be the largest in modulus of $v_1, v_2, \ldots, v_n$. Select the sth equation and divide by v_s, giving

$$\lambda_i = a_{s,1}\left(\frac{v_1}{v_s}\right) + a_{s,2}\left(\frac{v_2}{v_s}\right) + \ldots + a_{s,n}\left(\frac{v_n}{v_s}\right).$$

Therefore

$$|\lambda_i| \leqslant |a_{s,1}| + |a_{s,2}| + \ldots + |a_{s,n}|,$$

because

$$\left|\frac{v_i}{v_s}\right| \leqslant 1, \quad i = 1, 2, \ldots, n.$$

If this is not the largest row sum then $|\lambda_i| <$ the largest row sum. In particular this holds for $|\lambda_i| = \max|\lambda_s|$, $s = 1(1)N$.
Since the eigenvalues of the transpose of **A** are the same as those of **A** the theorem is also true for columns.

Gerschgorin's circle theorem or Brauer's theorem

Let P_s be the sum of the moduli of the elements along the sth row excluding the diagonal element $a_{s,s}$. Then each eigenvalue of $\mathbf{A}$ lies inside or on the boundary of at least one of the circles $|\lambda - a_{s,s}| = P_s$.

Proof

By the previous proof,

$$\lambda_i = a_{s,1}\left(\frac{v_1}{v_s}\right) + a_{s,2}\left(\frac{v_2}{v_s}\right) + \ldots + a_{s,s} + \ldots + a_{s,n}\left(\frac{v_n}{v_s}\right).$$

Hence

$$\begin{aligned}|\lambda_i - a_{s,s}| &= \left|a_{s,1}\left(\frac{v_1}{v_s}\right) + \ldots + 0 + \ldots + a_{s,n}\left(\frac{v_s}{v_s}\right)\right| \\ &\leqslant |a_{s,1}| + |a_{s,2}| + \ldots + 0 + \ldots + |a_{s,n}| \\ &= P_s,\end{aligned}$$

which completes the proof.

As an illustrative example consider the Crank–Nicolson equations with known boundary-values, namely

$$\begin{aligned}&(2\mathbf{I}_{N-1} - r\mathbf{T}_{N-1})\mathbf{u}_{j+1} \\ &\qquad = (2\mathbf{I}_{N-1} + r\mathbf{T}_{N-1})\mathbf{u}_j + \mathbf{b}_j = \{4\mathbf{I}_{N-1} - (2\mathbf{I}_{N-1} - r\mathbf{T}_{N-1})\}\mathbf{u}_j + \mathbf{b}_j,\end{aligned}$$

which can be written as

$$\mathbf{B}\mathbf{u}_{j+1} = (4\mathbf{I}_{N-1} - \mathbf{B})\mathbf{u}_j + \mathbf{b}_j,$$

giving

$$\mathbf{u}_{j+1} = (4\mathbf{B}^{-1} - \mathbf{I}_{N-1})\mathbf{u}_j + \mathbf{B}^{-1}\mathbf{b}_j,$$

where

$$\mathbf{B} = \begin{bmatrix} (2+2r) & -r & & & \\ -r & (2+2r) & -r & & \\ & \ddots & \ddots & \ddots & \\ & & -r & (2+2r) & -r \\ & & & -r & (2+2r) \end{bmatrix}$$

The equations will be stable in the Lax–Richtmyer sense when $\|4\mathbf{B}^{-1}-\mathbf{I}_{N-1}\|\leqslant 1$. Since $\mathbf{B}$ is real and symmetric it follows by Exercise 14 that $4\mathbf{B}^{-1}-\mathbf{I}_{N-1}$ is real and symmetric, so $\|4\mathbf{B}^{-1}-\mathbf{I}_{N-1}\|=\rho(4\mathbf{B}^{-1}-\mathbf{I}_{N-1})$. The stability condition will therefore be satisfied when the modulus of every eigenvalue of $4\mathbf{B}^{-1}-\mathbf{I}_{N-1}$ does not exceed one; that is, when

$$\left|\frac{4}{\lambda}-1\right|\leqslant 1, \quad \text{implying} \quad -1\leqslant\frac{4}{\lambda}-1\leqslant 1,$$

where λ is an eigenvalue of $\mathbf{B}$. This states that $\lambda\geqslant 2$.

For the matrix $\mathbf{B}$, $a_{s,s}=2+2r$, $\max P_s=2r$, so Gerschgorin's circle theorem leads to

$$|\lambda-2-2r|\leqslant 2r,$$

from which it follows that

$$-2r\leqslant\lambda-2-2r\leqslant 2r,$$

or

$$2\leqslant\lambda\leqslant 2+4r,$$

proving that the equations are unconditionally stable, since $\lambda\geqslant 2$ for all $r>0$.

Gerschgorin's circle theorem and the norm of matrix A

It should be noted that when the eigenvalues λ_i of matrix $\mathbf{A}$ are estimated by the circle theorem, the condition $|\lambda_i|\leqslant 1$ is equivalent to $\|\mathbf{A}\|_\infty\leqslant 1$ or $\|\mathbf{A}\|_1\leqslant 1$. The theorem states that

$$|\lambda-a_{s,s}|\leqslant P_s.$$

Hence,

$$-P_s\leqslant\lambda-a_{ss}\leqslant P_s,$$

so that

$$-P_s+a_{ss}\leqslant\lambda\leqslant P_s+a_{ss}.$$

The eigenvalue λ will therefore satisfy $-1\leqslant\lambda\leqslant 1$ if

$$-1\leqslant -P_s+a_{ss}\leqslant P_s+a_{ss}\leqslant 1, \quad s=1(1)N-1.$$

Remembering that P_s is the sum of the moduli of the elements of $\mathbf{A}$ along the sth row and that a_{ss} may be positive or negative, this inequality is equivalent to

$$\sum_{j=1}^{N-1} |a_{sj}| \leqslant 1, \quad s = 1(1)N-1,$$

i.e. to $\|\mathbf{A}\|_\infty \leqslant 1$ for rows or $\|\mathbf{A}\|_1 \leqslant 1$ for columns.

When both inequalities hold, it follows automatically that $\|\mathbf{A}\|_2 \leqslant 1$ because it is shown in Exercise 15 that $\|\mathbf{A}\|_2^2 \leqslant \|\mathbf{A}\|_1 \, \|\mathbf{A}\|_\infty$. Therefore, Gerschgorin's circle theorem can be used to establish conditions satisfying the Lax–Richtmyer definition of stability, a situation that frequently arises with derivative boundary-conditions.

Gerschgorin's third theorem

If p of the circles of Gerschgorin's circle theorem form a connected domain that is isolated from the other circles, then there are precisely p eigenvalues of matrix $\mathbf{A}$ within this connected domain.

In particular, an isolated Gerschgorin circle contains one eigenvalue.

A proof is given in reference 30.

Stability criteria for derivative boundary conditions

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1,$$

and the conditions,

$$\frac{\partial U}{\partial x} = h_1(U - v_1) \quad \text{at } x = 0, \quad t \geqslant 0,$$

$$\frac{\partial U}{\partial x} = -h_2(U - v_2) \quad \text{at } x = 1, t \geqslant 0,$$

where h_1, h_2, v_1, v_2 are constants, $h_1 \geqslant 0$, $h_2 \geqslant 0$.

When the boundary conditions are approximated by the central-difference equations

$$(u_{1,j} - u_{-1,j})/2\delta x = h_1(u_{0,j} - v_1),$$

$$(u_{N+1,j} - u_{N-1,j})/2\delta x = -h_2(u_{N,j} - v_2), \quad (N\delta x = 1),$$

and the differential equation by the explicit scheme

$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j},$$

elimination of $u_{-1,j}$, $u_{N+1,j}$, leads to the equations

$$\begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ \cdot \\ u_{N-1,j+1} \\ u_{N,j+1} \end{bmatrix} = \begin{bmatrix} \{1-2r(1+h_1\delta x)\} & 2r & & & \\ r & (1-2r) & r & & \\ & \cdot & \cdot & \cdot & \\ & & r & (1-2r) & r \\ & & & 2r & \{1-2r(1+h_2\delta x)\} \end{bmatrix} \times \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ \vdots \\ u_{N-1,j} \\ u_{N,j} \end{bmatrix} + \begin{bmatrix} 2rh_1v_1\delta x \\ 0 \\ \vdots \\ 0 \\ 2rh_2v_2\delta x \end{bmatrix}$$

As each component of the last column vector is a constant the matrix determining the propagation of the error is

$$\begin{bmatrix} \{1-2r(1+h_1\delta x)\} & 2r & & & \\ r & (1-2r) & r & & \\ & \cdot & \cdot & \cdot & \\ & & r & (1-2r) & r \\ & & & 2r & \{1-2r(1+h_2\delta x)\} \end{bmatrix}$$

Since the off-diagonal elements of this real matrix are one-signed, all its eigenvalues are real. (See Exercise 16.) Application of the circle theorem to this matrix, with

$$a_{ss}=1-2r(1+h_1\delta x) \quad \text{and} \quad P_s=2r,$$

shows that some of its eigenvalues λ may lie on or within the circle

$$|\lambda-\{1-2r(1+h_1\delta x)\}|\leqslant 2r.$$

Using Fig. 2.8,

$$\lambda_1=1-2r(2+h_1\delta x), \quad \lambda_2=1-2rh_1\delta x,$$

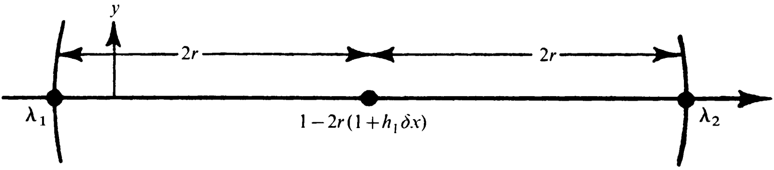


Fig. 2.8

and for stability,

$$|\lambda_1|\leqslant 1, \quad |\lambda_2|\leqslant 1.$$

Hence

$$-1\leqslant 1-2r(2+h_1\delta x)\leqslant 1, \quad \text{giving } r\leqslant 1/(2+h_1\delta x),$$

and

$$-1\leqslant 1-2rh_1\delta x, \quad \text{giving } r\leqslant 1/h_1\delta x.$$

The least of these is $r\leqslant 1/(2+h_1\delta x)$.
Similarly, we require $r\leqslant 1/(2+h_2\delta x)$.
For rows $2(1)N-1$, $a_{s,s}=1-2r$, $P_s=2r$, giving $r\leqslant \frac{1}{2}$.
For overall stability, $r\leqslant \min\{1/(2+h_1\delta x); 1/(2+h_2\delta x)\}$.

Direct application of $\|\mathbf{A}\|_\infty\leqslant 1$

As mentioned earlier, the circle theorem and $|\lambda|\leqslant 1$ is equivalent to $\|\mathbf{A}\|_\infty\leqslant 1$. This is easily verified for this example.
Row 1. If $1-2r(1+h_1\delta x)\geqslant 0$, the sum of the moduli of the terms along row $1=1-2rh_1\delta x$.

Since $\|\mathbf{A}\|_\infty\leqslant 1$ for Lax–Richtmyer stability,

$$0<1-2rh_1\delta x\leqslant 1 \quad \text{implies} \quad r<1/2h_1\delta x.$$

If $1-2r(1+h_1\delta x)\leqslant 0$, the sum of the moduli of the elements along row $1=2r(1+h_1\delta x)-1+2r=2r(2+h_1\delta x)-1$.

As before, $0<2r(2+h_1\delta x)-1\leqslant 1$ implies that $r\leqslant 1/(2+h_1\delta x)$.
Similarly for the other rows.

Crank–Nicolson equations

It is easily shown that the Crank–Nicolson equations for the problem just considered propagate any perturbation or rounding error by the recursion formula

$$\mathbf{e}_{j+1}=(4\mathbf{B}^{-1}-\mathbf{I})\mathbf{e}_j,$$

where

$$\mathbf{B}=\begin{bmatrix} \{2+2r(1+h_1\delta x)\} & -2r & & & \\ -r & (2+2r) & -r & & \\ & \cdot & \cdot & \cdot & \\ & & -r & (2+2r) & -r \\ & & & -2r & \{2+2r(1+h_2\delta x)\} \end{bmatrix}$$

As shown previously, the spectral radius of $4\mathbf{B}^{-1}-\mathbf{I}\leqslant 1$ if the eigenvalues of $\mathbf{B}\geqslant 2$. By the circle theorem,

$$|\lambda-\{2+2r(1+h_1\delta x)\}|\leqslant 2r,$$

giving

$$2+2rh_1\delta x\leqslant\lambda.$$

Similarly, the remaining rows give that $2\leqslant\lambda$ and $2+2rh_2\delta x\leqslant\lambda$. Hence the equations are unconditionally stable.

Stability condition allowing exponential growth (Constant coefficients)

If the solution of the partial differential equation increases exponentially with increasing t then the exact solution of the difference equations must do so also. In order to allow this we need to return to the original definition of stability for the equations

$$\mathbf{e}_j=\mathbf{A}^j\mathbf{e}_0,\quad j=1(1)J,\quad Jk=T,$$

which was that

$$\|\mathbf{A}^j\|\leqslant M,$$

where M is a positive number independent of j, h, and k.
Since $\|\mathbf{A}^j\|=\|\mathbf{A}\cdot\mathbf{A}^{j-1}\|\leqslant\|\mathbf{A}\|\,\|\mathbf{A}^{j-1}\|\leqslant\ldots\leqslant\|\mathbf{A}\|^j$,
the definition will be satisfied by

$$\|\mathbf{A}\|^j\leqslant M,\quad 0<j\leqslant J.$$

Assuming $M>1$,

$$\|\mathbf{A}\|\leqslant M^{1/j}=\exp(\ln M^{1/j})=\exp\left(\frac{1}{j}\ln M\right)$$

$$=1+\frac{1}{j}\ln M+\ldots$$

In particular, this must satisfy

$$\begin{aligned}\|\mathbf{A}\| &\leq 1+\frac{1}{J}\ln M+\ldots\\ &\geq 1+\frac{1}{J}\ln M\\ &= 1+\frac{k}{T}\ln M.\end{aligned}$$

In other words, a necessary and sufficient condition for stability is that

$$\|\mathbf{A}\| \leq 1+Kk=1+O(k),$$

where the positive number K is independent of h and k. (See Worked Example 2.11.) A much more rigorous treatment of this condition based on infinite sets of uniformly bounded linear operators is given in reference 25.

Stability by the Fourier series method (von Neumann's method)

Assume we are concerned with the stability of a linear two time-level difference equation in $u(x, t)$ in the time interval $0 \leq t \leq T=Jk$, T finite, as $\delta x=h \rightarrow 0$ and $\delta t=k \rightarrow 0$, i.e. as $J \rightarrow \infty$. The Fourier series or von Neumann method, first discussed in detail in reference 21, expresses the initial values at the mesh points along $t=0$ in terms of a finite Fourier series, then considers the growth of a function that reduces to this series for $t=0$ by a 'variables separable' method identical to that commonly used for solving partial differential equations.

The Fourier series can be formulated in terms of sines and cosines but the algebra is easier if the complex exponential form is used, i.e. with $\sum a_n \cos(n\pi x/l)$ or $\sum b_n \sin(n\pi x/l)$ replaced by the equivalent $\sum A_n e^{in\pi x/l}$, where $i=\sqrt{-1}$ and l is the x-interval throughout which the function is defined. Clearly we need to change our usual notation $u_{i,j}$ to $u(ph, qk)=u_{p,q}$. In terms of this notation,

$$A_n e^{in\pi x/l}=A_n e^{in\pi ph/Nh}=A_n e^{i\beta_n ph}$$

where $\beta_n=n\pi/Nh$ and $Nh=l$.

Denote the initial values at the pivotal points along $t=0$ by

$u(ph, 0) = u_{p,0}$, $p = 0(1)N$. Then the $N+1$ equations

$$u_{p,0} = \sum_{n=0}^{N} \mathbf{A}_n e^{i\beta_n ph}, \quad p = 0, 1, \ldots, N, \tag{2.33}$$

are sufficient to determine the $(N+1)$ unknowns $A_0, A_1, \ldots, A_N$ uniquely, showing that the initial mesh values can be expressed in this complex exponential form. As we are considering only linear-difference equations we need investigate the propagation of only one initial value, such as $e^{i\beta ph}$, because separate solutions are additive. The coefficient A_n is a constant and can be neglected.

To investigate the propagation of this term as t increases, put

$$u_{p,q} = e^{i\beta x} e^{\alpha t} = e^{i\beta ph} e^{\alpha qk} = e^{i\beta ph} \xi^q, \tag{2.34}$$

where $\xi = e^{\alpha k}$ and α, in general, is a complex constant. ξ is often called the amplification factor.

The finite-difference equations will be stable by the Lax–Richtmyer definition if $|u_{p,q}|$ remains bounded for all $q \leqslant J$ as $h \to 0$ and $k \to 0$, and for all values of β needed to satisfy the initial conditions.

If the exact solution of the difference equations does not increase exponentially with time, then a necessary and sufficient condition for stability is that

$$|\xi| \leqslant 1,$$

i.e.

$$-1 \leqslant \xi \leqslant 1. \tag{2.35}$$

If, however, $u_{p,q}$ does increase with t, (see Worked Example 2.11), then the necessary and sufficient condition for stability is, as shown in the preceding section,

$$|\xi| \leqslant 1 + Kk = 1 + O(k), \tag{2.36}$$

where the positive number K is independent of h, k and β.

It should be noted that this method applies only to linear equations with constant coefficients, and strictly speaking only to initial value problems with periodic initial data, of period l. For difference equations involving three or more time-levels or two or more dependent variables, the von Neumann conditions (2.35) and (2.36) are always necessary but may not be sufficient, reference 25. In practice, the method often gives useful results even when its application is not fully justified.

Example 2.10

Investigate the stability of the fully implicit difference equation

$$\frac{1}{k}(u_{p,q+1}-u_{p,q})=\frac{1}{h^2}(u_{p-1,q+1}-2u_{p,q+1}+u_{p+1,q+1})$$

approximating $\partial U/\partial t=\partial^2 U/\partial x^2$ at (ph, qk).

Substitution of $u_{p,q}=e^{i\beta ph}\xi^q$ into the difference equation shows that

$$e^{i\beta ph}\xi^{q+1}-e^{i\beta ph}\xi^{q}=r\{e^{i\beta(p-1)h}\xi^{q+1}-2e^{i\beta ph}\xi^{q+1}+e^{i\beta(p+1)h}\xi^{q+1}\},$$

where $r=k/h^2$. Division by $e^{i\beta ph}\xi^q$ leads to

$$\begin{aligned}\xi-1&=r\xi(e^{-i\beta h}-2+e^{i\beta ph})\\&=r\xi(2\cos\beta h-2)\\&=-4r\xi\sin^2(\beta h/2).\end{aligned}$$

Hence,

$$\xi=\frac{1}{1+4r\sin^2\beta h/2}.$$

Clearly, $0<\xi\leqslant 1$ for all $r>0$ and all β. Therefore the equations are unconditionally stable.

Example 2.11

Investigate the stability of the linear difference equation

$$\frac{1}{k}(u_{p,q+1}-u_{p,q})=\frac{a}{h^2}(u_{p-1,q}-2u_{p,q}+u_{p,+1,q})+bu_{p,q}$$

approximating the parabolic equation

$$\frac{\partial U}{\partial t}=a\frac{\partial^2 U}{\partial x^2}+bU$$

at the point (ph, qk), where a and b are positive constants.

The analytical solution of this equation increases with t because it is eventually dominated by the term

$$e^{bt}\int_0^1 f(x)\,dx,\quad \text{where}\quad U=f(x)\text{ when }t=0,\ 0\leqslant x\leqslant 1.$$

Substitution of $u_{p,q}=e^{i\beta ph}\xi^q$ into the difference equation leads to

$$\xi-1=ra(e^{-i\beta h}-2+e^{i\beta h})+kb,\quad r=k/h^2.$$

This leads to

$$\xi = 1 - 4ra\sin^2(\beta h/2) + kb, \quad b > 0.$$

Hence,

$$|\xi| \leqslant |1 - 4ra\sin^2(\beta h/2)| + kb$$

satisfies the von Neumann condition (2.36) if

$$|1 - 4ra\sin^2(\beta h/2)| \leqslant 1.$$

This leads to $r \leqslant 1/2a$, which is the same restriction on r as for $b = 0$. It is easily shown that the difference equation is consistent so it follows that it is also convergent.

Example 2.12

The hyperbolic equation $\partial^2 U/\partial t^2 = \partial^2 U/\partial x^2$ is approximated at (ph, qk) by the explicit difference scheme

$$(u_{p,q+1} - 2u_{p,q} + u_{p,q-1})/k^2 = (u_{p+1,q} - 2u_{p,q} + u_{p-1,q})/h^2.$$

Investigate its stability.

It is easily shown by the procedure of Example 2.10 that the equation for ξ is

$$\xi^2 - 2A\xi + 1 = 0,$$

where

$$A = 1 - 2r^2\sin^2(\beta h/2) \quad r = k/h. \tag{2.37}$$

Hence the values of ξ are

$$\xi_1 = A + (A^2 - 1)^{\frac{1}{2}} \quad \text{and} \quad \xi_2 = A - (A^2 - 1)^{\frac{1}{2}}.$$

As U does not increase exponentially with t and because the difference equation is a three time-level approximation, a necessary condition for stability is that

$$|\xi| \leqslant 1.$$

As r, k and β are real, $A \leqslant 1$ by (2.37).

When $A < -1$, $|\xi_2| > 1$, giving instability.

When

$$-1 \leqslant A \leqslant 1,\ A^2 \leqslant 1,\ \xi_1 = A + i(1 - A^2)^{\frac{1}{2}},\ \xi_2 = A - i(1 - A^2)^{\frac{1}{2}},$$

hence

$$|\xi_1| = |\xi_2| = \{A^2 + (1 - A^2)\}^{\frac{1}{2}} = 1,$$

showing that a necessary condition for stability is $-1 \leqslant A \leqslant 1$. By

(2.37),

$$-1 \leqslant 1 - 2r^2 \sin^2(\beta h/2) \leqslant 1.$$

The only useful inequality is

$$-1 \leqslant 1 - 2r^2 \sin^2(\beta h/2),$$

giving

$$r = k/h \leqslant 1.$$

In reference 25, this condition is also shown to be sufficient.

The global rounding error

For simplicity, assume that all boundary values are zero so that the finite-difference equations approximating the initial-value differential equation in the solution domain $0 < x < 1$, $t > 0$, can be written as

$$\mathbf{u}_j = \mathbf{A}\mathbf{u}_{j-1},$$

where $\mathbf{u}_0 = \mathbf{U}_0$ is the vector of known initial values and $\mathbf{A}$ is a square matrix of known elements of order $(N-1)$.

In general, the computer will not store the initial value $u_{i,0}$ exactly, but a numerical approximation $N_{i,0}$, so that

$$N_{i,0} = u_{i,0} - r_{i,0}, \quad \text{i.e.} \quad \mathbf{N}_0 = \mathbf{u}_0 - \mathbf{r}_0,$$

where $\mathbf{r}_0$ is the vector of initial rounding errors. As rounding errors will be introduced at every stage of the calculations the numerical solution values calculated by the computer at the first time-level will be

$$\mathbf{N}_1 = \mathbf{A}\mathbf{N}_0 - \mathbf{r}_1 = \mathbf{A}\mathbf{u}_0 - \mathbf{A}\mathbf{r}_0 - \mathbf{r}_1.$$

Finally, at the jth time-level, the computed solution will be

$$\mathbf{N}_j = \mathbf{A}^j\mathbf{u}_0 - \mathbf{A}^j\mathbf{r}_0 - \mathbf{A}^{j-1}\mathbf{r}_1 - \ldots - \mathbf{r}_j.$$

If there were no rounding errors the exact solution of the difference equations would be

$$\mathbf{u}_j = \mathbf{A}^j\mathbf{u}_0.$$

Hence the difference between the exact solution and the computed solution, i.e., the global rounding error $\mathbf{R}_j$, at the jth time-level is

$$\mathbf{u}_j - \mathbf{N}_j = \mathbf{A}^j\mathbf{r}_0 + \mathbf{A}^{j-1}\mathbf{r}_1 + \ldots + \mathbf{r}_j.$$

This shows that the local rounding error vector at each time-level propagates forward in the same way as the exact solution vector at that time-level. As proved earlier, the effect of each local rounding error will diminish with increasing j if $\max_i|\lambda_i|<1$, where λ_i, $i=1(1)(N-1)$, are the eigenvalues of $\mathbf{A}$, but the global rounding error cannot possibly tend to zero because of the terms $\mathbf{r}_j$, $\mathbf{Ar}_{j-1}, \ldots$.

Lax's equivalence theorem

Given a properly posed linear initial-value problem and a linear finite-difference approximation to it that satisfies the consistency condition, stability is the necessary and sufficient condition for convergence.

The proof of this theorem is beyond the scope of this book and interested readers should consult reference 25.

A simple example demonstrating the relationship between convergence, stability, and consistency

The classical explicit approximation to the heat conduction equation provides a comparatively simple illustration of Lax's equivalence theorem.

In general, a problem is properly posed if:

(i) The solution is unique when it exists.

(ii) The solution depends continuously on the initial data.

(iii) A solution always exists for initial data that is arbitrarily close to initial data for which no solution exists. (In heat flow problems, for example, discontinuous temperature distributions can be approximated by a sum of N continuous functions whose limiting value, as N tends to infinity, equals the discontinuous distribution except at the points of discontinuity.)

Let U satisfy the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad 0<x<1, \quad t>0, \tag{2.38}$$

have known continuous initial values when $t=0$, $0\leqslant x\leqslant 1$, and known continuous boundary values at $x=0$ and 1, $t>0$.

The classical explicit approximation to (2.38) is

$$\frac{u_{i,j+1}-u_{i,j}}{k}=\frac{u_{i-1,j}-2u_{i,j}+u_{i+1,j}}{h^2},\quad i=1(1)(N-1),\qquad(2.39)$$

where $x=ih$, $t=jk$, and $Nh=1$. The local truncation error of this difference scheme is defined by

$$T_{i,j}=\frac{U_{i,j+1}-U_{i,j}}{k}-\frac{U_{i-1,j}-2U_{i,j}+U_{i+1,j}}{h^2},$$

which may be written as

$$U_{i,j+1}=kT_{i,j}+rU_{i-1,j}+(1-2r)U_{i,j}+rU_{i+1,j},\qquad(2.40)$$

where $r=k/h^2$. For $i=1(1)(N-1)$, eqns (2.40) can be written in matrix form as

$$\begin{bmatrix}U_{1,j+1}\\U_{2,j+1}\\\cdot\\\cdot\\\cdot\\U_{N-1,j+1}\end{bmatrix}=k\begin{bmatrix}T_{1,j}\\T_{2,j}\\\cdot\\\cdot\\\cdot\\T_{N-1,j}\end{bmatrix}+\begin{bmatrix}(1-2r)&r&&\\r&(1-2r)&r&\\&\cdot&&\\&&\cdot&\\&&&\cdot\\&&r&(1-2r)\end{bmatrix}$$

$$\times\begin{bmatrix}U_{1,j}\\U_{2,j}\\\cdot\\\cdot\\\cdot\\U_{N-1,j}\end{bmatrix}+\begin{bmatrix}rU_{0,j}\\0\\\cdot\\\cdot\\\cdot\\0\\rU_{N,j}\end{bmatrix}$$

i.e. as

$$\mathbf{U}_{j+1}=k\mathbf{T}_j+\mathbf{A}\mathbf{U}_j+\mathbf{c}_j,\qquad(2.41)$$

where $\mathbf{c}_j$ is a vector of known boundary values. Applying this recursively,

$$\begin{aligned}\mathbf{U}_{j+1}&=k\mathbf{T}_j+\mathbf{A}(k\mathbf{T}_{j-1}+\mathbf{A}\mathbf{U}_{j-1}+\mathbf{c}_{j-1})+\mathbf{c}_j\\&=k(\mathbf{T}_j+\mathbf{A}\mathbf{T}_{j-1})+\mathbf{A}^2\mathbf{U}_{j-1}+(\mathbf{c}_j+\mathbf{A}\mathbf{c}_{j-1})\\&=\dots\\&=k(\mathbf{T}_j+\mathbf{A}\mathbf{T}_{j-1}+\dots+\mathbf{A}^j\mathbf{T}_0)+\mathbf{A}^{j+1}\mathbf{U}_0\\&\quad+(\mathbf{c}_j+\mathbf{A}\mathbf{c}_{j-1}+\dots+\mathbf{A}^j\mathbf{c}_0).\end{aligned}\qquad(2.42)$$

As the boundary and initial values for u are the same as for U, it follows from eqn (2.39), which can be written as

$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j},\ i=1(1)(N-1),$$

that

$$\mathbf{u}_{j+1}=\mathbf{A}\mathbf{u}_j+\mathbf{c}_j.$$

This leads as before to

$$\mathbf{u}_{j+1}=\mathbf{A}^{j+1}\mathbf{u}_0+(\mathbf{c}_j+\mathbf{A}\mathbf{c}_{j-1}+\ldots+\mathbf{A}^j\mathbf{c}_0). \tag{2.43}$$

Subtraction of (2.43) from (2.42) shows that

$$\mathbf{U}_{j+1}-\mathbf{u}_{j+1}=k(\mathbf{T}_j+\mathbf{A}\mathbf{T}_{j-1}+\ldots+\mathbf{A}^j\mathbf{T}_0)+\mathbf{A}^{j+1}(\mathbf{U}_0-\mathbf{u}_0).$$

But $\mathbf{U}_0=\mathbf{u}_0=$ vector of initial values. Hence

$$\mathbf{U}_{j+1}-\mathbf{u}_{j+1}=k(\mathbf{T}_j+\mathbf{A}\mathbf{T}_{j-1}+\ldots+\mathbf{A}^j\mathbf{T}_0). \tag{2.44}$$

This equation shows that the difference between the *exact* solution of the partial differential equation and the *exact* solution of the approximating difference equation at, say, the $(i, j+1)$th mesh point depends on the local truncation errors at certain mesh points on every preceding time-level and on the difference scheme used. Whether or not the accumulative effect at the $(i, j+1)$th mesh point of these preceding errors is a catastrophic build-up or a hoped-for decay as j increases depends clearly on the matrix $\mathbf{A}$ and the nature of the $T_{i,j}$, $i=1(1)N$, $j=0(1)J$, if the field of integration of the differential equation is the rectangle $0\leqslant x\leqslant 1$, $0\leqslant t=jk\leqslant T=Jk$, T finite.

Hence eqn (2.44) gives

$$\|\mathbf{U}_{j+1}-\mathbf{u}_{j+1}\|\leqslant k\{\|\mathbf{T}_j\|+\|\mathbf{A}\|\,\|\mathbf{T}_{j-1}\|+\ldots\ +\|\mathbf{A}^j\|\,\|\mathbf{T}_0\|\},\quad 0\leqslant j\leqslant J-1.$$

By the Lax–Richtmyer definition of stability,

$$\|\mathbf{A}^j\|\leqslant M,\quad j=1(1)J,$$

where M is a positive number independent of j, h and k. Therefore,

$$\|\mathbf{U}_{j+1}-\mathbf{u}_{j+1}\|\leqslant k\ \|\mathbf{T}_j\|+kM(\|\mathbf{T}_{j-1}\|+\ldots+\|\mathbf{T}_0\|).$$

If the maximum of all of the moduli of the components of $\mathbf{T}_0$, $\mathbf{T}_1,\ldots,\mathbf{T}_j$ is C, then the infinity norm of each $\mathbf{T}_s$, $s=1(1)j$, will be $\leqslant C$. Using this norm it follows that

$$\|\mathbf{U}_{j+1}-\mathbf{u}_{j+1}\|\leqslant kC+kMjC=kC+MtC.$$

But $jk=t\leqslant T=Jk$ is finite and $k\rightarrow 0$ as $J\rightarrow\infty$, so the first term of the right-hand side tends to zero irrespective of the magnitudes of M and C. The second term, however, tends to zero if and only if $C=\max\limits_{i,j}|T_{i,j}|$ tends to zero as k and $h=(k/r)^{\frac{1}{2}}$ tend to zero, which, by definition, is the condition for consistency. This proves, in this particular case, that the difference scheme is convergent when it is stable and consistent.

Finite difference approximations to $\partial U/\partial t=\nabla^2 U$ in cylindrical and spherical polar co-ordinates

The non-dimensional form of the equation for heat-conduction in three dimensions is $\partial U/\partial t=\nabla^2 U$, which, in cylindrical polar coordinates (r, θ, z) is

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial r^2}+\frac{1}{r}\frac{\partial U}{\partial r}+\frac{1}{r^2}\frac{\partial^2 U}{\partial \theta^2}+\frac{\partial^2 U}{\partial z^2}.$$

Assuming, for simplicity, that U is independent of z, this reduces to the two-dimensional equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial r^2}+\frac{1}{r}\frac{\partial U}{\partial r}+\frac{1}{r^2}\frac{\partial^2 U}{\partial \theta^2}. \tag{2.45}$$

For non-zero values of r there is no difficulty in expressing each derivative in terms of standard finite-difference approximations, as shown in Chapter 5, but at $r=0$ the right side appears to contain singularities. This complication can be dealt with by replacing the polar co-ordinate form of $\nabla^2 U$ by its Cartesian equivalent which transforms eqn (2.45) to the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}+\frac{\partial^2 U}{\partial y^2}. \tag{2.46}$$

Now construct a circle of radius δr, centre the origin, and denote the four points in which Ox, Oy meet this circle by 1, 2, 3, 4. Denote the corresponding function values by u_1, u_2, u_3, and u_4 and the value at the origin by u_0. Then

$$\nabla^2 U=\frac{(u_1+u_2+u_3+u_4-4u_0)}{(\delta r)^2}+O\{(\delta r)^2\}.$$

Rotation of the axes through a small angle clearly leads to a similar equation. Repetition of this rotation and the addition of all such equations then gives

$$\nabla^2 U = \frac{4(u_M - u_0)}{(\delta r)^2} + O\{(\delta r)^2\},$$

where u_M is a mean value of U round the circle. The best mean value available is given, of course, by adding all values and dividing by their number.

When a two-dimensional problem in cylindrical co-ordinates possesses circular symmetry, then $\partial^2 U/\partial\theta^2 = 0$, and eqn (2.45) simplifies to

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial r^2} + \frac{1}{r}\frac{\partial U}{\partial r}. \tag{2.47}$$

Assuming $\partial U/\partial r = 0$ at $r = 0$, which it will be if the problem is symmetrical with respect to the origin, it is seen that $(1/r)\partial U/\partial r$ assumes the indeterminate form 0/0 at this point.

By Maclaurin's expansion,

$$U'(r) = U'(0) + rU''(0) + \tfrac{1}{2}r^2 U'''(0) + \ldots,$$

but $U'(0) = 0$, so the limiting value of $(1/r)\partial U/\partial r$ as r tends to zero is the value of $\partial^2 U/\partial r^2$ at $r = 0$. Hence eqn (2.47) at $r = 0$ can be replaced by

$$\frac{\partial U}{\partial t} = 2\frac{\partial^2 U}{\partial r^2}. \tag{2.48}$$

This result can also be deduced from eqn (2.46) because $\partial^2 U/\partial x^2 = \partial^2 U/\partial y^2$ from the circular symmetry, and we can make the x-axis coincide with the direction of r. The finite-difference representation of (2.48) is further simplified by the condition $\partial U/\partial r = 0$ at $r = 0$ because this gives $u_{-1,j} = u_{1,j}$. For example, the explicit approximation

$$\frac{(u_{0,j+1} - u_{0,j})}{\delta t} = \frac{2(u_{1,j} - 2u_{0,j} + u_{-1,j})}{(\delta r)^2}$$

to eqn (2.48) simplifies to

$$\frac{(u_{0,j+1} - u_{0,j})}{\delta t} = \frac{4(u_{1,j} - u_{0,j})}{(\delta r)^2}. \quad \text{(See Example 2.13)}$$

A complication identical to the one above also arises at $r=0$ with the spherical polar form of $\nabla^2 U$, namely

$$\frac{\partial^2 U}{\partial r^2}+\frac{2}{r}\frac{\partial U}{\partial r}+\frac{\cot}{r}\frac{\partial U}{\partial \theta}+\frac{1}{r^2}\frac{\partial^2 U}{\partial \theta^2}+\frac{1}{r^2 \sin^2\theta}\frac{\partial^2 U}{\partial \phi^2}.$$

By the same argument as in the two-dimensional case, this can be replaced at $r=0$ by $\partial^2 U/\partial x^2+\partial^2 U/\partial y^2+\partial^2 U/\partial z^2$ and approximated by $6(u_M-u_0)/(\delta r)^2$, where u_M is the mean value of U over the sphere of radius δr, centre the origin. The factor 6 occurs because Ox, Oy, Oz meet the sphere in six points. If, however, the problem is symmetrical with respect to the origin, i.e. independent of θ and ϕ, $\nabla^2 U$ reduces to $\partial^2 U/\partial r^2+(2/r)\partial U/\partial r$, with $\partial U/\partial r$ zero at $r=0$. By either of the previous arguments it follows that the heat conduction equation at $r=0$ becomes

$$\frac{\partial U}{\partial t}=3\frac{\partial^2 U}{\partial r^2}.$$

It is of interest to note that symmetrical heat flow problems for hollow cylinders and spheres that *exclude* $r=0$ can be solved by simpler equations than those considered because the change of independent variable defined by $R=\log r$ transforms the cylindrical equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial r^2}+\frac{1}{r}\frac{\partial U}{\partial r} \quad \text{to} \quad e^{2R}\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial R^2},$$

and the change of dependent variable given by $U=w/r$ transforms the spherical equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial r^2}+\frac{2}{r}\frac{\partial U}{\partial r} \quad \text{to} \quad \frac{\partial w}{\partial t}=\frac{\partial^2 w}{\partial r^2}.$$

Example 2.13

The function U is a solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}+\frac{2}{x}\frac{\partial U}{\partial x}, \quad 0<x<1,\ t>0,$$

and satisfies the initial conditions

$$U=1-x^2 \text{ when } t=0, \quad 0\leqslant x\leqslant 1,$$

and the boundary conditions

$$\frac{\partial U}{\partial x}=0 \text{ at } x=0, \quad t>0; \quad U=0 \text{ at } x=1, \quad t>0.$$

Using a rectangular grid defined by $\delta x=0.1$ and $\delta t=0.001$, calculate a finite-difference solution to 4*D* by an explicit method at the points $(0, 0.001)$, $(0.1, 0.001)$ and $(0.9, 0.001)$ in the x–t plane. (See Chapter 2, Exercise 23, for the stability of the difference scheme.) At $x=0$, $(2/x)(\partial U/\partial x)$ is indeterminate. As

$$\lim_{x\to 0}\frac{2}{x}\frac{\partial U}{\partial x}=\lim_{x\to 0} 2\frac{\partial^2 U}{\partial x^2},$$

the equation can be replaced at $x=0$ by

$$\frac{\partial U}{\partial t}=3\frac{\partial^2 U}{\partial x^2}.$$

This may be approximated by the difference equation

$$\frac{u_{0,j+1}-u_{0,j}}{\delta t}=\frac{3(u_{-1,j}-2u_{0,j}+u_{1,j})}{(\delta x)^2}. \tag{2.49}$$

If $(\partial U/\partial x)_{i,j}$ is approximated by $(u_{i+1,j}-u_{i-1,j})/2(\delta x)$, it follows that $u_{-1,j}=u_{1,j}$ since $(\partial U/\partial x)_{0,j}=0$. Hence eqn (2.49) reduces to

$$u_{0,j+1}=u_{0,j}+3r(2u_{1,j}-2u_{0,j})=(1-6r)u_{o,j}+6ru_{1,j},$$

where $r=\delta t/(\delta x)^2=0.1$ in this example. Therefore

$$u_{0,j+1}=\tfrac{1}{5}(2u_{0,j}+3u_{1,j}). \tag{2.50}$$

At $x\neq 0$ the differential equation can be approximated by

$$\frac{1}{\delta t}(u_{i,j+1}-u_{i,j})=\frac{1}{(\delta x)^2}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})+\frac{2}{2i(\delta x)^2}(u_{i+1,j}-u_{i-1,j})$$

giving

$$u_{i,j+1}=r\left(1-\frac{1}{i}\right)u_{i-1,j}+(1-2r)u_{i,j}+r\left(1+\frac{1}{i}\right)u_{i+1,j}.$$

Therefore

$$u_{1,j+1}=\tfrac{1}{5}(4u_{1,j}+u_{2,j}) \tag{2.51}$$

and

$$u_{9,j+1}=\tfrac{4}{5}(\tfrac{1}{9}u_{8,j}+u_{9,j}) \text{ since } u_{10,j}=0. \tag{2.52}$$

By eqns (2.50), (2.51), and (2.52) the solution values at the points in question are as shown below

$x=$	0	0.1	0.2	...	0.8	0.9
$t=0$	1.0000	0.9900	0.9600	...	0.3600	0.1900
$t=0.001$	0.9940	0.9840				0.1840

Exercises and solutions

F.D.S. *means finite difference solution.*
A.S. *means analytical solution of the partial differential equation.*

1. Calculate a finite-difference solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2} \quad (0<x<1, t>0)$$

satisfying the initial condition

$$U=\sin \pi x \text{ when } t=0 \text{ for } 0\leqslant x\leqslant 1,$$

and the boundary condition

$$U=0 \text{ at } x=0 \text{ and } 1 \text{ for } t>0,$$

using an explicit method with $\delta x=0.1$ and $r=0.1$.

Show by the method of separation of the variables, or merely verify, that the analytical solution is $U=e^{-\pi^2 t}\sin \pi x$. Hence check the accuracy of the numerical solution for $t=0.005$.

Solutions (Tables 2.18–2.21)

Solutions for $r=0.1$ and 0.5 and comparisons with the analytical solution at $x=0.5$ are given overpage. They show clearly that when the initial function and all its derivatives are continuous, and the boundary values at $(0, 0)$ and $(1, 0)$ remain equal to the initial values at these points, then the finite-difference solution can be very accurate indeed. Only one-half of the solution is shown because the problem is symmetric with respect to $x=0.5$.

2. Calculate a numerical solution of the equation $\partial U/\partial t=\partial^2 U/\partial x^2$, $0<x<1$, satisfying the initial condition $U=1$ when $t=0$, $0<x<1$, and the boundary condition $U=0$ at $x=0$ and 1, $t\geqslant 0$.

TABLE 2.18

	t	$x=0$	0.1	0.2	0.3	0.4	0.5
(F.D.).	0.005	0	0.2942	0.5596	0.7702	0.9054	0.9520
(A.S.)	0.005	0	0.2941	0.5595	0.7701	0.9053	0.9519
	...	.					
	...	.					
(F.D.S.)	0.01	0	0.2801	0.5327	0.7332	0.8602	0.9063
(A.S.)	0.01	0	0.2800	0.5325	0.7330	0.8617	0.9060
	...	.					
F.D.S.	0.02	0	0.2538	0.4828	0.6645	0.7812	0.8214
(A.S.)	0.02	0	0.2537	0.4825	0.6641	0.7807	0.8209
	...	.					
	...	.					
(F.D.S.)	0.10	0	0.1156	0.2198	0.3025	0.3556	0.3739
(A.S.)	0.10	0	0.1152	0.2191	0.3015	0.3545	0.3727

TABLE 2.19

	Finite-difference solution ($x=0.5$)	Analytical solution ($x=0.5$)	Percentage error
$t=0.005$	0.9520	0.9519	negligible
$t=0.01$	0.9063	0.9060	negligible
$t=0.02$	0.8214	0.8209	negligible
$t=0.10$	0.3739	0.3727	0.3

r = 0.5 $u_{i,j+1}=\frac{1}{2}(u_{i-1,j}+u_{i+1,j})$.

TABLE 2.20

	t	$x=0$	0.1	0.2	0.3	0.4	0.5
(F.D.S.)	0.005	0	0.2939	0.5590	0.7694	0.9045	0.9511
(F.D.S.)	0.010	0	0.2795	0.5317	0.7318	0.8602	0.9045
	.	0					
(F.D.S.)	0.02	0	0.2528	0.4809	0.6619	0.7781	0.8181
	.						
	.						
	.						
(F.D.S.)	0.10	0	0.1133	0.2154	0.2965	0.3486	0.3665
(A.S.)	0.10	0	0.1152	0.2191	0.3015	0.3545	0.3727

TABLE 2.21

	Finite-difference solution ($x=0.5$)	Analytical solution ($x=0.5$)	Percentage error
$t=0.005$	0.9511	0.9519	−0.08
$t=0.01$	0.9045	0.9060	−0.17
$t=0.02$	0.8181	0.8209	−0.34
$t=0.10$	0.3665	0.3727	−1.66

2 (continued)

Show by the method of separation of the variables that the analytical solution is

$$U=\frac{4}{\pi}\sum_{n=0}^{\infty}\frac{1}{(2n+1)}\,\mathrm{e}^{-(2n+1)^2\pi^2 t}\sin(2n+1)\pi x.$$

Compare these solutions at $x=0.1$ for small values of t. (See the comments below the tables in the solution on p. 82.)

This problem concerns the temperature changes in a uniform heat-insulated rod that is initially at a constant temperature and which is cooled by having its ends reduced to zero temperature at zero time and subsequently kept at zero.

Solution

The finite-difference solution given by eqn (2.6) with $r=0.1$ is shown in Table 2.22, and the analytical solution in Table 2.23. Only one-half the solution is given because the problem is symmetrical with respect to $x=\frac{1}{2}$.

TABLE 2.22. Finite-difference solution

	$x=0$	0.1	0.2	0.3	0.4	0.5
$t=0.000$	0	1.0000	1.0000	1.0000	1.0000	1.0000
0.001	0	0.9000	1.0000	1.0000	1.0000	1.0000
0.002	0	0.8200	0.9900	1.0000	1.0000	1.0000
0.005	0	0.6566	0.9335	0.9927	0.9996	1.0000
0.010	0	0.5113	0.8283	0.9566	0.9919	0.9979
0.050	0	0.2429	0.4589	0.6263	0.7313	0.7669
0.100	0	0.1460	0.2776	0.3820	0.4490	0.4721
0.200	0	0.0546	0.1038	0.1428	0.1679	0.1766

TABLE 2.23. Analytical solution

$x=0$		0.1	0.2	0.3	0.4	0.5
$t=0.000$	0	1.0000	1.0000	1.0000	1.0000	1.0000
0.001	0	0.9747	1.0000	1.0000	1.0000	1.0000
0.002	0	0.8862	0.9984	1.0000	1.0000	1.0000
0.005	0	0.6827	0.9545	0.9973	0.9999	1.0000
0.010	0	0.5205	0.8427	0.9661	0.9953	0.9992
0.050	0	0.2442	0.4616	0.6304	0.7363	0.7723
0.100	0	0.1467	0.2790	0.3839	0.4513	0.4745
0.200	0	0.0547	0.1040	0.1431	0.1682	0.1769

Comments

An obvious difficulty arises in the solution domain at (0, 0) because the limiting value of the initial temperature is unity as x tends to zero, whereas the limiting value of the boundary temperatures is zero as t tends to zero. In other words the temperature is discontinuous at (0, 0) and its value could equally well have been chosen as 1 or $\frac{1}{2}$ instead of zero. Because of this discontinuity the finite-difference solution is a poor one near $x=0$ for small values of t. It will be noticed, however, on comparing Tables 2.22 and 2.23, that the accuracy of the finite-difference solution near $x=0$ improves as t increases. This is characteristic of parabolic equations and indicates that an implicit method would give a better solution in the neighbourhood of (0, 0) than an explicit one because it would not draw its information exclusively from the first row and column. In general, however, we cannot calculate an accurate solution near a point of discontinuity by finite-difference methods unless we remove the discontinuity by a suitable transformation such as below. An alternative approach is to calculate an analytical solution that is continuous near the discontinuity.

For the problem above it is possible to remove the discontinuity at (0, 0) by changing the independent variables from (x, t) to (X, T) by means of the equations

$$X = xt^{-\frac{1}{2}}; \quad T = t^{\frac{1}{2}}.$$

This expands the origin $x=t=0$ into the positive half of the X-axis and collapses the positive half of the x-axis into the point at infinity along the x-axis. Consequently the discontinuity in U at $x = t = 0$ is transformed into a smooth change along the positive half of the X-axis. Further details are given in reference 8.

3. Derive the Crank–Nicolson equations for the problem in Exercise 1 and solve them directly for at least two time-steps.

Evaluate the corresponding analytical solution and calculate the percentage error in the numerical solution.

Solution

The solution for $x = 0(0.1)0.5$ and $r = 1$ is given in Table 2.24.

TABLE 2.24

	t	$x=0$	0.1	0.2	0.3	0.4	0.5
(F.D.S.)	0.01	0	0.2802	0.5329	0.7335	0.8623	0.9067
(A.S.)	0.01	0	0.2800	0.5325	0.7330	0.8617	0.9060
(F.D.S.)	0.02	0	0.2540	0.4832	0.6651	0.7818	0.8221
(A.S.)	0.02	0	0.2537	0.4825	0.6641	0.7807	0.8209
(F.D.S.)	0.10	0	0.1160	0.2207	0.3037	0.3571	0.3754
(A.S.)	0.10	0	0.1152	0.2191	0.3015	0.3545	0.3727

TABLE 2.25

	Finite-difference solution ($x = 0.5$)	Analytical solution ($x = 0.5$)	Percentage error
$t = 0.01$	0.9067	0.9060	0.08
0.02	0.8221	0.8209	0.15
0.10	0.3754	0.3727	0.72

4. Prove that the back-substitution procedure of the non-pivoting elimination algorithm given on p. 24 for solving a tridiagonal system of linear equations is stable when $a_i > 0$, $b_i > 0$, $c_i > 0$, and $b_i > a_i + c_i$, $i = 1, 2, \ldots, (N-1)$, where $a_1 = c_{N-1} = 0$.

Solution

$$u_i = \frac{1}{\alpha_i}(S_i + c_i u_{i+1}) = p_{i+1}u_{i+1} + \frac{S_i}{\alpha_i}, \quad \text{(say)},$$

where

$$\alpha_1 = b_1, \quad \alpha_i = b_i - \frac{a_i c_{i-1}}{\alpha_{i-1}}, \quad S_1 = d_1, \quad S_i = d_i + \frac{a_i S_{i-1}}{\alpha_{i-1}},$$

$$a_1 = 0 = c_{N-1}, \quad i = 1, 2, \ldots, N-1.$$

There will be no build-up of errors in the back-substitution process if $|p_{i+1}| < 1$,
where

$$p_{i+1} = \frac{c_i}{\alpha_i} = \frac{c_i}{b_i - a_i p_i}, \quad i = 1(1)N-1).$$

Now $p_2 = (c_1/b_1)$, since $a_1 = 0$. Also $b_1 > c_1$ by hypothesis.
Therefore, $0 < p_2 < 1$.

$$p_3 = \frac{c_2}{b_2 - a_2 p_2}.$$

As $c_2 > 0$, $0 < p_2 < 1$, and $b_2 > a_2 > 0$, it follows that

$$0 < p_3 < \frac{c_2}{b_2 - a_2}.$$

By hypothesis, $b_2 > a_2 + c_2$. Hence

$$0 < p_3 < \frac{c_2}{(a_2 + c_2) - a_2} = 1.$$

Similarly, $0 < p_4, p_5, \ldots, p_{N-1} < 1$.

5. Use the Gaussian elimination algorithm on p. 24 to calculate a finite-difference approximation to Worked Example 2.1 for one time-step taking $\delta x = 0.1$ and $r = 1$ using either

(i) the fully implicit backward time-difference equation

$$u_{i,j+1} - u_{i,j} = r(u_{i-1,j+1} - 2u_{i,j+1} + u_{i+1,j+1}),$$

or (ii) the Douglas equations (3.45) of Chapter 3.

Solution

(i) $$-ru_{i-1,j+1} + (1+2r)u_{i,j+1} - ru_{i+1,j+1} = u_{i,j}.$$

The equations for the first time-step are $3u_1 - u_2 = 0.2$,

$-u_1+3u_2-u_3=0.4$, $-u_2+3u_3-u_4=0.6$, $-u_3+3u_4-u_5=0.8$, and $-2u_4+3u_5=1.0$.

$\alpha_1=3$, $\alpha_2=2.66667$, $\alpha_3=2.62500$, $\alpha_4=2.61905$, $\alpha_5=2.23636$. $S_1=0.2$, $S_2=0.46667$, $S_3=0.77500$, $S_4=1.09524$, $S_5=1.83636$. $u_5=S_5/\alpha_5$ and $u_i=(S_i+c_iu_{i+1})/\alpha_i$ giving $u_5=0.8211$, $u_4=0.7317$, $u_3=0.5740$, $u_2=0.3902$, $u_1=0.1967$.

(ii) $$(1-6r)u_{i-1,j+1}+(10+12r)u_{i,j+1}+(1-6r)u_{i+1,j+1}$$
$$=(1+6r)u_{i-1,j}+(10-12r)u_{i,j}+(1-6r)u_{i+1,j}.$$

The equations for the first time-step are $22u_1-5u_2=2.4$, $-5u_1+22u_2-5u_3=4.8$, $-5u_2+22u_3-5u_4=7.2$, $-5u_3+22u_4-5u_5=9.6$, and $-10u_4+22u_5=9.2$.

$\alpha_1=22$, $\alpha_2=20.86364$, $\alpha_3=20.80174$, $\alpha_4=20.79818$, $\alpha_5=19.59594$.

$S_1=2.4$, $S_2=5.34545$, $S_3=8.48104$, $S_4=11.63855$, $S_5=14.79594$.

$u_5=0.7551$, $u_4=0.7411$, $u_3=0.5858$, $u_2=0.3966$, and $u_1=0.1992$.

6. The function U satisfies the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}\ (0<x<1,\ t>0)$$

and the boundary conditions

$$\frac{\partial U}{\partial x}=h_1(U-v_1) \text{ at } x=0, \quad \frac{\partial U}{\partial x}=-h_2(U-v_2) \text{ at } x=1,$$

where h_1, h_2, v_1, v_2 are positive constants.

(a) When the boundary conditions are approximated by central differences (see Worked Example 2.3) show that one explicit difference scheme is

$$u_{0,j+1}=\{1-2r(1+h_1\delta x)\}u_{0,j}+2ru_{1,j}+2rh_1v_1\delta x,$$
$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j}, \quad (i=1,2,\ldots,N-1),$$
$$u_{N,j+1}=2ru_{N-1,j}+\{1-2r(1+h_2\delta x)\}u_{N,j}+2rh_2v_2\delta x,$$

where $N\delta x=1$ and $r=\delta t/(\delta x)^2$.

(b) When the boundary conditions are approximated by forward-differences at $x=0$ and backward-differences at $x=1$ (see Worked Example 2.4), show that another explicit difference scheme is

$$\begin{aligned}
u_{1,j+1} &= \{1-2r+r/(1+h_1\delta x)\}u_{1,j}+ru_{2,j}+rh_1v_1\delta x/(1+h_1\delta x),\\
u_{0,j+1} &= (u_{1,j+1}+h_1v_1\delta x)/(1+h_1\delta x),\\
u_{i,j+1} &= ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j}, \quad (i=2,3,\ldots,N-2),\\
u_{N-1,j+1} &= \{1-2r+r/(1+h_2\delta x)\}u_{N-1},\\
&\quad +ru_{N-2,j}+rh_2v_2\delta x/(1+h_2\delta x),\\
u_{N,j+1} &= (u_{N-1,j+1}+h_2v_2\delta x)/(1+h_2\delta x).
\end{aligned}$$

7. A uniform solid rod of one-half a unit of length is thermally insulated along its length and its initial temperature at zero time is 0 °C. One end is thermally insulated and the other supplied with heat at a steady rate. Show that the subsequent temperatures at points within the rod are given, in non-dimensional form, by the solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2} \quad (0<x<\tfrac{1}{2},\, t>0)$$

satisfying the initial condition

$$U=0 \text{ when } t=0 \quad (0\leqslant x\leqslant \tfrac{1}{2}),$$

and the boundary conditions

$$\frac{\partial U}{\partial x}=0 \text{ at } x=0, \quad t>0, \quad \frac{\partial U}{\partial x}=f \text{ at } x=\tfrac{1}{2}, \quad t>0$$

where f is a constant.

Solve this problem numerically for $f=1$ using
(a) an explicit method with $\delta x=0.1$ and $r=\frac{1}{4}$,
(b) an implicit method with $\delta x=0.1$ and $r=1$.

Solution

The solution given by the explicit method of Worked Example 2.3, for which the equations are

$$\begin{aligned}
u_{0,j+1} &= \tfrac{1}{2}(u_{0,j}+u_{1,j}),\\
u_{i,j+1} &= \tfrac{1}{4}(u_{i-1,j}+2u_{i,j}+u_{i+1,j}) \quad (i=1,2,3,4),\\
u_{5,j+1} &= \tfrac{1}{2}(u_{4,j}+u_{5,j}+0.1),
\end{aligned}$$

TABLE 2.26. Explicit method

$x=$	0	0.1	0.2	0.3	0.4	0.5
$t=0.005$	0.0000	0.0000	0.0000	0.0000	0.0125	0.0750
0.0075	0.0000	0.0000	0.0000	0.0031	0.0250	0.0938
0.01	0.0000	0.0000	0.0008	0.0078	0.0367	0.1094
0.02	0.0009	0.0027	0.0103	0.0313	0.0767	0.1571
0.03	0.0062	0.0104	0.0248	0.0554	0.1095	0.1934
0.05	0.0291	0.0364	0.0594	0.1007	0.1636	0.2509
0.10	0.1169	0.1265	0.1556	0.2044	0.2735	0.3631
0.20	0.3150	0.3250	0.3550	0.4050	0.4750	0.5650
0.50	0.9150	0.9250	0.9550	1.0050	1.0750	1.1650
1.00	1.9150	1.9250	1.9550	2.0050	2.0750	2.1650

is recorded in Table 2.26. The Crank–Nicolson solution, as in the Worked Example 2.5, is shown in Table 2.27. The analytical solution of the differential equation is

$$U = 2t + \frac{1}{2}\left\{\frac{12x^2-1}{6} - \frac{2}{\pi^2}\sum_{n=1}^{\infty}\frac{(-1)^n}{n^2}\,e^{-4\pi^2 n^2 t}\cos 2n\pi x\right\}$$

$$= 2\sqrt{t}\sum_{n=0}^{\infty}\left\{i\,\mathrm{erfc}\frac{(2n+1-2x)}{4\sqrt{t}} + i\,\mathrm{erfc}\frac{(2n+1+2x)}{4\sqrt{t}}\right\},$$

and is evaluated in Table 2.28. Comparisons are made in Table 2.29.

The finite-difference solutions are clearly very accurate except for small values of t.

A simple calculation shows that the effect of the exponential

TABLE 2.27. Crank–Nicolson method

$x=$	0	0.1	0.2	0.3	0.4	0.5
$t=0.01$	0.0003	0.0006	0.0022	0.0083	0.0309	0.1155
0.02	0.0023	0.0039	0.0108	0.0302	0.0770	0.1540
0.03	0.0077	0.0115	0.0252	0.0552	0.1080	0.1925
0.05	0.0301	0.0373	0.0597	0.1004	0.1627	0.2499
0.10	0.1172	0.1268	0.1557	0.2043	0.2732	0.3628
0.20	0.3150	0.3250	0.3550	0.4050	0.4750	0.5650
0.50	0.9150	0.9250	0.9550	1.0050	1.0750	1.1650
1.00	1.9150	1.9250	1.9550	2.0050	2.0750	2.1650

TABLE 2.28. Analytical solution

$x =$	0	0.1	0.2	0.3	0.4	0.5
$t = 0.0025$	0.0000	0.0000	0.0000	0.0001	0.0050	0.0564
0.0050	0.0000	0.0000	0.0001	0.0017	0.0167	0.0798
0.0075	0.0000	0.0000	0.0006	0.0053	0.0286	0.0977
0.01	0.0000	0.0002	0.0017	0.0101	0.0399	0.1128
0.02	0.0016	0.0035	0.0117	0.0333	0.0791	0.1596
0.03	0.0074	0.0117	0.0264	0.0573	0.1115	0.1954
0.05	0.0307	0.0381	0.0610	0.1023	0.1653	0.2526
0.10	0.1186	0.1282	0.1573	0.2061	0.2751	0.3647
0.20	0.3167	0.3267	0.3567	0.4067	0.4766	0.5666
0.50	0.9167	0.9267	0.9567	1.0067	1.0767	1.1667
1.00	1.9167	1.9267	1.9567	2.0067	2.0767	2.1667

component of the analytical solution is negligible for values of t in excess of 0.1.

A point of interest in this example is that the difference between the analytical solution and both finite-difference solutions for values of t in excess of 0.1 is $0.0017 = 0.01/6 = (\delta x)^2/6$. It can be proved that the transient component of the solution of any explicit or implicit finite-difference scheme for a parabolic equation satisfying the boundary conditions above does not tend to zero as t increases, as does the transient, i.e. exponential component of the solution of the differential equation, but tends to a value of $k(\delta x)^2$, k constant. In this example $k = \frac{1}{6}$.

TABLE 2.29

t	Analytical solution $x = 0.3$	Explicit solution $x = 0.3$	Percentage error	Crank–Nicolson solution $x = 0.3$	Percentage error
0.0075	0.0053	0.0031	−41.5		
0.01	0.0101	0.0078	−22.8	0.0083	−17.8
0.05	0.1023	0.1007	−1.56	0.1004	−1.85
0.10	0.2061	0.2044	−0.82	0.2043	−0.87
0.50	1.0067	1.0050	−0.17	1.0050	−0.17
1.00	2.0067	2.0050	−0.08	2.0050	−0.08

8. The function U satisfies the equation

$$\frac{\partial U}{\partial t}=x\frac{\partial^2 U}{\partial x^2},\quad 0<x<\tfrac{1}{2},\quad t>0,$$

the boundary conditions

$$U=0 \text{ at } x=0,\quad t>0,\quad \frac{\partial U}{\partial x}=-\tfrac{1}{2}U \text{ at } x=\tfrac{1}{2},\quad t>0,$$

and the initial condition $U=x(1-x)$ when $t=0$, $0\leqslant x\leqslant\frac{1}{2}$.

When all the derivatives with respect to x are expressed in terms of central-difference formulae show that the simplest explicit difference equations approximating this problem at the point (ih, jk) in the $x-t$ plane can be written as

$$u_{i,j+1}=irhu_{i-1,j}+(1-2irh)u_{i,j}+irhu_{i+1,j},\quad i=1(1)(N-1),$$

and

$$u_{N,j+1}=2Nrhr_{N-1,j}+(1-2Nrh-Nrh^2)u_{N,j},$$

where $r=k/h^2$ and $Nh=\frac{1}{2}$.

Take $h=0.1$ and $r=0.5$ and calculate a numerical solution to 4D at the points along the first time-level corresponding to $i=3$ and 5. Will the solution at points near the point $(\frac{1}{2}, 0)$ be very accurate? Give a reason for your answer. (The stability of these equations is considered in Chapter 2, Exercise 25.)

Solution

The simplest explicit approximation is

$$\frac{u_{i,j+1}-u_{i,j}}{k}=\frac{ih(u_{i-1,j}-2u_{i,j}+u_{i+1,j})}{h^2},$$

giving

$$u_{i,j+1}=irhu_{i-1,j}+(1-2irh)u_{i,j}+irhu_{i+1,j},\quad i=1(1)(N-1).$$

Mentally extend the interval of integration $0\leqslant x\leqslant\frac{1}{2}$ to $0\leqslant x\leqslant\frac{1}{2}+h$. Then the previous equation holds for $i=N$. Eliminate $u_{N+1,j}$ by means of the approximating boundary condition equation

$$(u_{N+1,j}-u_{N-1,j})/2h=-\tfrac{1}{2}u_{N,j}$$

to give

$$u_{N,j+1}=2Nrhu_{N-1,j}+(1-2Nrh-Nrh^2)u_{N,j}.$$

For $h=0.1$ and $r=\frac{1}{2}$,

$$u_{i,j+1}=0.05iu_{i-1,j}+(1-0.1i)u_{i,j}+0.05iu_{i+1,j}, \quad i=1(1)4,$$

and

$$u_{5,j+1}=0{\cdot}5u_{4,j}+0{\cdot}475u_{5,j}.$$

$i=$	0	1	2	3	4	5
$t=0$	0	0.0900	0.1600	0.2100	0.2400	0.2500
$t=0.005$	0			0.2070		0.2388

As $(x,0)\rightarrow(\frac{1}{2},0)$, $\partial U/\partial x=1-2x\rightarrow 0$. At $(\frac{1}{2},0)$, $U=\frac{1}{2}(1-\frac{1}{2})=\frac{1}{4}$. As $(\frac{1}{2},t)\rightarrow(\frac{1}{2},0)$, $\partial U/\partial x=-\frac{1}{2}U\rightarrow-\frac{1}{2}\cdot\frac{1}{4}=-\frac{1}{8}$. Therefore $\partial U/\partial x$ is discontinuous at $(\frac{1}{2},0)$ so the finite-difference solution will not be very accurate at mesh points near $(\frac{1}{2},0)$. The solution of the differential equation is continuous at $(\frac{1}{2},0)$ and the effect of this discontinuous derivative on the difference solution is very small a few mesh lengths from $(\frac{1}{2},0)$.

9. The function V satisfies the non-linear equation

$$\frac{\partial V}{\partial t}=\frac{\partial^2 V}{\partial x^2}+\left(\frac{\partial V}{\partial x}\right)^2, \quad 0<x<1, \quad t>0,$$

the initial condition $V=0$ when $t=0$, $0\leqslant x\leqslant 1$, and the boundary conditions

$$\frac{\partial V}{\partial x}=1 \text{ at } x=0,\ t>0; \quad V=0 \text{ at } x=1, \quad t>0.$$

Show that the change of dependent variable defined by $V=\log U$, $U\neq 0$, transforms the problem to the solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad 0<x<1, \quad t>0,$$

where U satisfies the conditions

$$U=1 \text{ when } t=0, \quad 0\leqslant x\leqslant 1,$$

$$\frac{\partial U}{\partial x}=U \text{ at } x=0,\ t>0; \quad U=1 \text{ at } x=1, \quad t>0.$$

Using a rectangular mesh defined by $\delta x = 0.1$ and $\delta t = 0.0025$, approximate the equation for U by the classical explicit scheme, the derivative boundary condition being approximated by a central-difference formula.

Hence calculate a numerical solution for V at the points $(0, 0.005)$ and $(0, 0.0075)$ in the $x-t$ plane.

Solution

$$\frac{\partial V}{\partial t}=\frac{1}{U}\frac{\partial U}{\partial t};\quad \frac{\partial V}{\partial x}=\frac{1}{U}\frac{\partial U}{\partial x};\quad \frac{\partial^2 V}{\partial x^2}=-\frac{1}{U^2}\left(\frac{\partial U}{\partial x}\right)^2+\frac{1}{U}\frac{\partial^2 U}{\partial x^2}.$$

By $V=\log U$, $U=1$, when $V=0$. As $\partial V/\partial x=(1/U)(\partial U/\partial x)$ and $\partial V/\partial x=1$ at $x=0$, $t>0$, therefore $\partial U/\partial x = U$ at $x=0$, $t>0$. The approximation equation is

$$u_{i,j+1}=\tfrac{1}{4}(u_{i-1,j}+2u_{i,j}+u_{i+1,j}),\quad i=1(1)9.$$

Putting $i=0$ and eliminating $u_{-1,j}$ by means of the derivative boundary equation $(u_{1,j}-u_{-1,j})/0.2=u_{0,j}$, leads to $u_{0,j+1}=0.45u_{0,j}+0.5u_{1,j}$.

$x=$	0	0.1	0.2
$t=0$	1.0000	1.0000	1.0000
$t=0.0025$	0.9500	1.0000	1.0000
$t=0.0050$	0.9275	0.9875	1.0000
$t=0.0075$	0.9111		

$V(0, 0.005)=\log_e 0.9275=-0.0753$. $V(0, 0.0075)=\log_e 0.9111=-0.0931$.

10. The equation

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0$$

is approximated at the point (ih, jk) by the difference equation

$$\theta\left(\frac{u_{i,j+1}-u_{i,j-1}}{2k}\right)+(1-\theta)\left(\frac{u_{i,j}-u_{i,j-1}}{k}\right)-\frac{1}{h^2}\,\delta_x^2 u_{i,j}=0.$$

Show that the truncation error at this point is given by

$$-\tfrac{1}{2}k(1-\theta)\frac{\partial^2 U}{\partial t^2}-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}+O(k^2, h^4).$$

Hence find the value of θ that will reduce this error to one of order k^2 and h^4.

Solution

$$T_{i,j}=\theta(U_{i,j+1}-U_{i,j-1})/2k+(1-\theta)(U_{i,j}-U_{i,j-1})/k \\ -(U_{i-1,j}-2U_{i,j}+U_{i+1,j})/h^2.$$

Expand each term by Taylor's series about (ih, jk) to get

$$T_{i,j}=-\tfrac{1}{2}k(1-\theta)\frac{\partial^2 U_{i,j}}{\partial t^2}-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}+O(k^2)+O(h^4),$$

where

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},$$

so that

$$\frac{\partial^2 U}{\partial t^2}=\frac{\partial^4 U}{\partial x^4}.$$

Hence,

$$T_{i,j}=\{-\tfrac{1}{2}k(1-\theta)-\tfrac{1}{12}h^2\}\frac{\partial^4 U_{i,j}}{\partial x^4}+O(k^2)+O(h^4).$$

Therefore,

$$\tfrac{1}{2}k(1-\theta)=-\tfrac{1}{12}h^2,\text{ i.e. }\theta=1+\frac{h^2}{6k}$$

gives the required result.

11. Show that the local truncation error at the point (ih, jk) of the Crank–Nicolson approximation to $\partial U/\partial t=\partial^2 U/\partial x^2$ is $O(h^2)+O(k^2)$.

Solution

$$T_{i,j}=\frac{1}{k}(U_{i,j+1}-U_{i,j})-\frac{1}{2h^2}(\delta_x^2 U_{i,j+1}+\delta_x^2 U_{i,j}),$$

where, by Taylor's expansion about the point (ih, jk),

$$U_{i+1,j}=\left[U+h\frac{\partial U}{\partial x}+\tfrac{1}{2}h^2\frac{\partial^2 U}{\partial x^2}+\tfrac{1}{6}h^3\frac{\partial^3 U}{\partial x^3}+\tfrac{1}{24}h^4\frac{\partial^4 U}{\partial x^4}+O(h^6)\right]_{i,j}$$

etc., giving

$$\delta_x^2 U_{i,j}=\left[h^2\frac{\partial^2 U}{\partial x^2}+\tfrac{1}{12}h^4\frac{\partial^4 U}{\partial x^4}+O(h^6)\right]_{i,j}.$$

Again, by Taylor's expansion,

$$\delta_x^2 U_{i,j+1}=\delta_x^2 U_{i,j}+k\frac{\partial}{\partial t}\delta_x^2 U_{i,j}+O(k^2).$$

Therefore,

$$T_{i,j}=\left[\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right]_{i,j}+\tfrac{1}{2}k\frac{\partial}{\partial t}\left[\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}\right]_{i,j}$$

$$-\tfrac{1}{12}h^2\left(\frac{\partial^4 U}{\partial x^4}\right)_{i,j}+O(k^2)+O(h^4)+O(kh^2),$$

where

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0.$$

Hence the result.

12. The equation

$$\alpha\frac{\partial U}{\partial t}+\frac{\partial U}{\partial x}-f(x,t)=0,\quad \alpha \text{ constant},$$

is approximated at the point (ih, jk) in the x–t plane by the difference scheme

$$\frac{\alpha}{k}\{u_{i,j+1}-\tfrac{1}{2}(u_{i+1,j}+u_{i-1,j})\}+\frac{1}{2h}(u_{i+1,j}-u_{i-1,j})-f_{i,j}=0.$$

Investigate the consistency of this scheme for (a) $k=rh$, and (b) $k=rh^2$, r a positive constant, where it is assumed that U is sufficiently smooth for third-order derivatives in x and second-order derivatives in t to exist.

If either is inconsistent with the differential equation obtain the equation it does approximate.

Solution

$$U_{i,j+1}=U_{i,j}+k\frac{\partial U_{i,j}}{\partial t}+\tfrac{1}{2}k^2\frac{\partial^2 U_{i,j+\theta_1}}{\partial t^2},$$

$$U_{i+1,j}=U_{i,j}+h\frac{\partial U_{ij}}{\partial x}+\tfrac{1}{2}h^2\frac{\partial^2 U_{i,j}}{\partial x^2}+\tfrac{1}{6}h^3\frac{\partial^3 U_{i+\theta_2,j}}{\partial x^3},$$

$$U_{i-1,j}=U_{i,j}-h\frac{\partial U_{i,j}}{\partial x}+\tfrac{1}{2}h^2\frac{\partial^2 U_{i,j}}{\partial x^2}-\tfrac{1}{6}h^3\frac{\partial^3 U_{i-\theta_3,j}}{\partial x^3},$$

where $0<\theta_1, \theta_2, \theta_3<1$. Hence

$$T_{i,j}=F_{i,j}(U)=\left(\alpha\frac{\partial U}{\partial t}+\frac{\partial U}{\partial x}-f\right)_{i,j}+\tfrac{1}{2}\alpha k\frac{\partial^2 U_{i,j+\theta_1}}{\partial t^2}$$
$$-\frac{\alpha h^2}{2k}\frac{\partial^2 U_{i,j}}{\partial x^2}-\frac{\alpha h^3}{12k}\left(\frac{\partial^3 U_{i+\theta_2,j}}{\partial x^3}-\frac{\partial^3 U_{i-\theta_3,j}}{\partial x^3}\right)+O(h^2).$$

(a) $k=rh$. Scheme is consistent.
(b) $k=rh^2$. The difference equation approximates

$$\alpha\frac{\partial U}{\partial t}+\frac{\partial U}{\partial x}-\frac{\alpha}{2r}\frac{\partial^2 U}{\partial x^2}-f=0$$

as $h\to 0$.

13. Prove that when the explicit finite-difference scheme

$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j},\quad \text{where } r=\frac{k}{h^2},$$

is used to approximate the equation $\partial U/\partial t=\partial^2 U/\partial x^2$, and it is assumed that U possesses continuous and finite derivatives up to order three in t and order six in x, then the discretization error is the solution of the difference equation

$$e_{i,j+1}=re_{i-1,j}+(1-2r)e_{i,j}+re_{i+1,j}+k\omega(x,t),$$

where

$$\omega(x,t)=\frac{h^2}{12}\left(6r\frac{\partial^2 U}{\partial t^2}-\frac{\partial^4 U}{\partial x^4}\right)_{i,j}+\frac{k^2}{6}\frac{\partial^3 U(x_i, t_j+\theta_j k)}{\partial t^3}$$
$$-\frac{h^4}{360}\frac{\partial^6 U(x_i+\theta_i h, t_j)}{\partial x^6},\quad -1<\theta_i<1 \quad\text{and}\quad 0<\theta_j<1.$$

If the maximum value of $|\omega|$ is M, deduce, for $0<r\leqslant\frac{1}{2}$, that $|e_{i,j}|\leqslant tM$. Hence show that the discretization error is of order h^2, except when $r=\frac{1}{6}$ in which case it is of order h^4.

Solution

As in the text, p. 46.

14. (a) If the $n\times n$ matrices $\mathbf{A}$ and $\mathbf{B}$ are symmetric, prove that $\mathbf{AB}$ and $\mathbf{BA}$ are symmetric if and only if $\mathbf{A}$ and $\mathbf{B}$ commute.

(b) If the non-singular matrix $\mathbf{A}$ is symmetric, prove that $\mathbf{A}^{-1}$ is also symmetric.

(c) If the non-singular matrices $\mathbf{A}$ and $\mathbf{B}$ commute, prove that $\mathbf{A}^{-1}$ and $\mathbf{B}$, $\mathbf{A}$ and $\mathbf{B}^{-1}$, and also $\mathbf{A}^{-1}$ and $\mathbf{B}^{-1}$ commute.

(d) If the $n\times n$ non-singular and symmetric matrices $\mathbf{A}$ and $\mathbf{B}$ commute, prove that $\mathbf{A}^{-1}\mathbf{B}$, $\mathbf{AB}^{-1}$ and $\mathbf{A}^{-1}\mathbf{B}^{-1}$ are symmetric.

Solution

(a) Assume $\mathbf{AB}=\mathbf{BA}$. Since $\mathbf{A}$ and $\mathbf{B}$ are symmetric, $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T=\mathbf{BA}=\mathbf{AB}$ by hypothesis. Hence $\mathbf{AB}$ is symmetric. Similarly for $\mathbf{BA}$. Assume $\mathbf{AB}$ is symmetric. Since $\mathbf{A}$ and $\mathbf{B}$ are symmetric, $(\mathbf{AB})^T=\mathbf{B}^T\mathbf{A}^T=\mathbf{BA}$. But $(\mathbf{AB})^T=\mathbf{AB}$ by hypothesis. Hence $\mathbf{AB}=\mathbf{BA}$, proving $\mathbf{A}$ and $\mathbf{B}$ commute.

(b) $\mathbf{I}=\mathbf{A}^{-1}\mathbf{A}=(\mathbf{AA}^{-1})^T=(\mathbf{A}^{-1})^T\mathbf{A}^T=(\mathbf{A}^{-1})^T\mathbf{A}$, since $\mathbf{A}^T=\mathbf{A}$. Hence $(\mathbf{A}^{-1})^T=\mathbf{A}^{-1}$, proving $\mathbf{A}^{-1}$ is symmetric.

(c) $\mathbf{AB}=\mathbf{BA}$, so $\mathbf{A}^{-1}(\mathbf{AB})\mathbf{A}^{-1}=\mathbf{A}^{-1}(\mathbf{BA})\mathbf{A}^{-1}$. Therefore $(\mathbf{A}^{-1}\mathbf{A})\mathbf{BA}^{-1}=\mathbf{A}^{-1}\mathbf{B}(\mathbf{AA}^{-1})$, proving that $\mathbf{BA}^{-1}=\mathbf{A}^{-1}\mathbf{B}$. Similarly for the other pairs.

(d) By (c), $\mathbf{A}^{-1}\mathbf{B}=\mathbf{BA}^{-1}$. Therefore $(\mathbf{A}^{-1}\mathbf{B})^T=(\mathbf{BA}^{-1})^T=(\mathbf{A}^{-1})^T\mathbf{B}^T$. But $\mathbf{A}^{-1}$ is symmetric by (b) and $\mathbf{B}$ is symmetric. Hence $(\mathbf{A}^{-1})^T\mathbf{B}^T=\mathbf{A}^{-1}\mathbf{B}$, proving that $\mathbf{A}^{-1}\mathbf{B}$ is symmetric. Similarly for the other pairs.

15. (a) Prove that the infinity norm of matrix $\mathbf{A}$ is equal to the maximum row sum of the moduli of the elements of $\mathbf{A}$, assuming $\|\mathbf{x}\|_\infty=1$.

(b) Assuming that $\|\mathbf{A}\|_2^2=\rho(\mathbf{A}^H\mathbf{A})$, prove that $\|\mathbf{A}\|_2^2\leqslant\|\mathbf{A}\|_1\,\|\mathbf{A}\|_\infty$. (For derivative boundary condition problems this can sometimes be used to establish stability in the 2-norm.)

Solution

(a) $\|\mathbf{A}\|_\infty = \max\limits_{\|\mathbf{x}\|_\infty=1} \|\mathbf{Ax}\|_\infty$ where

$$\|\mathbf{Ax}\|_\infty = \max_i \left|\sum_{j=1}^n a_{ij}x_j\right| \leqslant \max_i \sum_{j=1}^n |a_{ij}|\,|x_j| \leqslant \max_i \sum_{j=1}^n \left\{|a_{ij}| \max_j |x_j|\right\}.$$

But $\|\mathbf{x}\|_\infty = \max\limits_j |x_j| = 1$. Hence, $\|\mathbf{Ax}\|_\infty \leqslant \max\limits_i \sum_{j=1}^n |a_{ij}| =$ max row sum of the moduli of the elements of **A**. If the maximum row sum is given by $i = k$ then $\|\mathbf{Ax}\|_\infty$ will attain this max value by choosing $x_j = 1$, when $a_{kj} \geqslant 0$ and $x_j = -1$ when $a_{kj} < 0$.

(b) $\|\mathbf{A}\|^2 = \rho(\mathbf{A}^H\mathbf{A})$. By Gerschgorin's first theorem, $\rho(\mathbf{A}^H\mathbf{A}) \leqslant \|\mathbf{A}^H\mathbf{A}\|_1 \leqslant \|\mathbf{A}^H\|_1\,\|\mathbf{A}\|_1 = \|(\bar{\mathbf{A}})^T\|_1\,\|\mathbf{A}\|_1 = \|\mathbf{A}\|_\infty\,\|\mathbf{A}\|_1$.

16. Prove that a real tridiagonal matrix with either all its off-diagonal elements positive or all its off-diagonal elements negative is similar to a real symmetric tridiagonal matrix with non-zero off-diagonal elements. Deduce that the eigenvalues of such a matrix are real.

Solution

$$\mathbf{A} = \begin{bmatrix} a_1 & c_2 & & & \\ b_2 & a_2 & c_3 & & \\ \cdot & b_3 & a_3 & c_4 & \\ & & & \cdot & \\ & & & & \cdot \\ & & & b_n & a_n \end{bmatrix}$$

Let **D** be a real diagonal matrix with elements $d_1, d_2, \ldots, d_n$, i.e. $\mathbf{D} = \operatorname{diag}(d_1, d_2, \ldots, d_n)$. Calculate $\mathbf{DAD}^{-1}$. This matrix will be symmetric if

$$\frac{d_1^2}{d_2^2} = \frac{b_2}{c_2},\quad \frac{d_2^2}{d_3^2} = \frac{b_3}{c_3},\ \ldots,\ \frac{d_{n-1}^2}{d_n^2} = \frac{b_n}{c_n}.$$

Each right-hand side is positive. Assign a real value to d_1, then $d_2, d_3, \ldots, d_n$ are determined. Matrix **A** and the real symmetric $\mathbf{DAD}^{-1}$ are similar so they have the same eigenvalues. Hence the eigenvalues of **A** are real because the eigenvalues of a real symmetric matrix are real.

17. If the real non-symmetric matrix **A** is similar to the real symmetric matrix **B**, prove that the equations $\mathbf{u}_{j+1}=\mathbf{A}\mathbf{u}_j+\mathbf{b}_j$ are stable in the Lax–Richtmyer sense if $\|\mathbf{B}^j\|\leqslant M$, $j=1,2,\ldots$, where M is a positive number independent of h, k, and j. Show, in particular, that the equations are stable if $\rho(\mathbf{B})\leqslant 1$, but that the errors could be large. (See also Chapter 2, Exercise 16.)

Solution

Since **A** is similar to **B**, there exists a non-singular matrix **P** such that $\mathbf{B}=\mathbf{P}^{-1}\mathbf{A}\mathbf{P}$. Hence, $\mathbf{A}=\mathbf{P}\mathbf{B}\mathbf{P}^{-1}$, $\mathbf{A}^2=(\mathbf{P}\mathbf{B}\mathbf{P}^{-1})(\mathbf{P}\mathbf{B}\mathbf{P}^{-1})=\mathbf{P}\mathbf{B}^2\mathbf{P}^{-1}$ and $\mathbf{A}^j=\mathbf{P}\mathbf{B}^j\mathbf{P}^{-1}$. Therefore, $\|\mathbf{A}^j\|\leqslant\|\mathbf{P}\|\,\|\mathbf{B}^j\|\,\|\mathbf{P}^{-1}\|$. As $\|\mathbf{P}\|$ and $\|\mathbf{P}^{-1}\|$ are finite and $\|\mathbf{B}^j\|\leqslant M$, therefore $\|\mathbf{A}^j\|$ is bounded for all j, h, and k, proving stability. Using the 2-norm, $\|\mathbf{B}^j\|_2\leqslant\|\mathbf{B}\|_2^j=[\rho(\mathbf{B})]^j\leqslant 1$ by hypothesis and because **B** is real and symmetric. Therefore, $\|\mathbf{A}^j\|_2\leqslant\|\mathbf{P}\|_2\,\|\mathbf{P}^{-1}\|_2$ which is a finite number that could be large as j increases.

18. The equation

$$\frac{\partial U}{\partial t}=\alpha\frac{\partial^2 U}{\partial x^2}-\beta U,\quad 0<x<1,\quad t>0,$$

where α and β are real positive constants, is approximated at the point (ih, jk) by the explicit difference scheme

$$\frac{1}{k}\Delta_t u_{i,j}=\frac{\alpha}{h^2}\delta_x^2 u_{i,j}-\beta u_{i,j}.$$

Given that U has known continuous initial values throughout the interval $0\leqslant x\leqslant 1$, $t=0$, known boundary-values at $x=0$ and 1, $t>0$, and that $Nh=1$, find an upper bound for $r=k/h^2$ that will be sufficient to keep the difference equations stable, assuming that U does not increase with t.

Verify that Gerschgorin's circle theorem establishes the same sufficient condition for stability.

Solution

$$u_{i,j+1}-u_{i,j}=r\alpha(u_{i-1,j}-2u_{i,j}+u_{i+1,j})-k\beta u_{i,j}.$$

Hence $\mathbf{u}_{j+1}=\mathbf{A}\mathbf{u}_j+\mathbf{b}_j$, where $r=k/h^2$,

$$\mathbf{A}=\begin{bmatrix} (1-2r\alpha-k\beta & r\alpha & & \\ r\alpha & (1-2r\alpha-k\beta) & r\alpha & \\ & & \ddots & \\ & & r\alpha & (1-2r\alpha-k\beta) \end{bmatrix}$$

and $\mathbf{b}_j$ is a column vector of known values. The equations will be stable when $\|\mathbf{A}\|\leqslant 1$.

Since $\mathbf{A}$ is real and symmetric, $\|\mathbf{A}\|_2=\rho(\mathbf{A})=\max\limits_s|1-k\beta-4r\alpha\sin^2 s\pi/2N|$, $s=1(1)N-1$. The largest value for r is given by $-1\leqslant 1-rh^2\beta-4r\alpha\sin^2(N-1)\pi/2N$, which implies $r\leqslant 2/(4\alpha+\beta h^2)$. The 1-norm leads to the same result if $1-2r\alpha-k\beta\leqslant 0$, for then $\|\mathbf{A}\|_1=2r\alpha+k\beta-1+2r\alpha=4r\alpha+rh^2\beta-1\leqslant 1$ for stability, etc. If $1-2r\alpha-k\beta>0$, then $0<\|\mathbf{A}\|_1=1-2r\alpha-k\beta+2r\alpha=1-rh^2\beta\leqslant 1$, which implies that $r<1/h^2\beta>2/(4\alpha+h^2\beta)$.

19. The equation

$$\frac{\partial U}{\partial t}=a\frac{\partial^2 U}{\partial x^2},\quad 0<x<1,\quad t>0$$

where a is a positive constant, is approximated at the point (ih, jk) by the fully implicit backward-difference (or backward Euler) scheme

$$u_{i,j+1}=ra(u_{i-1,j+1}-2u_{i,j+1}+u_{i+1,j+1}),$$

where $r=k/h^2$ and $Nh=1$. Assuming that the boundary and initial values are known, prove that:

(a) The scheme is unconditionally stable in the Lax–Richtmyer sense.

(b) The local truncation error is $O(k)+O(h^2)$.

Comment

This is called a backward difference scheme because the time-difference relative to the time of the space-difference is a backward one. It is not as accurate as the Crank–Nicolson method because the truncation error for the latter is $O(k^2)+O(h^2)$.

Solution

(a) Show that $\mathbf{u}_{j+1}=(\mathbf{I}-ra\mathbf{T}_{N-1})^{-1}\mathbf{u}_j+\mathbf{b}_j$. The matrix is real and symmetric so

$$\|\mathbf{A}\|_2=\rho(\mathbf{A})=1/\{1+4ra\sin^2\pi/2N\}<1$$

for all $r>0$.

(b) $$T_{i,j}=k\left(\frac{1}{2}\frac{\partial^2U}{\partial t^2}-a\frac{\partial^3U}{\partial x^2\,\partial t}\right)_{i,j}-\tfrac{1}{12}ah^2\left(\frac{\partial^4U}{\partial x^4}\right)_{i,j}$$
$$+k^2\left(\frac{1}{6}\frac{\partial^3U}{\partial t^3}-\tfrac{1}{2}a\frac{\partial^4U}{\partial x^2\,\partial t^2}\right)_{i,j}+\ldots.$$

20. The equation $\partial U/\partial t=\partial^2U/\partial x^2$, $0<x<1$, $t>0$, is approximated by the difference scheme

$$u_{i,j+1}-u_{i,j}=r\{\theta(u_{i-1,j+1}-2u_{i,j+1}+u_{i+1,j+1})$$
$$+(1-\theta)(u_{i-1,j}-2u_{i,j}+u_{i+1,j})\},$$

where $0\leqslant\theta\leqslant1$, $r=k/h^2$ and $Nh=1$. Assuming that the initial values and boundary values are known, show that:

(a) The scheme is unconditionally stable in the Lax–Richtmyer sense for $\frac{1}{2}\leqslant\theta\leqslant1$ and stable for $0\leqslant\theta<\frac{1}{2}$ when $r\leqslant1/2(1-2\theta)$.

(b) The von Neumann method gives exactly the same results as (a).

Solution

(a) The matrix $\mathbf{A}$ of the equations is

$(\mathbf{I}-r\theta\mathbf{T}_{N-1})^{-1}\{\mathbf{I}+r(1-\theta)\mathbf{T}_{N-1}\}$. Since $(\mathbf{I}-r\theta\mathbf{T}_{N-1})$ and $\{\mathbf{I}+r(1-\theta\mathbf{T}_{N+1}\}$ are both symmetric and commute, matrix $\mathbf{A}$ is symmetric. (Exercise 14(d)). Hence its 2-norm is equal to its spectral radius. The equations will be stable when

$$-1\leqslant\max_s\frac{1-4r(1-\theta)\sin^2(s\pi/2N)}{1+4r\theta\sin^2(s\pi/2N)}\leqslant1.$$

The right-hand inequality is automatically satisfied for $r>0$, $0\leqslant\theta\leqslant1$. The left-hand inequality gives $2r(1-2\theta)\leqslant1$. Hence the result.

(b) Replacing $u_{p,q}$ by $e^{i\beta ph}\xi^q$ leads to

$$\xi=\{1-4r(1-\theta)\sin^2\beta h/2\}/\{1+4r\theta\sin^2\beta h/2\}.$$

Because $u_{p,q}$ does not increase exponentially with q, the condition for stability is that $-1 \leqslant \xi \leqslant 1$, etc.

21. A set of linear difference equations is given by $\mathbf{u}_{j+1} = \mathbf{A}\mathbf{u}_j + \mathbf{b}_j$, where the non-singular $N \times N$ matrix $\mathbf{A}$ has N distinct non-zero eigenvalues λ_i, $i = 1(1)N$, and a corresponding non-singular matrix of eigenvectors $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N]$. Show that for the matrix method of analysis in which h, k, and N are kept constant, the stability condition $\|\mathbf{A}^j\| \leqslant M$, where M is a positive number independent only of j, is equivalent to $|\lambda_i| \leqslant 1$, $i = 1(1)N$. (N.B. For the Lax–Richtmyer definition of stability M must be independent of j, h, and k.)

Solution

Any perturbation $\mathbf{e} = \mathbf{u} - \mathbf{u}^*$ satisfies $\mathbf{e}_j = \mathbf{A}^j \mathbf{e}_0$. Hence $\|\mathbf{e}_j\| \leqslant \|\mathbf{A}^j\|\,\|\mathbf{e}_0\|$ and will be bounded if $\|\mathbf{A}^j\| \leqslant M$, a constant that is independent of j, since h and k are fixed.

The set of all equations $\mathbf{A}\mathbf{x}_i = \lambda_i \mathbf{x}_i$, $i = 1(1)N$, can be written as $\mathbf{AX} = \mathbf{XD}$, where $\mathbf{D} = \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_N)$. (See Chapter 3, Eigenvector–eigenvalue solution). Hence $\mathbf{A} = \mathbf{XDX}^{-1}$, $\mathbf{A}^2 = \mathbf{XDX}^{-1}\mathbf{XDX}^{-1} = \mathbf{XD}^2\mathbf{X}^{-1}$ and $\mathbf{A}^j = \mathbf{XD}^j\mathbf{X}^{-1}$. Therefore, $\|\mathbf{A}^j\| \leqslant \|\mathbf{X}\|\,\|\mathbf{D}^j\|\,\|\mathbf{X}^{-1}\|$. But the $1, 2$, and ∞ norms of the diagonal matrix $\mathbf{D}^j = \operatorname{diag}(\lambda_1^j, \lambda_2^j, \ldots, \lambda_N^j)$ are $\max_i |\lambda_i|^j$. Therefore, a sufficient condition for $\|\mathbf{A}^j\| \leqslant M$ is that $\max_i |\lambda_i| \leqslant 1$.

22. (a) Show that one explicit finite-difference scheme approximating the equation $\partial U/\partial t = \partial^2 U/\partial x^2$ is

$$u_{i,j+1} = r u_{i-1,j} + (1-2r) u_{i,j} + r u_{i+1,j}.$$

Given that the initial values and boundary values are known, show that Gerschgorin's first and second theorems establish stability of this scheme for $r \leqslant \frac{1}{2}$, but give no useful result for $r > \frac{1}{2}$.

(b) Show that Gerschgorin's first theorem is inadequate for establishing the unconditional stability of the Crank–Nicolson equations approximating the equation $\partial U/\partial t = \partial^2 U/\partial x^2$, the boundary and initial values being assumed known.

Solution

(a) Gerschgorin's theorem gives $|\lambda| \leqslant 2|r| + |1-2r|$. Consider $0 \leqslant r \leqslant \frac{1}{2}$. Then $|\lambda| \leqslant 2r + (1-2r) = 1$. Hence the finite-difference equations are stable. $r>\frac{1}{2}$ leads to $|\lambda|\leqslant 4r-1>1$, which is inconclusive. The circle theorem gives $|\lambda-1+2r|\leqslant 2r$, since $r>0$. Hence $-2r\leqslant\lambda-1+2r\leqslant 2r$, giving $1-4r\leqslant\lambda\leqslant 1$. The equations will be stable when $-1\leqslant 1-4r$, i.e., $r\leqslant\frac{1}{2}$. When $r>\frac{1}{2}$, $\lambda\geqslant 1-4r<-1$, which is inconclusive.

(b) As shown in the text, the necessary condition for stability is $\lambda\geqslant 2$, where λ is an eigenvalue of $\mathbf{B}$. Gerschgorin's first theorem leads to $|\lambda|\leqslant 2+4r$, where $r>0$, which is insufficient to establish $\lambda\geqslant 2$.

23. The function U satisfies the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}+\frac{2}{x}\frac{\partial U}{\partial x},\quad 0<x<1,\quad t>0,$$

the initial condition $U=1-x^2$ when $t=0$, $0\leqslant x\leqslant 1$, and the boundary conditions

$$\frac{\partial U}{\partial x}=0 \text{ at } x=0,\ t>0;\quad U=0 \text{ at } x=1,\ t>0.$$

The partial differential equation is approximated at the point (ih, jk) in the x–t plane by the explicit difference equation

$$\frac{1}{k}\Delta_t u_{i,j}=\frac{1}{h^2}\delta_x^2 u_{i,j}+\frac{1}{xh}(\Delta_x u_{i,j}+\nabla_x u_{i,j}),$$

and the limiting form of the partial differential equation at $x=0$ is approximated in a similar manner. The derivative boundary condition is subsequently approximated by the usual central-difference formula.

Given that $Nh=1$ and that $k=rh^2$, $r>0$, show that the matrix A of the difference equations $u_{j+1}=Au_j$ is

$$A=\begin{bmatrix} (1-6r) & 6r & & & & & \\ 0 & (1-2r) & 2r & & & & \\ & r(1-\frac{1}{2}) & (1-2r) & r(1+\frac{1}{2}) & & & \\ & & & \cdot\quad\cdot\quad\cdot & & & \\ & & & r\left(1-\frac{1}{i}\right) & (1-2r) & r\left(1+\frac{1}{i}\right) & \\ & & & & & \cdot\quad\cdot\quad\cdot & \\ & & & & & r\left(1-\frac{1}{N-1}\right) & (1-2r) \end{bmatrix}$$

Deduce that one eigenvalue of A is $(1-6r)$. Derive from this result a necessary upper bound M on r for possible stability of the equations. Deduce that for $r \leqslant M$ the infinity norm of matrix A is 1. Using the largest possible value for r and taking $h=0.1$, calculate the solution values denoted by $u_{0,1}$, $u_{2,1}$, and $u_{9,1}$.

Solution

$$\frac{1}{k}(u_{i,j+1}-u_{i,j})=\frac{1}{h^2}(u_{i-1,j}-2u_{i,j}+u_{i+1,j})+\frac{1}{ih^2}(u_{i+1,j}-u_{i-1,j}),$$

giving

$$u_{i,j+1}=r\left(1-\frac{1}{i}\right)u_{i-1,j}+(1-2r)u_{i,j}+r\left(1+\frac{1}{i}\right)u_{i+1,j}.$$

At $x=0$, $\partial U/\partial t=3\partial^2 U/\partial x^2$ which can be approximated by $(u_{0,j+1}-u_{0,j})/k=3(u_{-1,j}-2u_{0,j}+u_{1,j})/h^2$ where $u_{-1,j}=u_{1,j}$ because $\partial U/\partial x=0$ at $x=0$. In matrix form,

$$\mathbf{u}_{j+1}=\begin{bmatrix} (1-6r) & 6r & & & \\ 0 & (1-2r) & 2r & & \\ & & & & \\ & r\left(1-\frac{1}{i}\right) & (1-2r) & r\left(1+\frac{1}{i}\right) & \\ & & & & \\ & & & r\left(1-\frac{1}{N-1}\right) & (1-2r) \end{bmatrix}\mathbf{u}_j, \text{ i.e.,}$$

$\mathbf{u}_{j+1}=\mathbf{A}\mathbf{u}_j$, where $\mathbf{u}_j^{\mathrm{T}}=(u_{0,j}, u_{1,j}, \ldots, u_{N-1,j})$. Expansion of the determinant of $(\mathbf{A}-\lambda\mathbf{I})$ by the first column gives $(1-6r)$ as an eigenvalue. For this eigenvalue, $-1 \leqslant 1-6r \leqslant 1$ for stability. Hence $r \leqslant \frac{1}{3}$. When $0<r\leqslant\frac{1}{3}$ every term in the matrix is positive and the sum of the moduli of the terms along every row except the last is 1. For the last row,

$$0<\sum_j |a_{N-1,j}|=1-r\left(1+\frac{1}{N-1}\right)<1.$$

Taking $r=\frac{1}{3}$, $h=0.1$, $u_{0,j+1}=-u_{0,j}+2u_{1,j}$ and $u_{i,j+1}=$

$\frac{1}{3}\{(1-(1/i))u_{i-1,j}+u_{i,j}+(1+(1/i))u_{i+1,j}\}$, $i=1(1)9$. Hence $u_{0,1}=0.98$, $u_{2,1}=0.94$, and $u_{9,1}=0.17$.

24. Prove that the equations of Exercise 6(b), Chapter 2, are stable in the Lax–Richtmyer sense for $0<r\leqslant\frac{1}{2}$.

Solution

$u_{0,j+1}$ and $u_{N,j+1}$ are given in terms of $u_{1,j+1}$ and $u_{N-1,j+1}$ respectively, so any errors in their solution are not propagated forward in time. Hence we need consider only the remaining equations, namely,

$$\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \\ \\ u_{N-1,j+1} \end{bmatrix} =$$

$$\begin{bmatrix} \left\{1-2r+\frac{r}{1+h_1\delta x}\right\} & r & & & \\ r & (1-2r) & r & & \\ & \ddots & & & \\ & & r & (1-2r) & r \\ & & & r & \left\{1-2r+\frac{r}{1+h_2\delta x}\right\} \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \\ \\ u_{N-1,j} \end{bmatrix} + \mathbf{c}$$

where $\mathbf{c}$ is a column vector of constants.

If $1-2r\geqslant 0$, i.e. $0<r\leqslant\frac{1}{2}$, the sum of the moduli of the elements along rows 2 to $N-2$ is 1 and along row $1=1-r+r/(1+h_1\delta x)=1-rh_1\delta x/(1+h_1\delta x)<1$ for $r>0$, $h_1>0$. Similarly for row $N-1$. Hence $\|\mathbf{A}\|_\infty=1$, proving that the equations are stable for $0<r\leqslant\frac{1}{2}$. When $r>\frac{1}{2}$, the sum of the moduli of the terms along rows 2 to $N-2=4r-1>1$. Therefore $\|\mathbf{A}\|_\infty>1$ and the equations will be unstable.

Comment

As mentioned in the text, Gerschgorin's circle theorem combined with $|\lambda|\leqslant 1$ is equivalent to $\|\mathbf{A}\|_\infty \leqslant 1$, and this method can be used to establish stability for $0<r\leqslant\frac{1}{2}$ as in the next exercise.

25. Prove that the solution and rounding errors of the equations in Exercise 8, Chapter 2, will not increase exponentially with increasing time if $r\leqslant 2/(2+\frac{1}{2}h)$.

Solution

The $N\times N$ matrix $\mathbf{A}$ of the equations is

$$\begin{bmatrix} (1-2rh) & rh & & & \\ 2rh & (1-4rh) & 2rh & & \\ & & \ddots & & \\ & & irh & (1-2irh) & irh \\ & & & r & (1-r-\frac{1}{2}rh) \end{bmatrix}$$

We can proceed either as in Exercise 24, or use Gerschgorin's circle theorem together with $|\lambda|\leqslant 1$, which is equivalent to $\|\mathbf{A}\|_\infty\leqslant 1$, as follows.

The circle theorem applied to row N gives that $|\lambda-(1-r-\frac{1}{2}rh)|\leqslant r$, i.e. $1-2r-\frac{1}{2}rh\leqslant\lambda\leqslant 1-\frac{1}{2}rh$. The right-hand inequality merely shows that $\lambda<1$ for r, $h>0$. The condition $-1\leqslant\lambda$ will be satisfied by $-1\leqslant 1-2r-\frac{1}{2}rh$, giving $r\leqslant 2/(2+\frac{1}{2}h)$. Similarly, row 1 gives $r\leqslant 2/3h$ and rows 2 to $N-1$ that $r\leqslant 2/4ih$, $i=2(1)N-1$, where $Nh=\frac{1}{2}$. The smallest range is $0<r\leqslant 2/(2+\frac{1}{2}h)$.

26. Use the Fourier series method to prove that

(a) $$u_{p,q+1}-u_{p,q}=r(u_{p-1,q}-2u_{p,q}+u_{p+1,q})$$

is stable for $0<r\leqslant\frac{1}{2}$. (The forward-difference explicit approximation to $\partial U/\partial t=\partial^2U/\partial x^2$.)

(b) $$u_{p,q+1}-u_{p,q-1}=2r(u_{p-1,q}-2u_{p,q}+u_{p+1,q})$$

is *unstable* for all positive values of r. (The central-difference explicit approximation to $\partial U/\partial t=\partial^2U/\partial x^2$, often called Richardson's method.)

(c) $$u_{p,q+1}-2u_{p,q}+u_{p,q-1}$$
$$=\tfrac{1}{2}r^2\{(u_{p+1,q+1}-2u_{p,q+1}+u_{p-1,q+1})$$
$$+(u_{p+1,q-1}-2u_{p,q-1}+u_{p-1,q-1})\}$$

is stable for all positive values of $r=\delta t/\delta x$. (As implicit approximation to the hyperbolic equation $\partial^2U/\partial t^2=\partial^2U/\partial x^2$.)

Solution

(a) $\xi=1-2r(1-\cos\beta h)$ where $|\xi|\leqslant 1$. This gives $r\leqslant 1/(1-\cos\beta h)$,
The least value of the right side is $\frac{1}{2}$.

(b) $\xi^2+8r\xi\sin^2(\beta h/2)-1=0$, so $\xi_1=-1/\xi_2$ and $\xi_1+\xi_2=-8r\sin^2(\beta h/2)$. For stability $|\xi_1|\leqslant 1$ and $|\xi_2|\leqslant 1$. When $|\xi_1|<1$, $|\xi_2|>1$, giving instability. When $\xi_1=1$, $\xi_2=-1$, and $\xi_1+\xi_2=0$ giving $r=0$.

(c) $\xi^2-2A\xi+1=0$ where $A=1/\{1+2r^2\sin^2(\beta h/2)\}>1$. Hence $\xi=A\pm i(1-A^2)^{\frac{1}{2}}$, giving $|\xi|=1$ for all real values of r.

27. The equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}+\frac{1}{x}\frac{\partial U}{\partial x},\quad 0<x<1,\quad t>0,$$

is approximated at the point (ph, qk) by the difference equation

$$\frac{1}{k}\Delta_t u_{p,q}=\frac{1}{h^2}\delta_x^2 u_{p,q}+\frac{1}{2xh}(\Delta_x u_{p,q}+\nabla_x u_{p,q}).$$

Use the von Neumann method of analysis to show that the difference equations are stable for $x>0$ when

$$\frac{k}{h^2}\leqslant\frac{2}{4+p^{-2}}.$$

Evaluate the form of the differential equation at $x=0$ given that $\partial U/\partial x=0$ at $x=0$, $t>0$.

Given also that U is constant at $x=1$, that $Nh=1$, and that the derivative boundary condition at $x=0$ is approximated by a central-difference formula, write out in matrix form the corresponding explicit difference equations approximating the partial differential equation and associated boundary conditions.

Given that $k/h^2 \leqslant \frac{1}{4}$, deduce that the errors will not increase exponentially with increasing q.

Solution

$$\frac{1}{k}(u_{p,q+1}-u_{p,q})=\frac{1}{h^2}(u_{p+1,q}-2u_{p,q}+u_{p-1,q})+\frac{1}{2ph^2}(u_{p+1,q}-u_{p-1,q}).$$

The error function $E_{p,q}=e^{i\beta ph}\xi^q$, $i=\sqrt{(-1)}$, satisfies

$$\xi=(1-4r\sin^2\tfrac{1}{2}\beta h)+i\frac{r}{p}\sin\beta h, \quad r=\frac{k}{h^2}, \quad p\geqslant 1.$$

$$|\xi|^2=1+16r^2\sin^4\tfrac{1}{2}\beta h-4r\sin^2\tfrac{1}{2}\,\beta h\left(2-\frac{r}{p^2}\cos^2\tfrac{1}{2}\beta h\right).$$

$$|\xi|\leqslant 1 \text{ if } 4r\sin^2\tfrac{1}{2}\beta h\left(2-\frac{r}{p^2}\cos^2\tfrac{1}{2}\beta h\right)\geqslant 16r^2\sin^4\tfrac{1}{2}\beta h,$$

giving

$$r\leqslant\frac{2}{4\sin^2\frac{1}{2}\beta h+\dfrac{1}{p^2}\cos^2\frac{1}{2}\beta h}>\frac{2}{4+p^{-2}}.$$

At $x=0$, $\partial U/\partial t=2\partial^2U/\partial x^2$ and this can be approximated by $(u_{0,q+1}-u_{0,q})/k=2(u_{-1,q}-2u_{0,q}+u_{1,q})/h^2$. As $\partial U/\partial x=0$ at $x=0$, $u_{1,q}=u_{-1,q}$. Hence $u_{0,q+1}=(1-4r)u_{0,q}+4ru_{1,q}$. In matrix form $\mathbf{u}_{q+1}=\mathbf{A}\mathbf{u}_q+\mathbf{c}_q$, where $\mathbf{c}_q$ is a column of constants and

$$\mathbf{A}=\begin{bmatrix}
(1-4r) & 4r & & & \\
(1-\frac{1}{2})r & (1-2r) & (1+\frac{1}{2})r & & \\
 & \cdot & & & \\
 & & \cdot & & \\
\left(1-\dfrac{1}{2p}\right)r & (1-2r) & \left(1+\dfrac{1}{2p}\right)r & & \\
 & & \cdot & & \\
 & & & \cdot & \\
 & & \left(1-\dfrac{1}{2(N-1)}\right)r & & (1-2r)
\end{bmatrix}.$$

When $r\leqslant\frac{1}{4}$ the sum of the moduli of the terms along each row $\leqslant 1$. Hence $\|\mathbf{A}\|_\infty\leqslant 1$.

28. The function U satisfies the diffusion–convection equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}-\lambda\frac{\partial U}{\partial x}, \quad 0<x<1, t>0, \lambda>0,$$

and the conditions

$$U(x,0)=0, \quad 0<x<1,$$

$$U(0,t)=1, \quad \frac{\partial U(1,t)}{\partial x}=0, \quad t>0.$$

If the equation is approximated at (ph, qk) by the central-difference scheme

$$\text{(a)} \qquad \frac{u_{p,q+1}-u_{p,q}}{k}=\frac{u_{p+1,q}-2u_{p,q}+u_{p-1,q}}{h^2}-\lambda\frac{u_{p+1,q}-u_{p-1,q}}{2h}$$

and by the 'upwind' scheme

$$\text{(b)} \qquad \frac{u_{p,q+1}-u_{p,q}}{k}=\frac{u_{p+1,q}-2u_{p,q}+u_{p-1,q}}{h^2}-\lambda\frac{u_{p,q}-u_{p-1,q}}{h},$$

use the von Neumann method to show that the Lax–Richtmyer definition of stability is satisfied by $0<r\leqslant\frac{1}{2}$ for Scheme (a) and by $0<r\leqslant 1/\{2+(\lambda^2/2K)\}<\frac{1}{2}$ for Scheme (b), K a positive number and $r=k/h^2$. (It must be assumed, with the usual notation, that the Fourier solution of the difference equations and boundary conditions at (ph, qk), $p=1(1)N$, $Nh=1$, is such that $\beta h=\pm\pi/N, \pm 2\pi/N, \ldots, \pm\pi$.)

Demonstrate that the upwind scheme is unstable for $r=\frac{1}{2}$ by showing that the initial values $u_{p,0}=(-1)^p$ are amplified by a factor of $(1+\lambda h)$ at each time-step.

If the growth of errors and of $u_{p,q}=e^{i\beta ph}\xi^q$ are restricted by the severer condition $|\xi|\leqslant 1$, show that:

(i) Scheme (a) is stable when $0<r\leqslant 2/\{\lambda^2h^2+(4-\lambda^2h^2)\sin^2\beta h/2\}$. Deduce that $r\leqslant\min\{\frac{1}{2}, 2/\lambda^2h^2\}$.

(ii) Scheme (b) is stable when $0<r\leqslant(2+\lambda h)/\{(2+\lambda h)^2\sin^2(\beta h/2)+\lambda^2h^2\cos^2\beta h/2\}$.

Deduce that $0<r\leqslant 1/(2+\lambda h)$. (See reference 19).

Solution

Substitution of $u_{p,q}=e^{i\beta ph}\xi^q$ into (a) leads to

$$\xi=1-4r\sin^2\beta h/2-i(\lambda k/h)\sin\beta h, \quad i=\sqrt{-1}.$$

The stability condition $|\xi|\leqslant 1+Kk$ for all h, k, and β, where $K>0$ is a positive number independent of h, k, and β, gives $|\xi|^2=(1-4r\sin^2\beta h/2)^2+\lambda^2 rk\sin^2\beta h\leqslant 1+2Kk+K^2k^2>1+2Kk$. This is satisfied by $|\xi|^2\leqslant 1+2Kk$. When $0<r\leqslant\frac{1}{2}$, $(1-4r\sin^2\beta h/2)^2\leqslant 1$ for all βh, and $\lambda^2 rk\sin^2\beta h\leqslant 2Kk$ for sufficiently large K.

Substitution of $u_{p,q}=\mathrm{e}^{i\beta ph}\xi^q$ into (b) leads to $\xi=1-2r(1-\cos\beta h)-\lambda rh(1-\mathrm{e}^{-i\beta h})=[1-(4r+2\lambda rh)\sin^2\beta h/2]-i\lambda rh\sin\beta h$. As in (a) this gives that

$$|\xi|^2=[1-(4r+2\lambda rh)\sin^2\beta h/2]^2+\lambda^2 rk\sin^2\beta h\leqslant 1+2Kk.$$

When $-1\leqslant 1-(4r+2\lambda rh)\sin^2\beta h/2$, the inequality is automatically satisfied for $|\beta h|<\pi$ and for sufficiently large K. This gives $r<1/(2+\lambda h)\sin^2\beta h/2$. When $\beta h=\pm\pi$, $\sin^2\beta h=0$ and the inequality is satisfied by $-1-Kk\leqslant 1-(4r+2\lambda rh)$, giving $r(2+\lambda h-\frac{1}{2}Kh^2)\leqslant 1$. Completing the square, $2+\lambda h-\frac{1}{2}Kh^2=-\frac{1}{2}K(h-\lambda/K)^2+2+(\lambda^2/2K)$ is seen to have a maximum value of $2+(\lambda^2/2K)$ when $h=\lambda/K$. Hence $r\leqslant 1/(2+\lambda^2/2K)$.

When $|\xi|^2\leqslant 1$, Scheme (a) gives

$$0\leqslant(1-4r\sin^2\beta h/2)^2+\lambda^2 rk\sin^2\beta h\leqslant 1.$$

The left-hand inequality is automatically satisfied. The right-hand inequality and $\sin\beta h=2\sin\beta h/2\cos\beta h/2$ lead to $r\{\lambda^2h^2+(4-\lambda^2h^2)\sin^2\beta h/2\}\leqslant 2$. The range $0<\beta h\leqslant\pi$ gives $r\leqslant\min\{\frac{1}{2},2/\lambda^2h^2\}$. (For the latter, let $\beta h\to 0$.)

Scheme (b) gives

$$[1-(4r+2\lambda rh)\sin^2\beta h/2]^2+\lambda^2r^2h^2\sin^2\beta h\leqslant 1.$$

The use of $\sin\beta h=2\sin\beta h/2\cos\beta h/2$ leads to $r\{(2+\lambda h)^2\sin^2\beta h/2+\lambda^2h^2\cos^2\beta h/2\}\leqslant 2+\lambda h$. Hence $r\leqslant(2+\lambda h)/\{(4+4\lambda h)\sin^2\beta h/2+\lambda^2h^2\}$ and its upper bound is given by $\beta h=\pi$, etc.

29. The function U is a solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial r^2}+\frac{1}{r}\frac{\partial U}{\partial r}+\frac{1}{r^2}\frac{\partial^2 U}{\partial\theta^2},\quad 0<r<1,\quad t>0,$$

at every point $P(r,\theta,t)$ of the open-bounded domain $0<r<1$, $t>0$, and satisfies the initial condition $U=r\sin\frac{1}{2}\theta$, $0\leqslant r\leqslant 1$,

$t=0$, and the boundary condition $\partial U \partial r=-U$ at $r=1$, $t>0$, where (r, θ, t) are the cylindrical polar co-ordinates of P.

Take $\delta r=0.1$, $\delta\theta=\pi/16$, and $\delta t=0.01(\delta r)^2$, and use an explicit method to calculate the numerical values of $u(o, \pi, \delta t)$ and $u(1, \pi, \delta t)$, where u is an approximation to U.

Solution

As shown on p. 75, the equation may be approximated at $(0, \theta_j, t_n)$ by $(u_{o,j,n+1}-u_{o,j,n})/\delta t=4(u_M-u_{0,j,n})/(\delta r)^2$, where u_M is the mean value of u round the circle $r=0.1$ at time $t_n=n\delta t$. Initially, $u(r, \theta, 0)=U(r, \theta, 0)$. Hence

$$u_M=\frac{1}{2\pi}\int_0^{2\pi} r\sin\frac{\theta}{2}\,\mathrm{d}\theta=\frac{0.1}{2\pi}\int_0^{2\pi}\sin\frac{\theta}{2}\,\mathrm{d}\theta=0.0637.$$

As $u_{0,j,0}=0$ for all j,

$$u(0, \pi, \delta t)=u_{0,16,1}=4\delta t u_M/(\delta r)^2=0.04u_M=0.0025.$$

As shown on p. 246, the equation can be approximated at the point (r_i, θ_j, t_n) by

$$\frac{u_{i,j,n+1}-u_{i,j,n}}{\delta t}=\frac{1}{(\delta r)^2}\left\{\left(1-\frac{1}{2i}\right)u_{i-1,j,n}+\left(1+\frac{1}{2i}\right)u_{i+1,j,n}\right.$$
$$\left.-2\left(1+\frac{1}{i^2(\delta\theta)^2}\right)u_{i,j,n}+\frac{1}{i^2(\delta\theta)^2}(u_{i,j-1,n}+u_{i,j+1,n})\right\}.$$

By the boundary condition, $(u_{11,j,n}-u_{9,j,n})/2(\delta r)=-u_{10,j,n}$. Hence $u_{11,j,n}=u_{9,j,n}-0.2u_{10,j,n}$. Putting $i=10$, $j=16$, and $n=0$ in the difference equation leads to

$$u_{10,16,1}=u_{10,16,0}+0.01\left\{\tfrac{19}{20}u_{9,16,0}+\tfrac{21}{20}(u_{9,16,0}-0.2u_{10,16,0})\right.$$
$$\left.-2\left(1+\frac{16^2}{100\pi^2}\right)u_{10,16,0}+\frac{16^2}{100\pi^2}(u_{10,15,0}+u_{10,17,0})\right\},$$

where $u_{10,16,0}=\sin(\pi/2)=1$, $u_{9,16,0}=0.9$, and $u_{10,15,0}=u_{10,17,0}=0.9952$. Therefore $u_{10,16,1}=1-0.01882=0.9812$ to 4*D*.

3

Parabolic equations: alternative derivation of difference equations and miscellaneous topics

This chapter is not necessary for any reader who would prefer to study, at this stage, the numerical solution of hyperbolic and/or elliptic equations.

Reduction to a system of ordinary differential equations

Consider the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad 0<x<X, \quad t>0, \tag{3.1}$$

where U satisfies the initial condition $U(x, 0)=g(x)$, $0\leqslant x\leqslant X$, and has known boundary values at $x=0$ and X, $t>0$.

If the x derivative at (x, t) is replaced by

$$\frac{1}{h^2}\{U(x-h, t)-2U(x, t)+U(x+h, t)\}+O(h^2)$$

and x is considered as a constant, eqn (3.1) can be written as the ordinary differential equation

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t}=\frac{1}{h^2}\{U(x-h, t)-2U(x, t)+U(x+h, t)\}+O(h^2). \tag{3.2}$$

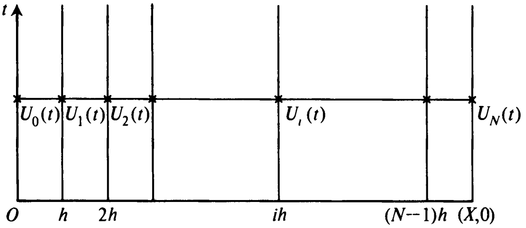


Fig. 3.1

Subdivide the interval $0 \leqslant x \leqslant X$ into N equal subintervals by the grid lines $x_i = ih$, $i = 0(1)N$, where $Nh = X$, and write down eqn (3.2) at every mesh point $x_i = ih$, $i = 1(1)N-1$, along time-level t. It then follows that the values $V_i(t)$ approximating $U_i(t)$ will be the exact solution values of the system of $(N-1)$ ordinary differential equations

$$\begin{aligned}
\frac{\mathrm{d}V_1(t)}{\mathrm{d}t} &= \frac{1}{h^2}(V_0 - 2V_1 + V_2)\\
\frac{\mathrm{d}V_2(t)}{\mathrm{d}t} &= \frac{1}{h^2}(V_1 - 2V_2 + V_3)\\
&\vdots\\
\frac{\mathrm{d}V_{N-1}}{\mathrm{d}t} &= \frac{1}{h^2}(V_{N-2} - 2V_{N-1} + V_N),
\end{aligned}$$

where V_0 and V_N are known boundary-values. These can be written in matrix form as

$$\frac{\mathrm{d}}{\mathrm{d}t}\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{N-2} \\ V_{N-1} \end{bmatrix} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & \cdot & \vdots & \vdots & \cdot \\ & & 1 & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}\begin{bmatrix} V_1 \\ V_2 \\ \vdots \\ V_{N-2} \\ V_{N-1} \end{bmatrix} + \frac{1}{h^2}\begin{bmatrix} V_0 \\ 0 \\ \vdots \\ 0 \\ V_N \end{bmatrix},$$

i.e. as

$$\frac{\mathrm{d}\mathbf{V}(t)}{\mathrm{d}t} = \mathbf{A}\mathbf{V}(t) + \mathbf{b}, \tag{3.3}$$

where $\mathbf{V}(t) = [V_1, V_2, \ldots, V_{N-1}]^{\mathrm{T}}$, $\mathbf{b}$ is a column vector of zeros and known boundary-values and matrix $\mathbf{A}$ of order $(N-1)$ is given by

$$\mathbf{A} = \frac{1}{h^2}\begin{bmatrix} -2 & 1 & & & \\ 1 & -2 & 1 & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & -2 & 1 \\ & & & 1 & -2 \end{bmatrix}. \tag{3.4}$$

The solution of the ordinary scalar differential equation

$$\frac{\mathrm{d}V}{\mathrm{d}t} = AV + b,$$

where A and b are independent of t and $V(t)$ satisfies the initial condition $V(0)=g$, is easily shown, by the method of separation of the variables, to be

$$V(t)=-\frac{b}{A}+\left(g+\frac{b}{A}\right)\exp(At).$$

At the end of this section it is shown that the solution of (3.3) satisfying the initial condition $\mathbf{V}(0)=[g_1, g_2, \ldots, g_{N-1}]^T=\mathbf{g}$, and where $\mathbf{b}$ is independent of t, is

$$\mathbf{V}(t)=-\mathbf{A}^{-1}\mathbf{b}+\{\exp(t\mathbf{A})\}(\mathbf{g}+\mathbf{A}^{-1}\mathbf{b}). \tag{3.5}$$

Hence,

$$\begin{aligned}\mathbf{V}(t+k)&=-\mathbf{A}^{-1}\mathbf{b}+\{\exp(t+k)\mathbf{A}\}(\mathbf{g}+\mathbf{A}^{-1}\mathbf{b})\\ &=-\mathbf{A}^{-1}\mathbf{b}+\{\exp(k\mathbf{A})\}\{\exp(t\mathbf{A})\}(\mathbf{g}+\mathbf{A}^{-1}\mathbf{b}).\end{aligned}$$

By eqn (3.5) this leads to

$$\mathbf{V}(t+k)=-\mathbf{A}^{-1}\mathbf{b}+\{\exp(k\mathbf{A})\}(\mathbf{V}(t)+\mathbf{A}^{-1}\mathbf{b}). \tag{3.6}$$

If all boundary values are zero,

$$\mathbf{V}(t+k)=\{\exp(k\mathbf{A})\}\mathbf{V}(t). \tag{3.7}$$

The boundary values can always be eliminated if we are concerned more, say, with stability than with a particular numerical solution. Perturb the vector of initial values from $\mathbf{g}$ to $\mathbf{g}^*$. By eqn (3.5), the solution $\mathbf{V}^*(t)$ is

$$\mathbf{V}^*(t)=-\mathbf{A}^{-1}\mathbf{b}+\{\exp(t\mathbf{A})\}(\mathbf{g}^*+\mathbf{A}^{-1}\mathbf{b}). \tag{3.8}$$

Equations (3.5) and (3.8) then show that

$$\mathbf{V}^*(t)-\mathbf{V}(t)=\{\exp(t\mathbf{A})\}(\mathbf{g}^*-\mathbf{g}).$$

Hence the perturbation vector $\mathbf{e}(t)=\mathbf{V}^*(t)-\mathbf{V}(t)$ at time t is related to the initial perturbation vector $\mathbf{e}(0)=\mathbf{g}^*-\mathbf{g}$ by

$$\mathbf{e}(t)=\{\exp(t\mathbf{A})\}\mathbf{e}(0).$$

As before, $\mathbf{e}(t+k)=\{\exp(k\mathbf{A})\}\mathbf{e}(t)$.

A note on the solution of $\mathrm{d}\mathbf{V}/\mathrm{d}t=\mathbf{AV}+\mathbf{b}$

Define the exponential matrix of the real $n\times n$ matrix $\mathbf{P}$ by

$$\exp\mathbf{P}=\mathrm{e}^{\mathbf{P}}=\mathbf{I}_n+\mathbf{P}+\frac{\mathbf{P}^2}{2!}+\frac{\mathbf{P}^3}{3!}+\cdots\sum_{m=0}^{\infty}\frac{\mathbf{P}^m}{m!}, \tag{3.9}$$

where $\mathbf{P}^0=\mathbf{I}_n$ is the unit matrix of order n.

If $\mathbf{Q}$ is a real $n \times n$ matrix such that $\mathbf{PQ} = \mathbf{QP}$, it can be proved by (3.9) that

$$e^{\mathbf{P}}e^{\mathbf{Q}} = e^{\mathbf{Q}}e^{\mathbf{P}} = e^{\mathbf{P}+\mathbf{Q}}.$$

Hence

$$e^{\mathbf{P}}e^{-\mathbf{P}} = e^{-\mathbf{P}}e^{\mathbf{P}} = e^{\mathbf{O}}.$$

But, by (3.9),

$$e^{\mathbf{O}} = \mathbf{I}_n.$$

Therefore

$$e^{\mathbf{P}}e^{-\mathbf{P}} = \mathbf{I}_n. \tag{3.10}$$

Premultiplication of both sides of (3.10) by the inverse $(e^{\mathbf{P}})^{-1}$ of $e^{\mathbf{P}}$, defined by $(e^{\mathbf{P}})^{-1}e^{\mathbf{P}} = \mathbf{I}_n$, then shows

$$e^{-\mathbf{P}} = (e^{\mathbf{P}})^{-1}.$$

On putting $\mathbf{P} = \mathbf{A}t$ into (3.9), where matrix $\mathbf{A}$ is independent of t, and differentiating with respect to t, it follows that

$$\frac{d}{dt}(e^{\mathbf{A}t}) = \mathbf{A}e^{\mathbf{A}t} = e^{\mathbf{A}t}\mathbf{A}.$$

Now consider $\mathbf{V}(t) = e^{\mathbf{A}t}\mathbf{g}$, where $\mathbf{g}$ is independent of t. This clearly satisfies the initial condition $\mathbf{V}(0) = \mathbf{g}$. Differentiation with respect to t gives that

$$\frac{d\mathbf{V}}{dt} = \mathbf{A}e^{\mathbf{A}t}\mathbf{g} = \mathbf{AV}.$$

In other words, the solution of

$$\frac{d\mathbf{V}}{dt} = \mathbf{AV} \tag{3.11}$$

which satisfies $\mathbf{V}(0) = \mathbf{g}$, is

$$\mathbf{V}(t) = e^{\mathbf{A}t}\mathbf{g}.$$

Similarly, the vector function

$$\mathbf{V}(t) = -\mathbf{A}^{-1}\mathbf{b} + e^{t\mathbf{A}}(\mathbf{g} + \mathbf{A}^{-1}\mathbf{b}),$$

which obviously satisfies the initial condition $\mathbf{V}(0) = \mathbf{g}$, is the solution of

$$\frac{d\mathbf{V}}{dt} = \mathbf{AV} + \mathbf{b},$$

provided vector **b** and matrix **A** are independent of t. The analytical solution of (3.11) in terms of the eigenvalues and eigenvectors of matrix **A** is given later.

Finite difference approximation via the ordinary differential equations

For simplicity, assume that the boundary values associated with

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad 0<x<X, \quad t>0,$$

are zero. By eqn (3.7) the recurrence relationship satisfied by the vector of values $\mathbf{V}(t)$ approximating $U(t)$ at the mesh points $x_i = ih$, $i = 1(1)N-1$, along time-level t, is

$$\mathbf{V}(t+k)=\{\exp(k\mathbf{A})\}\mathbf{V}(t), \quad t=0, k, 2k, \ldots, \tag{3.12}$$

where matrix **A** is defined by eqn (3.4).

In order to derive a set of finite difference equations from this it is necessary to approximate the exponential of $k\mathbf{A}$ by a finite algebraic function of $k\mathbf{A}$. Since $\exp(k\mathbf{A})$, by definition, is

$$\mathbf{I}+k\mathbf{A}+\tfrac{1}{2}k^2\mathbf{A}^2+\tfrac{1}{6}k^3\mathbf{A}^3+\ldots,$$

one obvious approximation is $\mathbf{I}+k\mathbf{A}$, with a leading error term of order k^2. The vector of values $\mathbf{u}=[u_1, u_2, \ldots, u_{N-1}]^T$ approximating **V** in eqn (3.12) will then be the solution of the finite-difference equations

$$\mathbf{u}(t+k)=(\mathbf{I}+k\mathbf{A})\mathbf{u}(t). \tag{3.13}$$

If $t=t_j=jk$ and $r=k/h^2$, these equations in detail, for zero boundary-values, are

$$\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix} = \begin{bmatrix} (1-2r) & r & & & \\ r & (1-2r) & r & & \\ & & \ddots & & \\ & & r & (1-2r) & r \\ & & & r & (1-2r) \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix}$$

The ith equation is the classical explicit approximation

$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j}, \quad i=1(1)N-1.$$

Other higher-order approximations are given by the Padé approximants to the exponential function.

The Padé approximants to e$^\theta$, *θ real*

Assume that e$^\theta$ is to be approximated by $(1+p_1\theta)/(1+q_1\theta)$, where p_1 and q_1 are constants. The determination of p_1 and q_1 requires two equations, which will come from the coefficients of θ and θ^2, so the leading error term will be of order θ^3. Hence

$$\mathrm{e}^\theta \equiv \frac{1+p_1\theta}{1+q_1\theta}+c_3\theta^3+c_4\theta^4+\dots.$$

Therefore,

$$(1+q_1\theta)(1+\theta+\tfrac{1}{2}\theta^2+\tfrac{1}{6}\theta^3+\dots)$$
$$\equiv 1+p_1\theta+(1+q_1\theta)(c_3\theta^3+c_4\theta^4+\dots).$$

Hence,

$$(1+q_1-p_1)\theta+(\tfrac{1}{2}+q_1)\theta^2+(\tfrac{1}{6}+\tfrac{1}{2}q_1-c_3)\theta^3$$
$$+\text{higher order terms}\equiv 0.$$

This is satisfied uniquely *to terms of order three* by

$$p_1=\tfrac{1}{2},\quad q_1=-\tfrac{1}{2}\quad\text{and}\quad c_3=-\tfrac{1}{12}.$$

The rational approximation $(1+\frac{1}{2}\theta)/(1-\frac{1}{2}\theta)$ is called the (1, 1) Padé approximation of order 2 to exp θ and has a leading error term of order 3.

In general, it is possible to approximate exp θ by

$$\mathrm{e}^\theta=\frac{1+p_1\theta+p_2\theta^2+\dots+p_T\theta^T}{1+q_1\theta+q_2\theta^2+\dots+q_S\theta^S}+c_{S+T+1}\theta^{S+T+1}+O(\theta^{S+T+2}),$$

where c_{S+T+1} is a constant. The rational function

$$R_{S,T}(\theta)=\frac{1+p_1\theta+\dots+p_T\theta^T}{1+q_1\theta+\dots+q_S\theta^S}=\frac{P_T(\theta)}{Q_S(\theta)}$$

is called the (S, T) Padé approximant of order $(S+T)$ to e$^\theta$. The following table gives the first eight Padé approximants to exp θ and their leading error terms.

TABLE 3.1

(S, T)	$R_{S,T}(\theta)$	Principal error term
$(0, 1)$	$1+\theta$	$\frac{1}{2}\theta^2$
$(0, 2)$	$1+\theta+\frac{1}{2}\theta^2$	$\frac{1}{6}\theta^3$
$(1, 0)$	$\dfrac{1}{1-\theta}$	$-\frac{1}{2}\theta^2$
$(1, 1)$	$\dfrac{1+\frac{1}{2}\theta}{1-\frac{1}{2}\theta}$	$-\frac{1}{12}\theta^3$
$(1, 2)$	$\dfrac{1+\frac{2}{3}\theta+\frac{1}{6}\theta^2}{1-\frac{1}{3}\theta}$	$-\frac{1}{72}\theta^4$
$(2, 0)$	$\dfrac{1}{1-\theta+\frac{1}{2}\theta^2}$	$\frac{1}{6}\theta^3$
$(2, 1)$	$\dfrac{1+\frac{1}{3}\theta}{1-\frac{2}{3}\theta+\frac{1}{6}\theta^2}$	$\frac{1}{72}\theta^4$
$(2, 2)$	$\dfrac{1+\frac{1}{2}\theta+\frac{1}{12}\theta^2}{1-\frac{1}{2}\theta+\frac{1}{12}\theta^2}$	$\frac{1}{720}\theta^5$

Standard finite difference equations via the Padé approximants

The classical explicit approximation

It is seen by Table 3.1 and eqn (3.13) that the classical explicit approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$ is given by approximating $\exp(k\mathbf{A})$ by its (0, 1) Padé approximant.

The classical implicit approximation

The (1, 0) Padé approximant approximates

$$\mathbf{V}(t+k) = \{\exp(k\mathbf{A})\}\mathbf{V}(t)$$

by

$$\mathbf{u}(t+k)=(\mathbf{I}-k\mathbf{A})^{-1}\mathbf{u}(t).$$

Premultiplication of both sides by the matrix $(\mathbf{I}-k\mathbf{A})$ yields

$$(\mathbf{I}-k\mathbf{A})\mathbf{u}(t_j+k)=\mathbf{u}(t_j), \quad j=0,1,2,\ldots,$$

where $\mathbf{u}(t_j+k)=[u_{1,j+1}, u_{2,j+1}, \ldots, u_{N-1,j+1}]^T$ and matrix $\mathbf{A}$ is defined by (3.4). In detail, for zero boundary-conditions,

$$\begin{bmatrix} (1+2r) & -r & & & \\ -r & (1+2r) & -r & & \\ & & \ddots & & \\ & & -r & (1+2r) & -r \\ & & & -r & (1+2r) \end{bmatrix}$$

$$\times\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix}=\begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix}.$$

The ith equation gives the implicit or backward-difference scheme

$$-ru_{i-1,j+1}+(1-2r)u_{i,j+1}-ru_{i+1,j+1}=u_{i,j}, \quad i=1(1)N-1.$$

This is unconditionally stable for all $r=k/h^2>0$. The leading error terms are of order h^2 in x because of the central-difference approximation to $\partial^2U/\partial x^2$ and of order k in t. (The leading error term of the $(1,0)$ Padé approximant to $\exp(k\mathbf{A})$ is $O(k^2)$ but eqn (3.12) was derived by integrating the ordinary differential for $\mathbf{V}(t)$ with respect to t.) The method is said to be first order accurate in t.

The Crank–Nicolson equations

Table 3.1 shows that the $(1,1)$ Padé approximant replaces

$$\mathbf{V}(t+k)=\{\exp(k\mathbf{A})\}\mathbf{V}(t)$$

by

$$\mathbf{u}(t+k)=(\mathbf{I}-\tfrac{1}{2}k\mathbf{A})^{-1}(\mathbf{I}+\tfrac{1}{2}k\mathbf{A})\mathbf{u}(t).$$

For numerical calculations this needs to be written as

$$(\mathbf{I}-\tfrac{1}{2}k\mathbf{A})\mathbf{u}(t+k)=(\mathbf{I}+\tfrac{1}{2}k\mathbf{A})\mathbf{u}(t).$$

This gives the Crank–Nicolson scheme

$$\begin{aligned}-ru_{i-1,j+1}+2(1+r)u_{i,j+1}-ru_{i+1,j+1}&\\ =ru_{i-1,j}+2(1-r)u_{i,j}+ru_{i+1,j},&\quad i=1(1)N-1.\end{aligned}$$

It is second-order accurate in t, having an error term via its Padé approximant of order k^3, and may be used with larger time-intervals than the backward difference method. As shown later, it can, however, produce unwanted finite oscillations near points of discontinuity if $k/h>X/\pi$. (See Fig. 3.2.)

A_0-stability, L_0-stability and the symbol of the method

Assume that the boundary values associated with the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad 0<x<X,\quad t>0,$$

are zero and that the vector of initial values is

$$\mathbf{U}(0)=\mathbf{g}=[g_1, g_2, \ldots, g_{N-1}]^T,\quad Nh=X.$$

Then the vector of values $\mathbf{V}(t)$ approximating $\mathbf{U}(t)$ at the mesh points $x_i=ih$, $i=1(1)N-1$, along time-level $t_j=jk$ satisfies the recurrence relationship

$$\mathbf{V}(t_j+k)=\{\exp(k\mathbf{A})\}\mathbf{V}(t_j),\quad j=0, 1, 2, \ldots.$$

If the exponential of $k\mathbf{A}$ is approximated by its (S, T) Padé approximant $R_{S,T}(k\mathbf{A})$, the resulting set of finite difference equations is

$$\mathbf{u}(t_j+k)=R_{S,T}(k\mathbf{A})\mathbf{u}(t_j),$$

which can be written equivalently as

$$\mathbf{u}(t_j)=R_{S,T}(k\mathbf{A})\mathbf{u}(t_{j-1}).$$

Applied recursively this leads to

$$\mathbf{u}(t_j)=[R_{S,T}(k\mathbf{A})]^j\mathbf{u}(0), \tag{3.14}$$

where $\mathbf{u}(0)=\mathbf{U}(0)=\mathbf{g}$. Now the eigenvalues of matrix $\mathbf{A}$ are

$$\lambda_s=-\frac{4}{h^2}\sin^2\frac{s\pi}{2N},\quad s=1(1)N-1,$$

and are all different. Hence the $(N-1)$ eigenvectors $\mathbf{v}_s$ of $\mathbf{A}$ are linearly independent and can be used as a basis for the $(N-1)$ dimensional space of the vector $\mathbf{g}$ of initial values. In other words, $\mathbf{g}$ can be expressed as

$$\mathbf{g}=\sum_{s=1}^{N-1}c_s\mathbf{v}_s,$$

where the c_s are constants. Equation (3.14) can therefore be written as

$$\begin{aligned}\mathbf{u}(t_j)&=[R_{S,T}(k\mathbf{A})]^j\sum_{s=1}^{N-1}c_s\mathbf{v}_s\\&=\sum_{s=1}^{N-1}c_s[R_{S,T}(k\mathbf{A})]^j\mathbf{v}_s.\end{aligned}\tag{3.15}$$

Remembering that $\mathbf{A}\mathbf{v}_s=\lambda_s\mathbf{v}_s$ and that $f(\mathbf{A})\mathbf{v}_s=f(\lambda_s)\mathbf{v}_s$, it follows by eqn (3.15) that the solution of the finite-difference equations can be expressed as

$$\mathbf{u}(t_j)=\sum_{s=1}^{N-1}c_s[R_{S,T}(k\lambda_s)]^j\mathbf{v}_s.\tag{3.16}$$

Equation (3.16) shows that $\mathbf{u}(t_j)$ will tend to the null vector as $j\to\infty$ if and only if $|R_{S,T}(k\lambda_s)|<1$, $s=1(1)N-1$. All rounding errors will also tend to zero because they are subject to the same arithmetic operations as the components of $\mathbf{u}(t_j)$, $j=1,2,\ldots$. If this condition depends on the value of $r=k/h^2$, the equations are conditionally stable. This definition corresponds exactly to conditional stability defined by the matrix method for fixed h and k, when the solution tends to zero with increasing t.

When $|R_{S,T}(k\lambda_s)|<1$ *for all* $r>0$, *the equations are said to be* A_0*-stable.* This definition corresponds exactly to unconditional stability defined by the matrix method in the sense that $\mathbf{u}(t_j)$ tends to the null vector as $j\to\infty$ and is not merely bounded for fixed h and k.

The suffix '0' refers to the fact that all the eigenvalues of matrix $\mathbf{A}$ are real, negative and non-zero, so that λ_s, considered as a complex number, is such that $\pi+0=\arg\lambda_s=\pi-0$. (If the eigenvalues are within the wedge defined by $\pi+\alpha>\arg\lambda_s>\pi-\alpha$, $0<\alpha<\pi/2$, then the method is $A(\alpha)$-stable when, within this wedge, $|R_{S,T}(k\lambda_s)|<1$, $s=1(1)N-1$, for all $r>0$.)

Although A_0-stability implies that $-1<R_{S,T}(k\lambda_s)<1$ for real $R_{S,T}(k\lambda_s)$, it is possible that some values of $R_{S,T}(k\lambda_s)$ might be close to -1 for particular values of $k\lambda_s$. The corresponding values of $[R_{S,T}(k\lambda_s)]^j$ in eqn (3.16) will then alternate in sign as j increases and diminish in magnitude only very slowly. If the corresponding terms on the right-hand side of (3.16) are large in comparison with the remaining terms, the numerical solution could then oscillate finitely as j increases. This phenomenon is particularly pronounced in the x-neighbourhoods of points of discontinuity either in the initial values or between boundary values and initial values, as is illustrated in the next section.

The real coefficients $R_{S,T}(k\lambda_s)$ in (3.16) would clearly give a solution that is stable and free of unwanted oscillations if $0<R_{S,T}(k\lambda_s)<1$, $s=1(1)N-1$, and $R_{S,T}(k\lambda_s)\rightarrow 0$ monotonically as $k\lambda_s$ increases in magnitude. The $(1,0)$ Padé approximant, for which $R_{1,0}(k\lambda_s)=\dfrac{1}{1-k\lambda_s}$, λ_s real and negative, is of this character. In practice, this is an unnecessarily severe set of conditions and it is sufficient for $R_{S,T}(k\lambda_s)$ to tend to zero through positive and/or negative values as $k\lambda_s\rightarrow-\infty$, $\lambda_s<0$ and real.

As a consequence, a set of difference equations is said to be L_0*-stable if* $|R_{S,T}(k\lambda_s)|<1$, $s=1(1)N-1$, *and* $R_{S,T}(k\lambda_s)\rightarrow 0$ *as* $k\lambda_s\rightarrow-\infty$, *where* λ_s *is real, negative, and non-zero.*

Following Gourlay and Morris 1980, reference 10, it is usual to put $k\lambda_s=-z$, which makes z positive since $\lambda_s=-(4/h^2)\sin^2 s\pi/2N$, and to call $R_{S,T}(-z)$ *the symbol of the method.*

A set of difference equations would then be L_0-stable if $|R_{S,T}(-z)|<1$ for all $z>0$ and $R_{S,T}(-z)\rightarrow 0$ as $z\rightarrow\infty$. For the $(1,0)$ Padé approximant, $R_{1,0}(-z)=1/(1+z)$ and obviously satisfies both conditions, showing that the backward-difference scheme is L_0-stable. (See Fig. 3.3.)

For the Crank–Nicolson scheme,

$$R_{1,1}(-z)=\frac{1-\frac{1}{2}z}{1+\frac{1}{2}z}=\frac{2/z-1}{2/z+1}. \tag{3.17}$$

Clearly, $|R_{1,1}(-z)|<1$ for all $z>0$, but $R_{1,1}(-z)\rightarrow-1$ as $z\rightarrow\infty$, showing that the scheme is A_0-stable.

In general, L_0-stable schemes are preferable to A_0-stable ones because possible unwanted finite oscillations in the numerical solution are rapidly dampened and this eliminates constraints on the time-step k in relation to the space-step h, as occurs with the Crank–Nicolson method. (See the next section.) With A_0-stable methods it is also common for unwanted finite oscillations to increase in magnitude when any attempt is made to improve accuracy by decreasing h. This occurs if the term $R_{S,T}(k\lambda_s)$ of eqn (3.16) approaches -1 more closely because of the decrease in h. The Crank–Nicolson equations exhibit this phenomenon because a decrease in h increases the magnitude of $\lambda_s = (-4/h^2)\sin^2 s\pi/2N$, i.e. increases z and brings $R_{1,1}(-z)$ of eqn (3.17) closer to -1. (See also the section on Stiff Equations.)

As a consequence, a great deal of research since 1978 has been directed towards the generation of L_0-stable methods of high-order accuracy in t. (Reference Lawson and Morris 1978, Gourlay and Morris 1980, Twizell and Khaliq 1982.)

The Padé approximation schemes for parabolic equations are L_0-stable when $S>T$ and A_0-stable when $S=T$.

A necessary constraint on the time-step for the Crank–Nicolson method

Numerical studies indicate that very slowly decaying finite oscillations can occur with the Crank–Nicolson method in the neighbourhood of discontinuities in the initial values or between initial values and boundary values.

Assume that the boundary values for

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}, \quad 0<x<X, \quad t>0,$$

are zero and that the vector of initial values is $\mathbf{U}_0=\mathbf{g}$. Then the Crank–Nicolson or (1, 1) Padé approximant solution corresponding to eqn (3.16) will be

$$\mathbf{u}_j = \sum_{s=1}^{N-1} c_s \mu_s^j \mathbf{v}_s, \tag{3.18}$$

where

$$\mu_s = \frac{1+\frac{1}{2}k\lambda_s}{1-\frac{1}{2}k\lambda_s}, \quad s=1(1)N-1, \tag{3.19}$$

are the eigenvalues of the amplification matrix $(\mathbf{I}-\frac{1}{2}k\mathbf{A})^{-1}(\mathbf{I}+\frac{1}{2}k\mathbf{A})$, λ_s and $\mathbf{v}_s$ are the eigenvalues and corresponding eigenvectors of matrix $\mathbf{A}$ of (3.4), $Nh = X$, $\lambda_s = -(4/h^2)\sin^2 s\pi/2N$, and

$$\mathbf{g} = \sum_{s=1}^{N-1} c_s \mathbf{v}_s. \tag{3.20}$$

The eigenvalue μ_s, often called the growth factor associated with $\mathbf{v}_s$, is less than one in modulus for all s because λ_s is negative. Hence eqn (3.18) shows that $\mathbf{u}_j$ tends to the null vector as $j \to \infty$. In this sense the equations are unconditionally stable. By eqn (3.19) it is seen, however, that μ_s will be close to -1 when $k\lambda_s = -4r\sin^2 s\pi/2N$ is large, where $r = k/h^2$. This will occur when r is large and $s\pi/2N \simeq \pi/2$, implying N and s both large, i.e. $s = N-1, N-2, \ldots$.

By eqns (3.20) and (3.18) it is seen that the high-frequency components $c_{N-1}\mathbf{v}_{N-1}$, $c_{N-2}\mathbf{v}_{N-2}, \ldots$, of the initial values have been transformed to $c_{N-1}\mu^j_{N-1}\mathbf{v}_{N-1}$, $c_{N-2}\mu^j_{N-2}\mathbf{v}_{N-2}, \ldots$, respectively, at the jth time-level. As j increases these components will alternate in sign and decay only very slowly. If $c_{N-1}, c_{N-2}, \ldots$, are large, as tends to occur when there are discontinuities between initial values and boundary values, the solution will oscillate finitely near the points of discontinuity. (See Fig. 3.2.)

Lawson and Morris, reference 14, have pointed out, however, that these oscillatory terms will not have disastrous consequences provided the highest-frequency component $c_{N-1}\mathbf{v}_{N-1}$ decays to zero faster than the lowest-frequency component $c_1\mathbf{v}_1$. The condition for this, by (3.18), is that $|\mu_{N-1}| < |\mu_1|$, i.e. $-\mu_1 < \mu_{N-1} < \mu_1$, giving

$$\frac{-1-\frac{1}{2}k\lambda_1}{1-\frac{1}{2}k\lambda_1} < \frac{1+\frac{1}{2}k\lambda_{N-1}}{1-\frac{1}{2}k\lambda_{N-1}} < \frac{1+\frac{1}{2}k\lambda_1}{1-\frac{1}{2}k\lambda_1}.$$

The right-hand inequality is automatically satisfied by $\lambda_s = -(4/h^2)\sin^2 s\pi/2N$. The left-hand inequality leads to

$$k^2\lambda_1\lambda_{N-1} < 4. \tag{3.21}$$

For large N, $\lambda_1 \simeq -4\pi^2/4h^2N^2 = -\pi^2/X^2$, and

$$\lambda_{N-1} \simeq -(4/h^2)\sin^2\pi/2 = -4/h^2.$$

Hence by (3.21),

$$k/h < X/\pi \text{ approximately, or } k < hX/\pi.$$

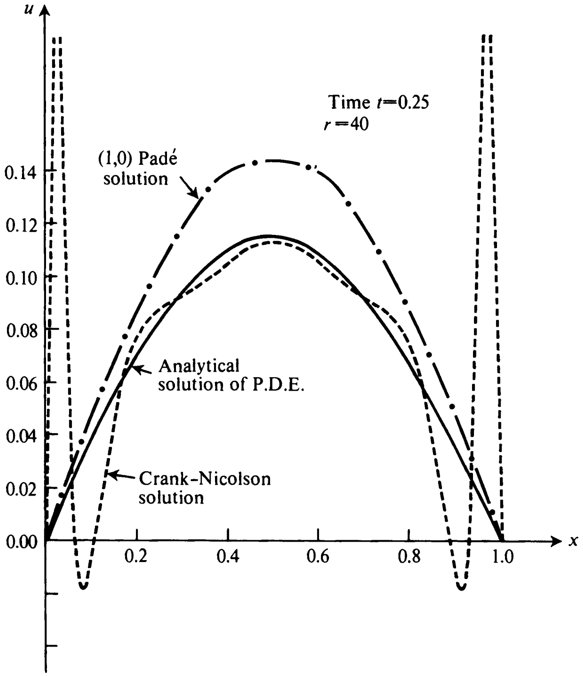


Fig. 3.2

Figure 3.2 exhibits the analytical solution at $t=0.25$ of $\partial U/\partial t=\partial^2 U/\partial x^2$, $0<x<1$, $t>0$, satisfying $U(0,t)=U(1,t)=0$, $t>0$, and $U(0,x)=1$, $0\leqslant x\leqslant 1$, $t=0$, together with the Crank–Nicolson and backward-difference or $(1,0)$ Padé approximant solutions for $h=0.025$ and $r=40$. For these values, $k=0.025>hX/\pi=0.00796$.

The local truncation errors associated with the Padé approximants

When the exponential of matrix $k\mathbf{A}$, defined by (3.4), is approximated by its (S,T) Padé approximant, the difference equation approximating

$$\frac{\partial U}{\partial t}-\frac{\partial^2 U}{\partial x^2}=0$$

at the point (i, j) is the ith row of the equations

$$\frac{1}{k}\{\mathbf{u}(t_j+k)-\mathbf{R}_{S,T}(k\mathbf{A})\mathbf{u}(t_j)\}=0, \quad i=1(1)N-1,$$

where $\mathbf{R}_{S,T}(k\mathbf{A})=\mathbf{Q}_S^{-1}(k\mathbf{A})\mathbf{P}_T(k\mathbf{A})$. Assuming that $\mathbf{Q}_s^{-1}$ is nonsingular, it follows that the difference equations can be written as

$$\frac{1}{k}\{\mathbf{Q}_S(k\mathbf{A})\mathbf{u}_{j+1}-\mathbf{P}_T(k\mathbf{A})\mathbf{u}_j\}=0, \quad i=1(1)N-1.$$

The definition of the local truncation error $T_{i,j}(U)$ then gives

$$T_{i,j}(U)=i\text{th row of } \frac{1}{k}\{\mathbf{Q}_S(k\mathbf{A})\mathbf{U}_{j+1}-\mathbf{P}_T(k\mathbf{A})\mathbf{U}_j\},$$
$$i=2(1)N-2.$$

If all the terms on the right-hand side are expanded about the point (i, j) by Taylor's series, (see Exercise 4), it can be shown that the principal part of the local truncation error at (i, j) is

$$\left[-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}+C_q k^{q-1}\frac{\partial^q U}{\partial t^q}\right]_{i,j},$$

where $q=S+T+1$. Some of the constants C_q are given in Table 3.2. (Reference 13.) The component $-\frac{1}{12}h^2\,\partial^4U/\partial x^4$ arises from

TABLE 3.2

(S, T)	q and r	C_q	E_r
$(0, 1)$	2	$\frac{1}{2}$	$\frac{4}{3}$
$(1, 1)$	3	$-\frac{1}{12}$	$\frac{1}{10}$
$(1, 0)$	2	$-\frac{1}{2}$	$\frac{4}{3}$
$(2, 0)$	3	$\frac{1}{6}$	$-\frac{1}{3}$
$(2, 1)$	4	$\frac{1}{72}$	$-\frac{8}{945}$
$(2, 2)$	5	$\frac{1}{720}$	$-\frac{1}{1890}$
$(1, 2)$	4	$-\frac{1}{72}$	$-\frac{8}{945}$

the approximation of $\partial^2U/\partial x^2$ by $\{u(x-h, t)-2u(x, t)+u(x+h, t)\}/h^2$ and will appear in the principal part of the local truncation error of every difference scheme that approximates $\exp(k\mathbf{A})$ by a Padé approximant. The C_q term is determined by (S, T).

Stiff equations

The magnitude of the component $-h^2(\partial^4U/\partial x^4)/12$ of the local truncation error can obviously be decreased by increasing N, since $Nh=X$, a fixed number. This inevitably increases the number and range of the eigenvalues $\lambda_s=-(4/h^2)\sin^2 s\pi/2N$, $s=1(1)N-1$, of matrix $\mathbf{A}$. As a consequence, the analytical solution of the difference equations, namely

$$\mathbf{u}_j=\sum_{s=1}^{N-1} c_s[R_{S,T}(k\lambda_s)]^j\mathbf{v}_s,$$

may contain a large number of components with widely varying rates of decay, assuming $|R_{S,T}(k\lambda_s)|<1$. Equations giving rise to this phenomenon are said to be 'stiff'. If, in general, the eigenvalues of matrix $\mathbf{A}$ are $\lambda_s=-\mu_s+i\nu_s$, where μ_s, ν_s are real and $\mu_s>0$, the 'stiffness ratio' of the equations is measured by $\max \mu_s/\min \mu_s$. For $\mathbf{A}$ defined by (3.4), the stiffness ratio for large N is

$$\{\sin^2(N-1)\pi/2N\}/\sin^2\pi/2N\simeq 4N^2/\pi^2.$$

An extrapolation method for improving accuracy in *t*

Consider the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad 0<x<X,\quad t>0,$$

satisfying the boundary conditions $U(0, t)=U(X, t)=0$, $t>0$. As shown previously, the vector of values $\mathbf{V}$ approximating U at the mesh points $x_i=ih$, $i=1(1)N-1$, along time-levels t and $t+k$ satisfy

$$\mathbf{V}(t+k)=\{\exp(k\mathbf{A})\}\mathbf{V}(t),\quad t=0, k, 2k, \ldots. \tag{3.22}$$

If the exponential is approximated by its (1, 0) Padé approximant, the vector of values $\mathbf{u}=[u_1, u_2, \ldots, u_{N-1}]^T$ approximating $\mathbf{V}$ will be the solution of the implicit backward difference equations

$$\mathbf{u}(t+k)=(\mathbf{I}-k\mathbf{A})^{-1}\mathbf{u}(t), \tag{3.23}$$

where matrix $\mathbf{A}$ is defined by (3.4). Over a time-interval of $2k$ this gives

$$\mathbf{u}^{(1)}(t+2k)=(\mathbf{I}-2k\mathbf{A})^{-1}\mathbf{u}(t). \tag{3.24}$$

Alternatively, the application of eqn (3.23) twice, each over a time-interval of k, leads to the implicit equations

$$\mathbf{u}^{(2)}(t+2k)=(\mathbf{I}-k\mathbf{A})^{-1}\mathbf{u}(t+k)=(\mathbf{I}-k\mathbf{A})^{-1}(\mathbf{I}-k\mathbf{A})^{-1}\mathbf{u}(t)$$

i.e.

$$\mathbf{u}^{(2)}(t+2k)=(\mathbf{I}-k\mathbf{A})^{-2}\mathbf{u}(t). \tag{3.25}$$

Equations (3.24) and (3.25) are two different backward-difference schemes for calculating approximations to $U_i(t+2k)$, $i=1(1)N-1$. The detail of the arithmetic is given later.

The binomial expansion of the matrix inverses of (3.24) and (3.25) show that

$$\mathbf{u}^{(1)}(t+2k)=(\mathbf{I}+2k\mathbf{A}+4k^2\mathbf{A}^2)\mathbf{u}(t)+O(k^3) \tag{3.26}$$

and

$$\mathbf{u}^{(2)}(t+2k)=(\mathbf{I}+2k\mathbf{A}+3k^2\mathbf{A}^2)\mathbf{u}(t)+O(k^3). \tag{3.27}$$

But the Maclaurin expansion of $\exp(2k\mathbf{A})$ in

$$\mathbf{V}(t+2k)=\{\exp(2k\mathbf{A})\}\mathbf{V}(t),$$

an equation giving a more accurate approximation to $U_i(t+2k)$ than either (3.26) or (3.27), $i=1(1)N-1$, results in

$$\mathbf{V}(t+2k)=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2)\mathbf{V}(t)+O(k^3). \tag{3.28}$$

Comparison of eqns (3.26), (3.27), and (3.28) shows that neither (3.26) nor (3.27) is accurate to terms of order k^2. A simple linear combination of (3.26) and (3.27) will, however, produce an extrapolated vector $\mathbf{u}^{(E)}$ that is second-order accurate in t, i.e. with a leading error term $O(k^3)$, namely,

$$\begin{aligned}\mathbf{u}^{(E)}(t+2k)&=2\mathbf{u}^{(2)}(t+2k)-\mathbf{u}^{(1)}(t+2k)\\&=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2)\mathbf{u}(t)\end{aligned}$$

The algorithm for the extrapolation is therefore

$$(\mathbf{I}-2k\mathbf{A})\mathbf{u}^{(1)}(t+2k)=\mathbf{u}(t), \tag{3.29}$$

$$(\mathbf{I}-k\mathbf{A})^2\mathbf{u}^{(2)}(t+2k)=\mathbf{u}(t) \tag{3.30}$$

and

$$\mathbf{u}^{(E)}(t+2k)=2\mathbf{u}^{(2)}-\mathbf{u}^{(1)}. \tag{3.31}$$

Naturally, the extrapolated solution values are used as the starting values for the extrapolation procedure over the next two time-levels.

The symbol for the extrapolation method

If eqn (3.31) is written in the form

$$\mathbf{u}^{(E)}(t+2k)=S_{1,0}(k\mathbf{A})\mathbf{u}(t)=\{2(\mathbf{I}-k\mathbf{A})^{-2}-(\mathbf{I}-2k\mathbf{A})^{-1}\}\mathbf{u}(t),$$

then

$$S_{1,0}(k\mathbf{A})=2(\mathbf{I}-k\mathbf{A})^{-2}-(\mathbf{I}-2k\mathbf{A})^{-1}.$$

Therefore the symbol $S_{1,0}(-z)$ of the extrapolation method is

$$S_{1,0}(-z)=\frac{2}{(1+z)^2}-\frac{1}{1+2z}=\frac{1+2z-z^2}{1+4z+5z^2+2z^3}.$$

Division of the numerator and denominator by z^2 shows that $S_{1,0}(-z)\rightarrow 0$ as $z\rightarrow\infty$, whilst Fig. 3.3 demonstrates that $|S_{1,0}(-z)|<1$ for all $z>0$. Hence the scheme is L_0-stable. The symbol is small and negative for $z>1+\sqrt{2}$, which implies that small oscillations or fluctuations could occur in the numerical solution for $z=-k\lambda_s>1+\sqrt{2}$. They would, of course, be heavily

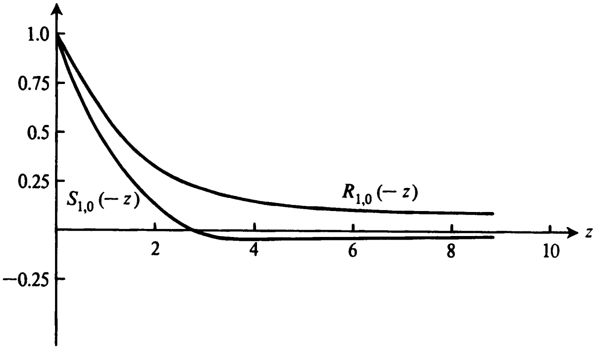


Fig. 3.3

damped in subsequent calculations because $S_{1,0}(-z)$ is very small for these values of z.

In practice, only L_0-stable methods are worth extrapolating.

The arithmetic of the extrapolation method

As

$$\mathbf{I}-2k\mathbf{A}=\begin{bmatrix}(1+4r) & -2r & & \\ -2r & (1+4r) & -2r & \\ & & \ddots & \\ & & -2r & (1+4r)\end{bmatrix},$$

it follows that eqn (3.29), with zero boundary-values, yields the equations

$$\begin{array}{rlll} & (1+4r)u^{(1)}_{1,j+2} & -2ru^{(1)}_{2,j+2} & =u_{1,j} \\ -2ru^{(1)}_{1,j+2} & +(1+4r)u^{(1)}_{2,j+2} & -2ru^{(1)}_{3,j+2} & =u_{2,j} \\ \vdots & \vdots & \vdots & \vdots \\ -2ru^{(1)}_{N-2,j+2} & +(1+4r)u^{(1)}_{N-1,j+2} & & =u_{N-1,j}. \end{array}$$

These can be solved by the algorithm on p. 24. Similarly, eqn (3.30) can be treated as two tridiagonal systems, namely,

$$(\mathbf{I}-k\mathbf{A})\mathbf{u}^{(2)}(t+k)=\mathbf{u}(t) \tag{3.32}$$

and

$$(\mathbf{I}-k\mathbf{A})\mathbf{u}^{(2)}(t+2k)=\mathbf{u}^{(2)}(t+k), \tag{3.33}$$

where the equations for (3.32) are

$$\begin{array}{rlll} & (1+2r)u^{(2)}_{1,j+1} & -ru^{(2)}_{2,j+1} & =u_{1,j} \\ -ru^{(2)}_{1,j+1} & +(1+2r)u^{(2)}_{2,j+1} & -ru^{(2)}_{3,j+1} & =u_{2,j} \\ & & & \\ -ru^{(2)}_{N-2,j+1} & +(1+2r)u^{(2)}_{N-1,j+1} & & =u_{N-1,j+1} \end{array}$$

Similarly for (3.33). Alternatively, eqn (3.30) can be written as

$$(\mathbf{I}-2k\mathbf{A}+k^2\mathbf{A}^2)\mathbf{u}^{(2)}(t+2k)=\mathbf{u}(t), \tag{3.34}$$

where

$$\mathbf{I}-2k\mathbf{A}+k^2\mathbf{A}^2$$

$$=\mathbf{I}-2r\begin{bmatrix} -2 & 1 & & & & \\ 1 & -2 & 1 & & & \\ & 1 & -2 & 1 & & \\ & & & \ddots & & \\ & & & & 1 & -2 \end{bmatrix}+r^2\begin{bmatrix} 5 & -4 & 1 & & & & & \\ -4 & 6 & -4 & 1 & & & & \\ 1 & -4 & 6 & -4 & 1 & & & \\ & 1 & -4 & 6 & -4 & 1 & & \\ & & & & \ddots & & & \\ & & & & 1 & -4 & 6 & -4 \\ & & & & & 1 & -4 & 5 \end{bmatrix}$$

$$=\begin{bmatrix} (1+4r+5r^2) & (-2r-4r^2) & r^2 & & & \\ (-2r-4r^2) & (1+4r+6r^2) & (-2r-4r^2) & r^2 & & \\ r^2 & (-2r-4r^2) & (1+4r+6r^2) & (-2r-4r^2) & r^2 & \\ & r^2 & (-2r-4r^2) & (1+4r+6r^2) & (-2r-4r^2) & r^2 \\ & & & \ddots & & \\ & & r^2 & (-2r-4r^2) & (1+4r+6r^2) & (-2r-4r^2) \\ & & & r^2 & (-2r-4r^2) & (1+4r+5r^2) \end{bmatrix}$$

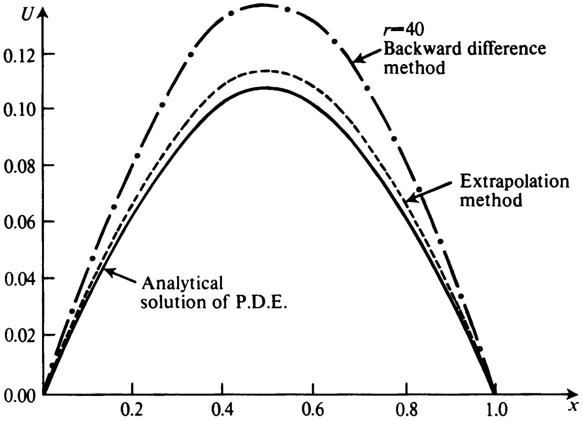


Fig. 3.4

Equation (3.34) can then be solved directly for the components of $\mathbf{u}^{(2)}(t+2k)$ by means of a quindiagonal solver.

For the heat-flow problem $\partial U/\partial t=\partial^2U/\partial x^2$, $0<x<1$, $t>0$, $U(0,t)=U(1,t)=0$, $t>0$, and $U(x,0)=1$, $0\leqslant x\leqslant 1$, Fig. 3.4 shows the graphs of the solutions at $t=0.25$ by the backward-difference method, eqn (3.23), and the extrapolation method, eqn (3.31) taking $h=k=0.025$ and $r=40$. The maximum errors, at $x=0.5$, are 0.0324 (backward-difference) and 0.0061 (extrapolation).

Further comments

If the classical explicit method is extrapolated it is easily shown that $|S_{0,1}(-z)|\leqslant 1$ for $0<r\leqslant 1/4$. (Exercise 2.) In general, there is little advantage in the extrapolation of conditionally stable and A_0-stable methods, especially if discontinuities exist in the initial values or between initial values and boundary values.

As one would expect, L_0-stable difference methods of third and fourth accuracy in t can be achieved by extrapolating over three or four time-levels respectively and two such schemes are considered in Gourlay and Morris, reference 10. The extrapolation of (2, 0), (2, 1), (2, 2), and (3, 0) Padé approximations are given in references 28 and 13.

The local truncation errors and amplification symbols of extrapolated schemes

In general, the extrapolation formula for the (S, T) Padé approximant is

$$\mathbf{u}^{(E)}(t+2k)=\alpha\mathbf{u}^{(2)}-(\alpha-1)\mathbf{u}^{(1)},$$

where $\alpha=2^{S+T}/(2^{S+T}-1)$.

The principal part of the local truncation error of the extrapolated scheme at (i, j) is

$$\left[-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}+E_r k^{r-1}\frac{\partial^r U}{\partial t^r}\right]_{i,j}, \quad i=2(1)N-2,$$

where $r=S+T+2$. Some of the constants E_r are given in Table 3.2. The amplification symbol of the extrapolated method associated with the (S, T) Padé approximant is

$$S_{S,T}(-z)=\alpha[R_{S,T}(-z)]^2-(\alpha-1)R_{S,T}(-2z).$$

As an example, consider the $(2, 0)$ Padé approximant to $\exp\theta$, namely $R_{2,0}(\theta)=1/(1-\theta+\frac{1}{2}\theta^2)$. For this,

$$\alpha=2^2/(2^2-1)=4/3, \quad \mathbf{u}^{(E)}=\tfrac{4}{3}\mathbf{u}^{(2)}-\tfrac{1}{3}\mathbf{u}^{(1)} \quad \text{and} \quad r=4.$$

The principal part of the local truncation error is

$$\left[-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}-\tfrac{1}{3}k^3\frac{\partial^4 U}{\partial t^4}\right]_{i,j}$$

and

$$S_{2,0}(-z)=\frac{4}{3(1+z+\frac{1}{2}z^2)}-\frac{1}{3(1+2z+2z^2)},$$

which is easily shown to give L_0-stability.

The eigenvalue–eigenvector solution of a system of O.D.E.'s

Preliminary results

(i) If matrix $\mathbf{A}$ of order N has N linearly independent eigenvectors $\mathbf{x}_i$ corresponding to the N different eigenvalues λ_i, $i=1(1)N$, then the N equations

$$\mathbf{A}\mathbf{x}_i=\lambda_i\mathbf{x}_i, \quad i=1(1)N,$$

may be written in matrix form as

$$\begin{aligned}\mathbf{A}[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]&=[\mathbf{A}\mathbf{x}_1,\mathbf{A}\mathbf{x}_2,\ldots,\mathbf{A}\mathbf{x}_N]\\&=[\lambda_1\mathbf{x}_1,\lambda_2\mathbf{x}_2,\ldots,\lambda_N\mathbf{x}_N]\\&=[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N\end{bmatrix}.\end{aligned}$$

Putting the matrix of eigenvectors $[\mathbf{x}_1,\mathbf{x}_2,\ldots,\mathbf{x}_N]=\mathbf{X}$, often called the modal matrix, and

$$\begin{bmatrix}\lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_N\end{bmatrix}=\operatorname{diag}(\lambda_1,\lambda_2,\ldots,\lambda_N)=\mathbf{D},$$

the eigenvalue–eigenvector equations can be expressed very compactly as

$$\mathbf{AX}=\mathbf{XD}.$$

Hence

$$\mathbf{X}^{-1}\mathbf{AX}=\mathbf{D}. \tag{3.35}$$

(ii) By eqn (3.35),

$$\begin{aligned}(\mathbf{X}^{-1}\mathbf{AX})^2=(\mathbf{X}^{-1}\mathbf{AX})(\mathbf{X}^{-1}\mathbf{AX})&=\mathbf{X}^{-1}\mathbf{A}^2\mathbf{X}=\mathbf{D}^2\\&=\operatorname{diag}(\lambda_1^2,\lambda_2^2,\ldots,\lambda_N^2).\end{aligned}$$

Similarly, $\mathbf{X}^{-1}\mathbf{A}^m\mathbf{X}=\mathbf{D}^m=\operatorname{diag}(\lambda_1^m,\lambda_2^m,\ldots,\lambda_N^m)$.

(iii) By definition of exp **A**,

$$\begin{aligned}\mathbf{X}^{-1}(\exp\mathbf{A})\mathbf{X}&=\mathbf{X}^{-1}\left[\mathbf{I}_N+\mathbf{A}+\frac{1}{2!}\mathbf{A}^2+\ldots\right]\mathbf{X}\\&=\mathbf{I}_N+\mathbf{D}+\frac{1}{2!}\mathbf{D}^2+\ldots\end{aligned}$$

i.e.

$$\mathbf{X}^{-1}(\exp\mathbf{A})\mathbf{X}=\exp\mathbf{D}.$$

(iv) Using 2×2 matrices for simplicity, the definition of

$\exp(t\mathbf{D})$, t a scalar, gives

$$\begin{aligned}\exp(t\mathbf{D}) &= \mathbf{I}_2 + t\mathbf{D} + \frac{1}{2!}(t\mathbf{D})^2 + \ldots + \frac{1}{m!}(t\mathbf{D})^m + \ldots\\ &= \begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix} + \begin{bmatrix}t\lambda_1 & 0\\ 0 & t\lambda_2\end{bmatrix} + \ldots + \frac{1}{m!}\begin{bmatrix}t^m\lambda_1^m & 0\\ 0 & t^m\lambda_2^m\end{bmatrix} + \ldots\end{aligned}$$

$$= \begin{bmatrix}\left(1 + t\lambda_1 + \cdots + \frac{1}{m}t^m\lambda_1^m + \cdots\right) & \\ & \left(1 + t\lambda_2 + \cdots + \frac{1}{m!}t^m\lambda_2^m + \cdots\right)\end{bmatrix},$$

i.e.

$$\exp(t\mathbf{D}) = \begin{bmatrix}\exp(t\lambda_1) & 0\\ 0 & \exp(t\lambda_2)\end{bmatrix}. \tag{3.36}$$

The eigenvalue–eigenvector solution of $\mathrm{d}\mathbf{V}/\mathrm{d}t = \mathbf{A}\mathbf{V}$

As shown earlier, the solution of

$$\frac{\mathrm{d}\mathbf{V}}{\mathrm{d}t} = \mathbf{A}\mathbf{V},$$

where $\mathbf{A}$ is independent of t, is

$$\mathbf{V}(t) = \{\exp(t\mathbf{A})\}\mathbf{V}(0). \tag{3.37}$$

Put

$$\mathbf{V}(t) = \mathbf{X}\mathbf{Y}(t), \tag{3.38}$$

where $\mathbf{X}$ is the modal matrix of $\mathbf{A}$, of order $(N-1)$, say. Then, by eqns (3.37) and (3.38),

$$\mathbf{X}\mathbf{Y}(t) = \{\exp(t\mathbf{A})\}\mathbf{V}(0) = \{\exp(t\mathbf{A})\}\mathbf{X}\mathbf{Y}(0).$$

Hence by (iii) and (iv),

$$\begin{aligned}\mathbf{Y}(t) &= \mathbf{X}^{-1}\{\exp(t\mathbf{A})\}\mathbf{X}\mathbf{Y}(0)\\ &= \{\exp(t\mathbf{D})\}\mathbf{Y}(0)\\ &= \begin{bmatrix}\exp(t\lambda_1) & & & \\ & \exp(t\lambda_2) & & \\ & & \ddots & \\ & & & \exp(t\lambda_{N-1})\end{bmatrix}\mathbf{Y}(0),\end{aligned}$$

from which it follows that the sth component $Y_s(t)$ of $\mathbf{Y}(t)$ is

$$Y_s(t) = \{\exp(t\lambda_s)\}Y_s(0).$$

Therefore eqn (3.38) gives the solution

$$\begin{bmatrix} V_1(t) \\ V_2(t) \\ \vdots \\ V_{N-1}(t) \end{bmatrix} = \mathbf{X}\begin{bmatrix} Y_1(0)\exp(t\lambda_1) \\ Y_2(0)\exp(t\lambda_2) \\ \vdots \\ Y_{N-1}(0)\exp(t\lambda_{N-1}) \end{bmatrix}, \qquad (3.39)$$

where $\mathbf{XY}(0) = \mathbf{V}(0)$. For known eigenvectors the latter could be solved for the components of $\mathbf{Y}(0)$ by Gaussian elimination.

Solution (3.39) will also hold for problems with derivative boundary-conditions at $x = 0$ and L except that $\mathbf{V}(t) = [V_0(t), V_1(t), \ldots, V_N(t)]^T$, $Nh = L$, and matrix $\mathbf{A}$ will be of order $(N+1)$.

Equation (3.39) can be used to give an approximate solution for sufficiently large positive values of t when the eigenvalues are negative, and I am indebted to D. Drew, Brunel University, for the following example.

Consider Worked Example 2.2, for which $h = 0.1$ and $\lambda_s = -(4/h^2)\sin^2 s\pi/20$, so that

$$\lambda_1 = -9.7887,\ \lambda_2 = -38.197, \ldots, \lambda_9 = -390.2.$$

Then

$$\begin{bmatrix} V_1(t) \\ V_2(t) \\ \vdots \\ V_9(t) \end{bmatrix} = \mathbf{X}\begin{bmatrix} Y_1(0)\exp(-9.8t) \\ Y_2(0)\exp(-38.2t) \\ \vdots \\ \end{bmatrix} \simeq \mathbf{X}\begin{bmatrix} Y_1(0)\exp(-9.8t) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}, \qquad (3.40)$$

for $t > 0.1$, to about 3 significant place accuracy.

As $\mathbf{V}(0) = \mathbf{XY}(0)$, $\mathbf{Y}(0) = \mathbf{X}^{-1}\mathbf{V}(0)$.

But matrix $\mathbf{A}$ is symmetric, so $\mathbf{X}^T\mathbf{X} = \mathbf{I}$, assuming each eigenvector has been normalized to 1, i.e. divided by the square root of the sum of the squares of its components. Therefore,

$$\mathbf{Y}(0) = \mathbf{X}^T\mathbf{V}(0). \qquad (3.41)$$

If $\mathbf{x}_1^T = [x_{11}, x_{21}, \ldots, x_{91}]$, eqn (3.41) gives

$$\begin{aligned} Y_1(0) &= [x_{11}, x_{21}, \ldots, x_{91}]\mathbf{V}(0) \\ &= \sum_{i=1}^{9} x_{i1} V_i(0). \end{aligned}$$

Hence by (3.40),

$$\begin{aligned} V_i(t) &\simeq x_{i1} Y_1(0) \exp(-9.7887t) \\ &= x_{i1}[\exp(-9.7887t)] \sum_{i=1}^{9} x_{i1} V_i(0). \end{aligned} \tag{3.42}$$

In Worked Example 2.2, which is symmetric about $i = 5$, $V_i(0) = 0.2i$, $i = 1(1)5$, and the normalized component x_{i1} of $\mathbf{x}_1$ is $x_{i1} = (\sin i\pi/10)/\sqrt{5}$, $i = 1(1)9$. These values give that $\sum_1^9 x_{i1} V_i(0) = 4.0863/\sqrt{5}$. For $t = 0.1$,

$$\begin{aligned} V_i(0.1) &\simeq \frac{1}{5}\left(\sin \frac{i\pi}{10}\right)[\exp(-0.9789)](4.0863) \\ &= 0.3071 \sin i\pi/10. \end{aligned}$$

Table 3.3 shows that these approximate values differ from the Crank–Nicolson solution for $r = 1$ only in the fourth decimal place.

If matrix **A** is non-symmetric but satisfies the conditions of Exercise 16, Chapter 2, the same method can be used by transforming **A** to the symmetric matrix $\mathbf{DAD}^{-1}$, where **D** is the diagonal matrix of Exercise 16.

TABLE 3.3

x	0.1	0.2	0.3	0.4	0.5
Approximate V_i	0.0949	0.1805	0.2484	0.2920	0.3071
$C-N$ solution	0.0948	0.1803	0.2482	0.2918	0.3069
Solution of P.D.E.	0.0934	0.1776	0.2444	0.2873	0.3021

Miscellaneous methods for improving accuracy

(i) ***Reduction of the local truncation error*** (Douglas equations)

All derivatives can be expressed exactly in terms of infinite series of forward, backward, or central-differences. For example

$$\frac{\partial^2 U}{\partial x^2}=\frac{1}{h^2}(\delta_x^2 U-\tfrac{1}{12}\delta_x^4 U+\tfrac{1}{90}\delta_x^6 U+\ldots) \qquad (3.43)$$

where the subscript x denotes differencing in the x-direction and the central-differences are defined by

$$\delta_x U_{i,j}=U_{i+\frac{1}{2},j}-U_{i-\frac{1}{2},j}$$

and

$$\delta_x^2 U_{i,j}=\delta_x(\delta_x U_{i,j})=U_{i+1,j}-2U_{i,j}+U_{i-1,j},\ \text{etc.} \qquad (3.44)$$

In the approximation methods already considered the right-hand side of (3.43) has been truncated after the first term. If it is truncated after two or more terms the accuracy of the approximation method will always be improved but this normally increases the number of unknowns in an implicit method and complicates the boundary procedure. For equations involving second-order derivatives however it is possible to eliminate the fourth-order central-differences yet leave the number of unknowns unchanged. For example, if the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}$$

is approximated at the point $(i, j+\frac{1}{2})$ by

$$\frac{1}{k}(u_{i,j+1}-u_{i,j})=\frac{1}{2}\left\{\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j+1}+\left(\frac{\partial^2 u}{\partial x^2}\right)_{i,j}\right\}$$

$$=\frac{1}{2h^2}(\delta_x^2-\tfrac{1}{12}\delta_x^4+\tfrac{1}{90}\delta_x^6+\ldots)(u_{i,j+1}+u_{i,j}),$$

then the terms involving δ_x^4 can be eliminated by operating on both sides with $(1+\frac{1}{12}\delta_x^2)$. This gives that

$$(1+\tfrac{1}{12}\delta_x^2)(u_{i,j+1}-u_{i,j})=\tfrac{1}{2}r(\delta_x^2 u_{i,j+1}+\delta_x^2 u_{i,j})+O(\delta_x^6)$$

which can be written as

$$\{1+(\tfrac{1}{12}-\tfrac{1}{2}r)\delta_x^2\}u_{i,j+1}=\{1+(\tfrac{1}{12}+\tfrac{1}{2}r)\delta_x^2\}u_{i,j}$$

when terms of order δ_x^6 are neglected. By (3.44) it follows that the differential equation at the point $(i, j+\frac{1}{2})$ can be approximated by the implicit algebraic equation

$$(1-6r)u_{i-1,j+1}+(10+12r)u_{i,j+1}+(1-6r)u_{i+1,j+1}$$
$$=(1+6r)u_{i-1,j}+(10-12r)u_{i,j}+(1+6r)u_{i+1,j}, \quad (3.45)$$

where $r=k/h^2$. The resulting tridiagonal system of equations can be solved by the algorithm on p. 24 and requires exactly the same amount of arithmetic as the Crank–Nicolson method. Whereas, however, the local truncation error of the Crank–Nicolson equation is $O(h^2)+O(k^2)$, it is $O(h^4)+O(k^2)$ for the Douglas equation. As proved in Chapter 3, Exercise 6, the equations are stable and consistent for all positive r. The numerical solution of Example 2.1 by eqns (3.45) for $h=0.1$ and $r=1$ at $t=0.1$ is compared with the Crank–Nicolson solution for $r=1$ in Table 3.4.

An explicit difference equation of h^4 accuracy is developed in Exercise 7.

(ii) *Use of three time-level difference equations*

The finite-difference approximation of a parabolic equation needs only two time-levels. Three (or more) time-level schemes can be constructed but naturally this is done only to achieve some advantage over two-level schemes, such as a smaller local truncation error, greater stability, or the transformation of a non-linear problem to a linear one as is demonstrated further on in this chapter. For example, the three-level difference equation

$$\frac{3}{2}\frac{(u_{i,j+1}-u_{i,j})}{k}-\frac{1}{2}\frac{(u_{i,j}-u_{i,j-1})}{k}=\frac{(u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1})}{h^2}$$

TABLE 3.4

$x=$	0.1	0.2	0.3	0.4	0.5
Solution of P.D.E.	0.0934	0.1776	0.2444	0.2873	0.3021
Douglas solution	0.0941	0.1789	0.2463	0.2895	0.3044
C–N solution	0.0948	0.1803	0.2482	0.2918	0.3069

approximating $\partial U/\partial t = \partial^2 U/\partial x^2$ has a truncation error of the same order as the Crank–Nicolson equation, namely $O(k^2)+O(h^2)$, but is the better one to use when the initial data is discontinuous or varies very rapidly with x. The Crank–Nicolson approximation should be used when the initial data and its derivatives are continuous (reference 25). In order to solve the first set of equations for $u_{i,2}$ it is necessary to calculate a solution along the first time-level by some other method, it being assumed that the initial data along $t=0$ are known. This first time-level solution must be of the same accuracy as that given by the three-levels equation. A three-level variation of the Douglas equation is

$$\begin{aligned}&\frac{1}{12k}\{\tfrac{3}{2}(u_{i+1,j+1}-u_{i+1,j})-\tfrac{1}{2}(u_{i+1,j}-u_{i+1,j-1})\}\\&\quad+\frac{5}{6k}\{\tfrac{3}{2}(u_{i,j+1}-u_{i,j})-\tfrac{1}{2}(u_{i,j}-u_{i,j-1})\}\\&\quad+\frac{1}{12k}\{\tfrac{3}{2}(u_{i-1,j+1}-u_{i-1,j})-\tfrac{1}{2}(u_{i-1,j}-u_{i-1,j-1})\}\\&=\frac{1}{h^2}(u_{i+1,j+1}-2u_{i,j+1}+u_{i-1,j+1}),\end{aligned}$$

and like the Douglas equation its truncation error is $O(k^2)+O(h^4)$. A number of such schemes for constant and variable coefficients and for one and two space dimensions are discussed in references 18 and 25.

(iii) *Deferred correction method*

In this method the approximating difference equations are solved as usual. Their solution is then used to calculate a correction term, at each mesh point of the solution domain, which is added to the approximating difference equation at each mesh point. The corrected equations are then re-solved and the process repeated if necessary. The correction terms are numbers obtained by differencing the numerical solution in either the x-direction or the t-direction, or both directions. One method for deriving a correction term for the Crank–Nicolson equations is given in Chapter 3, Exercise 9, but the following is better as it is based on a general result. Define the averaging operator μ by $\mu f_{j+\frac{1}{2}} = \frac{1}{2}(f_j + f_{j+1})$ and use the following results which are proved in most introductory books to numerical analysis, namely,

$$k\frac{\partial}{\partial t}\equiv 2\sinh^{-1}(\tfrac{1}{2}\delta_t) \quad \text{and} \quad \mu_t\equiv(1+\tfrac{1}{4}\delta_t^2)^{\frac{1}{2}}.$$

From these it follows that

$$\tfrac{1}{2}k\left\{\left(\frac{\partial U}{\partial t}\right)_{i,j}+\left(\frac{\partial U}{\partial t}\right)_{i,j+1}\right\}=k\mu_t\left(\frac{\partial U}{\partial t}\right)_{i,j+\frac{1}{2}}$$
$$=\{(1+\tfrac{1}{4}\delta_t^2)^{\frac{1}{2}}2\sinh^{-1}(\tfrac{1}{2}\delta_t)\}U_{i,j+\frac{1}{2}}.$$

The expansion of the right-hand side into positive powers of δ_t leads to

$$\tfrac{1}{2}k\left\{\left(\frac{\partial U}{\partial t}\right)_{i,j}+\left(\frac{\partial U}{\partial t}\right)_{i,j+1}\right\}=(\delta_t+\tfrac{1}{12}\delta_t^3-\tfrac{1}{120}\delta_t^5+\ldots)U_{i,j+\frac{1}{2}}$$

which can be rearranged as

$$\delta_t U_{i,j+\frac{1}{2}}=\tfrac{1}{2}k\left\{\left(\frac{\partial U}{\partial t}\right)_{i,j}+\left(\frac{\partial U}{\partial t}\right)_{i,j+1}\right\}+C_t U_{i,j+\frac{1}{2}},$$

giving

$$U_{i,j+1}-U_{i,j}=\tfrac{1}{2}k\frac{\partial}{\partial t}(U_{i,j}+U_{i,j+1})+C_t U_{i,j+\frac{1}{2}}, \qquad (3.46)$$

where

$$C_t\equiv-\tfrac{1}{12}\delta_t^3+\tfrac{1}{120}\delta_t^5+\ldots.$$

Equation (3.46) is a general result relating the value of a continuous function at the $(j+1)$th time-level to its value at the jth time-level in terms of first time-derivatives and central-differences in the t-direction. For the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2}$$

it follows that

$$\frac{\partial}{\partial t}\equiv\frac{\partial^2}{\partial x^2}.$$

Hence by (3.46)

$$U_{i,j+1}-U_{i,j}=\tfrac{1}{2}k\frac{\partial^2}{\partial x^2}(U_{i,j}+U_{i,j+1})+C_t U_{i,j+\frac{1}{2}}.$$

Using eqn (3.43) it is seen that

$$U_{i,j+1}-U_{i,j}=\frac{1}{2}\frac{k}{h^2}(\delta_x^2-\tfrac{1}{12}\delta_x^4+\tfrac{1}{90}\delta_x^6+\ldots)(U_{i,j+1}+U_{i,j})+C_tU_{i,j+\frac{1}{2}}$$
$$=\tfrac{1}{2}r(\delta_x^2U_{i,j+1}+\delta_x^2U_{i,j})+C, \tag{3.47}$$

where

$$C=\tfrac{1}{2}r\{(-\tfrac{1}{12}\delta_x^4U_{i,j+1}+\tfrac{1}{90}\delta_x^6U_{i,j+1}+\ldots)$$
$$+(-\tfrac{1}{12}\delta_x^4U_{i,j}+\tfrac{1}{90}\delta_x^6U_{i,j}+\ldots)\}+(-\tfrac{1}{12}\delta_t^3U_{i,j+\frac{1}{2}}+\tfrac{1}{120}\delta_t^5U_{i,j+\frac{1}{2}}).$$

The first approximation to the solution of (3.47) would be found by putting $C=0$ and solving the Crank–Nicolson equations

$$u_{i,j+1}-u_{i,j}=\tfrac{1}{2}r(\delta_x^2u_{i,j+1}+\delta_x^2u_{i,j})$$

over the solution domain $[0<x<1]\times[0<t\leqslant T]$, say. The correction term at each mesh point would then be calculated from a truncated approximation to C such as

$$C'=-\frac{r}{24}(\delta_x^4u_{i,j+1}+\delta_x^4u_{i,j})-\tfrac{1}{12}\delta_t^3u_{i,j+\frac{1}{2}}.$$

This method, in effect, includes higher-order difference terms in the approximations to the derivatives but keeps the matrix of coefficients of the approximation equations tridiagonal which allows the algorithm on p. 24 to be used.

(iv) *Richardson's deferred approach to the limit*

For this method two or more solutions approximating the problem must be known for two or more different mesh sizes and the difference between the solution of the partial differential equation and the solution of the approximating equations must be known as a function of the mesh lengths.

Let U represent the solution of the differential equation and $u_{r,s}$ represent the solution of the finite-difference equations for a mesh of size (h_r, k_s). Now assume, for example, that the discretization error

$$U-u(h,k)=Ak+Bh^2+Ck^2+Dh^4+\ldots,$$

as it is for the classical explicit method for finite t.

If two solutions $u_{1,1}$ and $u_{2,2}$ are known then

$$U - u_{1,1} = Ak_1 + Bh_1^2 + Ck_1^2 + Dh_1^4 + \dots$$

and

$$U - u_{2,2} = Ak_2 + Bh_2^2 + Ck_2^2 + Dh_2^4 + \dots.$$

Hence B (or A, but not both) can be eliminated to give that

$$U = \frac{1}{h_2^2 - h_1^2}(h_2^2 u_{1,1} - h_1^2 u_{2,2}) + A\frac{k_1 h_2^2 - k_2 h_1^2}{h_2^2 - h_1^2} + \dots.$$

If the term involving A is negligible then U is an improvement on $u_{1,1}$ and $u_{2,2}$. For the special case $h_2 = 2h_1$, $k_2 = 2k_1$,

$$U = \tfrac{1}{3}(4u_{1,1} - u_{2,2}) + \tfrac{2}{3}kA + \dots.$$

If three different solutions are known then A and B can be eliminated. For the Crank–Nicolson equations,

$$U - u(h, k) = Ah^2 + Bk^2 + Ch^4 + \dots.$$

If $h_2 = \lambda h_1$ and $k_2 = \lambda k_1$ it is easily shown that

$$U = u_{2,2} + \frac{\lambda^2}{1 - \lambda^2}(u_{2,2} - u_{1,1}) + O(h^4).$$

Solution of non-linear parabolic equations

There is no difficulty in formally applying finite-difference methods to non-linear parabolic equations. The difficulties are associated with the difference equations themselves. If they are linear they can usually be solved quite easily, although we still have the problem of determining the conditions that must be satisfied for stability and convergence because the coefficients of the unknowns will be functions of the solution at earlier time-levels. If they are non-linear we have also the problem of their solution. Direct methods, in general, are difficult, so they are usually solved iteratively after being linearized in some way. Taylor's expansion provides a standard way of doing this and the method is usually referred to as Newton's method.

Linearization by Newton's method

Let

$$f_i(u_1, u_2, \dots, u_N) = 0, \quad i = 1(1)N, \tag{3.48}$$

represent N equations in the N dependent variables $u_1, u_2, \ldots, u_N$.

Let V_i be a known approximation to the exact solution value u_i, $i = 1(1)N$.

Put $u_i = V_i + \varepsilon_i$ and substitute into eqn (3.48). Then by Taylor's expansion to first-order terms in ε_i, $i = 1(1)N$,

$$f_i(V_1, V_2, \ldots, V_N) + \left[\frac{\partial f_i}{\partial u_1}\varepsilon_1 + \frac{\partial f_i}{\partial u_2}\varepsilon_2 + \ldots + \frac{\partial f_i}{\partial u_N}\varepsilon_N\right]_{u_i = V_i} = 0,$$
$$i = 1(1)N. \quad (3.49)$$

The subscript notation on the second bracket indicates that the dependent variables $u_1, u_2, \ldots, u_N$ appearing in the coefficients of $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ are replaced by $V_1, V_2, \ldots, V_N$ respectively after the differentiations. Equation (3.49) represents N linear equations for the N unknowns $\varepsilon_1, \varepsilon_2, \ldots, \varepsilon_N$ because V_1, V_2, V_N are known. When the ε's have been calculated the process is repeated, the starting values of the dependent variables for the next iteration being $(V_i + \varepsilon_i)$, $i = 1(1)N$. This process of successive approximations is continued until the u_i's have been found to the required degree of accuracy, such as $|\varepsilon_i| < 10^{-8}$, $i = 1(1)N$. Some numerical results for a particular problem are given on p. 148.

Example 3.1

The function U satisfies the non-linear equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U^2}{\partial x^2}, \quad 0 < x < 1, \quad t > 0,$$

the initial condition $U = 4x(1-x)$, $0 < x < 1$, $t = 0$, and the boundary conditions $U = 0$ at $x = 0$ and 1, $t \geqslant 0$.

If the equation is approximated at the point $\{ih, (j+\frac{1}{2})k\}$ in the x–t plane by the difference scheme

$$\frac{1}{k}\delta_t u_{i,j+\frac{1}{2}} = \frac{1}{2h^2}(\delta_x^2 u_{i,j+1}^2 + \delta_x^2 u_{i,j}^2),$$

use Newton's method to derive a set of linear equations giving an improved value $(V_i + \varepsilon_i)$ to the approximate value V_i at the mesh points defined by $x_i = \frac{1}{6}i$, $i = 1(1)5$, $t = k = \frac{1}{36}$.

If the V_i are taken equal to the initial values at $x_i = \frac{1}{6}i$,

$i=1(1)5$, $t=0$, show that these equations reduce to

$$-19\varepsilon_1+8\varepsilon_2+\tfrac{14}{9}=0,$$
$$5\varepsilon_1-25\varepsilon_2+9\varepsilon_3-\tfrac{22}{9}=0,$$

and

$$16\varepsilon_2-27\varepsilon_3-\tfrac{34}{9}=0.$$

The approximation equation is

$$\frac{1}{k}(u_{i,j+1}-u_{i,j})=\frac{1}{2h^2}\{(u^2_{i-1,j+1}-2u^2_{i,j+1}+u^2_{i+1,j+1})$$
$$+(u^2_{i-1,j}-2u^2_{i,j}+u^2_{i+1,j})\}.$$

Put $p=h^2/k$ and denote $u_{i,j+1}$ by u_i. The equation can then be written as

$$u^2_{i-1}-2(u^2_i+pu_i)+u^2_{i+1}+\{u^2_{i-1,j}-2(u^2_{i,j}-pu_{i,j})+u^2_{i+1,j}\}$$
$$=0\equiv f_i(u_{i-1},u_i,u_{i+1}).$$

By eqn (3.49),

$$\left[\frac{\partial f_i}{\partial u_{i-1}}\varepsilon_{i-1}+\frac{\partial f_i}{\partial u_i}\varepsilon_i+\frac{\partial f_i}{\partial u_{i+1}}\varepsilon_{i+1}\right]_{u_i=V_i}+f_i(V_{i-1},V_i,V_{i+1})=0,$$

hence

$$2V_{i-1}\varepsilon_{i-1}-2(2V_i+p)\varepsilon_i+2V_{i+1}\varepsilon_{i+1}+[\{V^2_{i-1}-2(V^2_i+pV_i)+V^2_{i+1}\}$$
$$+\{u^2_{i-1,j}-2(u^2_{i,j}-pu_{i,j})+u^2_{i+1,j}\}]=0, \quad (3.50)$$

where V_i is an approximation to $u_{i,j+1}$.
The problem is symmetric with respect to $x=\frac{1}{2}$. When the V_i are taken equal to $u_{i,0}$, eqn (3.50) for $j=0$ reduces to

$$2u_{i-1,0}\varepsilon_{i-1}-2(2u_{i,0}+p)\varepsilon_i+2u_{i+1,0}\varepsilon_{i+1}$$
$$+\{2u^2_{i-1,0}-4u^2_{i,0}+2u^2_{i+1,0}\}=0.$$

This gives the equations quoted for $p=1$ and the initial values indicated in Fig. 3.5 since $\varepsilon_0=0$.

Richtmyer's linearization method

Richtmyer, reference 25, considers the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U^m}{\partial x^2},\quad m \text{ a positive integer} \geqslant 2,$$

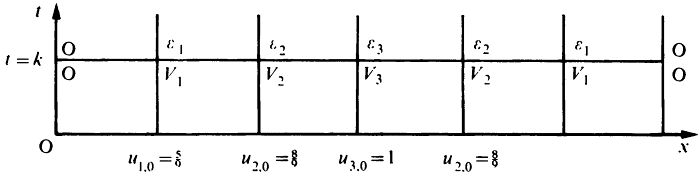


Fig. 3.5

which he approximates by the implicit weighted average difference scheme

$$\frac{1}{k}(u_{i,j+1}-u_{i,j})=\frac{1}{h^2}[\theta\,\delta_x^2(u_{i,j+1}^m)+(1-\theta)\,\delta_x^2(u_{i,j}^m)]. \qquad (3.51)$$

By Taylor's expansion about the point (i, j),

$$u_{i,j+1}^m=u_{i,j}^m+k\frac{\partial u_{i,j}^m}{\partial t}+\dots$$

$$=u_{i,j}^m+k\frac{\partial u_{i,j}^m}{\partial u_{i,j}}\frac{\partial u_{i,j}}{\partial t}+\dots.$$

Hence to terms of order k,

$$u_{i,j+1}^m=u_{i,j}^m+mu_{i,j}^{m-1}(u_{i,j+1}-u_{i,j}),$$

a result which replaces the non-linear unknown $u_{i,j+1}^m$ by an approximation linear in $u_{i,j+1}$.

Putting $\omega_i=u_{i,j+1}-u_{i,j}$ in eqn (3.51) leads to

$$\frac{1}{k}\omega_i=\frac{1}{h^2}[\theta\,\delta_x^2(u_{i,j}^m+mu_{i,j}^{m-1}\omega_i)+(1-\theta)\,\delta_x^2u_{i,j}^m]$$

$$=\frac{1}{h^2}[m\theta\,\delta_x^2u_{i,j}^{m-1}\omega_i+\delta_x^2u_{i,j}^m]$$

$$=\frac{1}{h^2}[m\theta(u_{i-1,j}^{m-1}\omega_{i-1}-2u_{i,j}^{m-1}\omega_i+u_{i+1,j}^{m-1}\omega_{i+1})$$

$$+(u_{i-1,j}^m-2u_{i,j}^m+u_{i+1,j}^m)],$$

which gives a set of linear equations for the ω_i. The solution at the $(j+1)$th time-level is obtained from $u_{i,j+1}=\omega_i+u_{i,j}$. For known boundary values at $x=0$ and 1, where $Nh=1$ and $r=$

k/h^2, the equations in matrix form for $m=2$ are

$$\begin{bmatrix} (1+4r\theta u_{1,j}) & -2r\theta u_{2,j} & & & \\ -2r\theta u_{1,j} & (1+4r\theta u_{2,j}) & -2r\theta u_{3,j} & & \\ & & & & \\ & & -2r\theta u_{N-3,j} & (1+4r\theta u_{N-2,j}) & -2r\theta u_{N-1,j} \\ & & & -2r\theta u_{N-2,j} & (1+4r\theta u_{N-1,j}) \end{bmatrix} \begin{bmatrix} \omega_1 \\ \omega_2 \\ \\ \omega_{N-2} \\ \omega_{N-1} \end{bmatrix}$$

$$= \begin{bmatrix} -2ru_{1,j} & ru_{2,j} & & & \\ ru_{1,j} & -2ru_{2,j} & ru_{3,j} & & \\ & & \cdot & & \\ & & & \cdot & \\ & & & & \cdot \\ & & & ru_{N-2,j} & -2ru_{N-1,j} \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \\ \\ \\ u_{N-1,j} \end{bmatrix}$$

$$+ \begin{bmatrix} ru_{0,j}^2 + 2r\theta u_{0,j}(u_{0,j+1}-u_{0,j}) \\ 0 \\ 0 \\ \vdots \\ 0 \\ ru_{N,j}^2 + 2r\theta u_{N,j}(u_{N,j+1}-u_{N,j}) \end{bmatrix} \tag{3.52}$$

These are easily solved by the algorithm on p. 24. Numerical results for a particular problem are given on p. 148.

A three time-level method

Lees, reference 16, considered the non-linear equation

$$b(U)\frac{\partial U}{\partial t} = \frac{\partial}{\partial x}\left\{a(U)\frac{\partial U}{\partial x}\right\}, \quad a(U)>0, \quad b(U)>0, \tag{3.53}$$

and investigated a difference scheme that

(i) achieved linearity in the unknowns $u_{i,j+1}$ by evaluating all coefficients of $u_{i,j+1}$ at a time-level of known solution values,

(ii) preserved stability by averaging $u_{i,j}$ over three time-levels, and

(iii) maintained accuracy by using central-difference approximations.

As

$$\left(\frac{\partial U}{\partial x}\right)_{i,j} \simeq \frac{1}{h}(U_{i+\frac{1}{2},j} - U_{i-\frac{1}{2},j}) = \frac{1}{h}\,\delta_x U_{i,j},$$

an obvious central-difference approximation to (3.53) is

$$b(u_{i,j})\frac{1}{2k}(u_{i,j+1}-u_{i,j-1})=\frac{1}{h}\delta_x\left\{a(u_{i,j})\frac{1}{h}\delta_x u_{i,j}\right\}$$

$$=\frac{1}{h^2}\delta_x\{a(u_{i,j})(u_{i+\frac{1}{2},j}-u_{i-\frac{1}{2},j})\}$$

$$=\frac{1}{h^2}\{a(u_{i+\frac{1}{2},j})(u_{i+1,j}-u_{i,j})-a(u_{i-\frac{1}{2},j})(u_{i,j}-u_{i-1,j})\},$$

but this is certainly unstable for $a=b=1$ (reference Chapter 2, Exercise 26(b)). If, however, $u_{i+1,j}$, $u_{i,j}$, and $u_{i-1,j}$ are replaced by $\frac{1}{3}(u_{i+1,j+1}+u_{i+1,j}+u_{i+1,j-1})$, $\frac{1}{3}(u_{i,j+1}+u_{i,j}+u_{i,j-1})$, and $\frac{1}{3}(u_{i-1,j+1}+u_{i-1,j}+u_{i-1,j-1})$ respectively, and the coefficients $a(u_{i+\frac{1}{2},j})$ and $a(u_{i-\frac{1}{2},j})$ are replaced by $a\{\frac{1}{2}(u_{i+1,j}+u_{i,j})\}$ and $a\{\frac{1}{2}(u_{i,j}+u_{i-1,j})$ respectively in order to avoid midpoint values of u, Lees proved that for sufficiently small values of h and k,

$$\max_{i,j}|U_{i,j}-u_{i,j}|\leq A(h^2+k^2),$$

where A is a constant.

For this method the equation considered in the preceding section needs to be written as

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U^m}{\partial x^2}=\frac{\partial}{\partial x}\left(mU^{m-1}\frac{\partial U}{\partial x}\right).$$

A comparison of results for a particular problem

If it is assumed that $U=U(x-vt)$, v constant, is a solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U^2}{\partial x^2},\quad 0<x<1,\quad t>0, \tag{3.54}$$

substitution into (3.54) and integration with respect to $(x-vt)$ gives that

$$\frac{A}{v}\log\left(U-\frac{A}{v}\right)+U=B-\tfrac{1}{2}v(x-vt),$$

where A and B are constants. Choosing $A=1$ and $v=2$, then $B=U(0,0)=1.5$ leads to the particular solution

$$(2U-3)+\log(U-\tfrac{1}{2})=2(2t-x). \tag{3.55}$$

For any given x and t eqn (3.55) can be solved iteratively for U by, for example, the Newton–Raphson method. In a similar way initial values for this particular solution can be found by putting $t = 0$ in (3.55), and $x = 0(0.1)1$, say. Boundary values at $x = 0$ and 1 for known values of t can be found in the same manner. These initial values and boundary values can then be used as the boundary data for approximation methods whose accuracy can be checked against the analytical solution of the differential equation.

The solution of eqns (3.51) for $\theta = \frac{1}{2}$, $m = 2$, $h = 0.1$, $r = k/h^2 = \frac{1}{2}$, and $t = 0.5$, which corresponds to 100 time-steps, by Newton's, Richtmyer's, and the three time-levels methods, are given below. All agree with the analytical solution of the differential equation to five decimal places after rounding, and with each other to 6*D*.

x	Solution of the P.D.E.	Newton's, Richtmyer's, and the three time-levels methods
0.1	2.149703	2.149701
0.3	1.997951	1.997948
0.5	1.849962	1.849958
0.7	1.706244	1.706240
0.9	1.567391	1.567389

The stability of three of more time-level difference equations (Fixed mesh lengths)

The following theorem is useful for the matrix method of analysis of the stability of three or more time-level difference equations and is easier to use than one might at first think.

Theorem: If the matrix $\mathbf{A}$ can be written as

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1} & \mathbf{A}_{1,2} & \cdots & \mathbf{A}_{1,m} \\ \mathbf{A}_{2,1} & \mathbf{A}_{2,2} & \cdots & \mathbf{A}_{2,m} \\ \vdots & & & \vdots \\ \mathbf{A}_{m,1} & \mathbf{A}_{m,2} & \cdots & \mathbf{A}_{m,m} \end{bmatrix},$$

where each $\mathbf{A}_{i,j}$ is an $n \times n$ matrix, and all the $\mathbf{A}_{i,j}$ have a common set of n linearly independent eigenvectors, then the eigenvalues of **A** are given by the eigenvalues of the matrices

$$\begin{bmatrix} \lambda_{1,1}^{(k)} & \lambda_{1,2}^{(k)} & \cdots & \lambda_{1,m}^{(k)} \\ \lambda_{2,1}^{(k)} & \lambda_{2,2}^{(k)} & \cdots & \lambda_{2,m}^{(k)} \\ \vdots & & & \vdots \\ \lambda_{m,1}^{(k)} & \lambda_{m,2}^{(k)} & \cdots & \lambda_{m,m}^{(k)} \end{bmatrix}, \quad k = 1(1)n,$$

where $\lambda_{i,j}^{(k)}$ is the kth eigenvalue of $\mathbf{A}_{i,j}$ corresponding to the kth eigenvector $\mathbf{v}_k$ common to all the $\mathbf{A}_{i,j}$'s.

Proof

Let $\mathbf{v}_k$ be an eigenvector common to all the submatrices $\mathbf{A}_{i,j}$, $i, j = 1(1)m$, and denote the corresponding eigenvalues of $\mathbf{A}_{1,1}, \mathbf{A}_{2,1}, \ldots$ by $\lambda_{1,1}^{(k)}, \lambda_{2,1}^{(k)}, \ldots$ respectively. For simplicity consider $i, j = 1(1)2$ and denote $\mathbf{v}_k$ by $\mathbf{v}$, $\lambda_{i,j}^{(k)}$ by $\lambda_{i,j}$. Then

$$\mathbf{A}_{1,1}\mathbf{v} = \lambda_{1,1}\mathbf{v}, \quad \mathbf{A}_{1,2}\mathbf{v} = \lambda_{1,2}\mathbf{v},$$
$$\mathbf{A}_{2,1}\mathbf{v} = \lambda_{2,1}\mathbf{v}, \quad \mathbf{A}_{2,2}\mathbf{v} = \lambda_{2,2}\mathbf{v}.$$

Multiply these equations respectively by the non-zero constants α_1, α_2, α_1, and α_2 and write them as

$$\begin{bmatrix} \mathbf{A}_{1,1}\mathbf{A}_{1,2} \\ \mathbf{A}_{2,1}\mathbf{A}_{2,2} \end{bmatrix} \begin{bmatrix} \alpha_1\mathbf{v} \\ \alpha_2\mathbf{v} \end{bmatrix} = \begin{bmatrix} (\lambda_{1,1}\alpha_1 + \lambda_{1,2}\alpha_2)\mathbf{v} \\ (\lambda_{2,1}\alpha_1 + \lambda_{2,2}\alpha_2)\mathbf{v} \end{bmatrix}. \tag{3.56}$$

Assume now that

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_{1,1}\mathbf{A}_{1,2} \\ \mathbf{A}_{2,1}\mathbf{A}_{2,2} \end{bmatrix}$$

has an eigenvalue μ corresponding to the eigenvector

$$\begin{bmatrix} \alpha_1\mathbf{v} \\ \alpha_2\mathbf{v} \end{bmatrix}$$

so that

$$\begin{bmatrix} \mathbf{A}_{1,1}\mathbf{A}_{1,2} \\ \mathbf{A}_{2,1}\mathbf{A}_{2,2} \end{bmatrix} \begin{bmatrix} \alpha_1\mathbf{v} \\ \alpha_2\mathbf{v} \end{bmatrix} = \mu \begin{bmatrix} \alpha_1\mathbf{v} \\ \alpha_2\mathbf{v} \end{bmatrix}. \tag{3.57}$$

By the right-hand sides of eqns (3.56) and (3.57),

$$(\lambda_{1,1} - \mu)\alpha_1 + \lambda_{1,2}\alpha_2 = 0$$

and

$$\lambda_{2,1}\alpha_1 + (\lambda_{2,2} - \mu)\alpha_2 = 0.$$

These two equations will have a non-trivial solution for α_1 and α_2 if and only if

$$\det\begin{bmatrix}(\lambda_{1,1}-\mu) & \lambda_{1,2}\\ \lambda_{2,1} & (\lambda_{2,2}-\mu)\end{bmatrix}=0,$$

i.e. if and only if μ is an eigenvalue of the matrix

$$\begin{bmatrix}\lambda_{1,1} & \lambda_{1,2}\\ \lambda_{2,1} & \lambda_{2,2}\end{bmatrix}.$$

Matrices with common eigenvector systems

Proofs of the following theorems can be found in reference 30.

(i) If the $N\times N$ matrix $\mathbf{A}$ has N distinct eigenvalues λ_s it has N unique linearly independent eigenvectors $\mathbf{v}_s$, $s=1(1)N$. As proved earlier, any polynomial $f(\mathbf{A})$ of $\mathbf{A}$ has the same set of eigenvectors $\mathbf{v}_s$ and a corresponding set of eigenvalues $f(\lambda_s)$.

(ii) All $N\times N$ Hermitian matrices, which includes real symmetric matrices, have N linearly independent eigenvectors.

(iii) If the matrices $\mathbf{A}$ and $\mathbf{B}$ commute and have linear elementary divisors then they have a common system of eigenvectors. In particular, all matrices with distinct eigenvalues, all Hermitian, and therefore all real symmetric matrices, have linear elementary divisors.

(iv) Let $\mathbf{A}$ and $\mathbf{B}$ be matrices with a common system of eigenvectors. Let λ and μ be the eigenvalues of $\mathbf{A}$ and $\mathbf{B}$ respectively corresponding to the common eigenvector $\mathbf{v}$. Then $\mathbf{v}$ is an eigenvector of $\mathbf{AB}$ and $\mathbf{A}^{-1}\mathbf{B}$ and the corresponding eigenvalues are $\lambda\mu$ and $\lambda^{-1}\mu$ respectively. These results are easily proved. By hypothesis, $\mathbf{Bv}=\mu\mathbf{v}$ and $\mathbf{Av}=\lambda\mathbf{v}$. Therefore $\mathbf{ABv}=\mu\mathbf{Av}=\mu\lambda\mathbf{v}$. Also $\mathbf{A}^{-1}\mathbf{Bv}=\mu\mathbf{A}^{-1}\mathbf{v}=(\mu/\lambda)\mathbf{v}$.

Example 3.2

Investigate the stability of the Du Fort and Frankel approximation to the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad 0<x<1,\quad t>0,$$

for fixed mesh lengths, given that U is known on the boundaries $x=0$ and 1 for $t>0$. (See Worked Example 2.7.)

The approximation at the point (ih, jk) is

$$\frac{1}{2k}(u_{i,j+1}-u_{i,j-1})=\frac{1}{h^2}\{u_{i-1,j}-(u_{i,j-1}+u_{i,j+1})+u_{i+1,j}\},$$

which may be written as

$$(1+2r)u_{i,j+1}=2r(u_{i-1,j}+u_{i+1,j})+(1-2r)u_{i,j-1},$$

where $r=k/h^2$. For known boundary-values and $Nh=1$ these equations in matrix form are

$$(1+2r)\begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ u_{3,j+1} \\ \vdots \\ u_{N-1,j+1} \end{bmatrix} = 2r\begin{bmatrix} 0 & 1 & & & \\ 1 & 0 & 1 & & \\ & 1 & 0 & 1 & \\ & & & \ddots & \\ & & & 1 & 0 \end{bmatrix}\begin{bmatrix} u_{1,j} \\ u_{2,j} \\ u_{3,j} \\ \vdots \\ u_{N-1,j} \end{bmatrix}$$

$$+(1-2r)\begin{bmatrix} u_{1,j-1} \\ u_{2,j-1} \\ u_{3,j-1} \\ \vdots \\ u_{N-1,j-1} \end{bmatrix} + 2r\begin{bmatrix} u_{0,j} \\ 0 \\ 0 \\ \vdots \\ u_{N,j} \end{bmatrix}$$

giving

$$\mathbf{u}_{j+1}=\frac{2r}{1+2r}\mathbf{A}\mathbf{u}_j+\frac{1-2r}{1+2r}\mathbf{u}_{j-1}+\mathbf{c}_j, \tag{3.58}$$

where $\mathbf{A}$ is as displayed and $\mathbf{c}_j$ is a vector of known values. Put

$$\mathbf{v}_j=\begin{bmatrix} \mathbf{u}_j \\ \mathbf{u}_{j-1} \end{bmatrix}.$$

Then eqn (3.58) and the identity $\mathbf{u}_j=\mathbf{u}_j$ can be written as

$$\left[\begin{array}{c} \mathbf{u}_{j+1} \\ \hdashline \mathbf{u}_j \end{array}\right] = \left[\begin{array}{c:c} \dfrac{2r}{1+2r}\mathbf{A} & \dfrac{(1-2r)}{(1+2r)}\mathbf{I}_{N-1} \\ \hdashline \mathbf{I}_{N-1} & \mathbf{O} \end{array}\right]\left[\begin{array}{c} \mathbf{u}_j \\ \hdashline \mathbf{u}_{j-1} \end{array}\right]+\left[\begin{array}{c} \mathbf{c}_j \\ \hdashline \mathbf{O} \end{array}\right],$$

where $\mathbf{I}_{N-1}$ is the unit matrix of order $(N-1)$, i.e. as

$$\mathbf{v}_{j+1}=\mathbf{P}\mathbf{v}_j+\mathbf{d}_j,$$

where $\mathbf{P}$ is the matrix shown and $\mathbf{d}_j$ a column vector of known constants. This technique has reduced a three-level difference equation to a two-level one. The equations will be stable when each eigenvalue of $\mathbf{P}$ has a modulus $\leqslant 1$. The matrix $\mathbf{A}$ has $(N-1)$ different eigenvalues so it has $(N-1)$ linearly independent eigenvectors $\mathbf{v}_s$, $s=1(1)(N-1)$. Although the matrix $\mathbf{I}_{N-1}$ has $(N-1)$ eigenvalues each equal to 1 it has $(N-1)$ linearly independent eigenvectors which may be taken as $\mathbf{v}_s$, $s=1(1)(N-1)$, because the eigenvalue equation $\mathbf{Bx}=\lambda\mathbf{x}$ is clearly satisfied by $\mathbf{I}_{N-1}\mathbf{v}_s=1\cdot\mathbf{v}_s$. Hence the eigenvalues λ of P are the eigenvalues of the matrix

$$\begin{bmatrix} \dfrac{2r\lambda_k}{1+2r} & \dfrac{1-2r}{1+2r} \\ 1 & 0 \end{bmatrix},$$

where λ_k is the kth eigenvalue of $\mathbf{A}$. For such a simple case we can work from first principles and find λ by evaluating

$$\det\begin{bmatrix} \left\{\dfrac{2r\lambda_k}{1+2r}-\lambda\right\} & \dfrac{1-2r}{1+2r} \\ 1 & -\lambda \end{bmatrix}=0,$$

giving

$$\lambda^2-\frac{2r\lambda_k}{1+2r}\lambda-\frac{1-2r}{1+2r}=0.$$

By the formula on p. 59

$$\lambda_k=2\cos(k\pi/N),\quad k=1(1)(N-1).$$

Hence

$$\lambda=\left\{2r\cos\frac{k\pi}{N}\pm\left(1-4r^2\sin^2\frac{k\pi}{N}\right)^{\frac{1}{2}}\right\}\Big/(1+2r).$$

Case (*i*) $\quad 1>1-4r^2\sin^2\dfrac{k\pi}{N}\geqslant 0.$

Then

$$|\lambda|<\frac{2r+1}{1+2r}=1.$$

Case (*ii*) $1-4r^2\sin^2\dfrac{k\pi}{N}<0.$

Then

$$|\lambda|^2=\frac{1}{(2r+1)^2}\left\{\left(2r\cos\frac{k\pi}{N}\right)^2+4r^2\sin^2\frac{k\pi}{N}-1\right\}$$

$$=\frac{4r^2-1}{4r^2+4r+1}<1\quad\text{since}\quad r>0.$$

Therefore the equations, for a fixed mesh size, are unconditionally stable for all positive r. It is shown in reference 25 that these equations are also stable for all $r>0$ as the mesh lengths tend to zero.

Brief introduction to the analytical solution of homogeneous finite-difference equations

Linear equations with constant coefficients

Consider the difference equation

$$u_{j+2}+au_{j+1}+bu_j=0,\quad j=0,1,2,\ldots,\tag{3.59}$$

where a and b are real constants.

Assume that

$$u_j=Am^j$$

is a solution, where A and m are non-zero constants. Substitution into (3.59) shows that m is a root of the quadratic equation

$$m^2+am+b=0.\tag{3.60}$$

Case (*i*) Roots real and distinct, $m=m_1$ and $m=m_2$, say. One solution is $u_j=Am_1^j$ and another is $u_j=Bm_2^j$ where A and B are arbitrary constants. As eqn (3.59) is linear in u its general solution is

$$u_j=Am_1^j+Bm_2^j.$$

Case (*ii*) Repeated roots, $m=m_1$ twice, say. Clearly one solution is $u_j=Am_1^j$.

Put $u_j=m_1^jf(j)$. Substitution into (3.59) and the use of $a=-2m_1$, $b=m_1^2$ leads to

$$f(j+2)-2f(j+1)+f(j)=0.$$

By inspection it is seen that $f(j)=j$ satisfies this equation. Therefore a second solution of (3.59) is $u_j = Bjm_1^j$. Hence the solution of eqn (3.59) in this case is

$$u_j = (A+Bj)m_1^j.$$

Case (iii) Complex roots.
Because a and b are real the roots of (3.60) will be conjugate complex numbers, $m_1 = re^{i\theta}$ and $m_2 = re^{-i\theta}$, say, where $i=\sqrt{(-1)}$.

Hence $a = -r(e^{i\theta}+e^{-i\theta}) = -2r\cos\theta$ and $b=r^2$. As in Case (i) the solution of (3.59) is

$$u_j = Ar^j e^{ij\theta} + Br^j e^{-ij\theta} = r^j\{(A+B)\cos j\theta + i(A-B)\sin j\theta\}.$$

Since A and B are arbitrary constants and $r=b^{\frac{1}{2}}$, this can be written as

$$u_j = b^{\frac{1}{2}j}(C\cos j\theta + D\sin j\theta),$$

where C and D are arbitrary constants and $\cos\theta = -a/2r = -a/2\sqrt{b}$. Methods for deriving particular integrals for non-homogeneous difference equations are given in *Finite Difference Equations* by H. Levy and F. Lessman (Pitman).

The eigenvalues and vectors of a common tridiagonal matrix

Let

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

be a square matrix of order N, where a, b, and c may be real or complex numbers.

Let λ represent an eigenvalue of $\mathbf{A}$ and $\mathbf{v}$ the corresponding eigenvector with components $v_1, v_2, \ldots, v_N$. Then the eigenvalue equation $\mathbf{A}\mathbf{v} = \lambda\mathbf{v}$ gives

$$(a-\lambda)v_1 + bv_2 = 0$$
$$cv_1 + (a-\lambda)v_2 + bv_3 = 0$$
$$\cdots$$
$$cv_{j-1} + (a-\lambda)v_j + bv_{j+1} = 0$$

and

$$cv_{N-1}+(a-\lambda)v_N=0.$$

If we define $v_0=v_{N+1}=0$ then these N equations can be combined into the single difference equation

$$cv_{j-1}+(a-\lambda)v_j+bv_{j+1}=0, \quad j=1(1)N. \tag{3.61}$$

As shown, previously, the solution of (3.61) is

$$v_j=Bm_1^j+Cm_2^j, \tag{3.62}$$

where B and C are arbitrary constants and m_1, m_2 are the roots of the equation

$$c+(a-\lambda)m+bm^2=0. \tag{3.63}$$

(It is proved later that the roots cannot be equal.)
By eqn (3.62) it follows, since $v_0=v_{N+1}=0$, that,

$$0=B+C$$

and

$$0=Bm_1^{N+1}+Cm_2^{N+1}.$$

Hence

$$\left(\frac{m_1}{m_2}\right)^{N+1}=1=\mathrm{e}^{i2s\pi}, \quad s=1(1)N,$$

where $i=\sqrt{(-1)}$. Therefore

$$\frac{m_1}{m_2}=\mathrm{e}^{i2s\pi/(N+1)}. \tag{3.64}$$

By eqn (3.63)

$$m_1m_2=\frac{c}{b}, \tag{3.65}$$

and elimination of m_2 between (3.64) and (3.65) leads to

$$m_1=\left(\frac{c}{b}\right)^{\frac{1}{2}}\mathrm{e}^{is\pi/(N+1)}.$$

Similarly,

$$m_2=\left(\frac{c}{b}\right)^{\frac{1}{2}}\mathrm{e}^{-is\pi/(N+1)}.$$

Again, by eqn (3.63)

$$m_1+m_2=(\lambda-a)/b,$$

giving that

$$\lambda = a + b\left(\frac{c}{b}\right)^{\frac{1}{2}}(e^{is\pi/(N+1)} + e^{-is\pi/(N+1)}).$$

Hence the N eigenvalues are given by

$$\lambda_s = a + 2b\left(\frac{c}{b}\right)^{\frac{1}{2}} \cos\frac{s\pi}{N+1}, \quad s = 1(1)N.$$

The jth component of the eigenvector is

$$\begin{aligned} v_j = Bm_1^j + Cm_2^j &= B\left(\frac{c}{b}\right)^{\frac{1}{2}j}(e^{ijs\pi/N+1} - e^{-ijs\pi/N+1}) \\ &= 2iB\left(\frac{c}{b}\right)^{\frac{1}{2}j} \sin\frac{js\pi}{N+1}, \end{aligned}$$

so the eigenvector $\mathbf{v}_s$ corresponding to λ_s can be taken as

$$\begin{aligned} \mathbf{v}_s^T = \Bigg\{\left(\frac{c}{b}\right)^{\frac{1}{2}} \sin\frac{s\pi}{N+1}, \frac{c}{b}\sin\frac{2s\pi}{N+1}, \left(\frac{c}{b}\right)^{\frac{3}{2}} \\ \times \sin\frac{3s\pi}{N+1}, \ldots, \left(\frac{c}{b}\right)^{N/2} \sin\frac{Ns\pi}{N+1}\Bigg\}. \end{aligned}$$

It is easily shown that the roots of eqn (3.63) cannot be equal because if we assume $m_1 = m_2$ the solution of (3.63) is then

$$v_j = (B + Cj)m_1^j$$

and $v_0 = v_{N+1} = 0$ implies that $B = C = 0$, giving $\mathbf{v} = 0$, which is not possible.

An analytical solution of the classical explicit approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1, \quad t > 0,$$

where $U = 0$ at $x = 0$ and 1, $t > 0$, and U is known when $t = 0$, $0 \leqslant x \leqslant 1$.

The classical explicit approximation to the differential equation is

$$u_{i,j+1} = ru_{i-1,j} + (1-2r)u_{i,j} + ru_{i+1,j}, \tag{3.66}$$

where

$$x = ih, \quad t = jk, \quad r = k/h^2, \quad \text{and} \quad Nh = 1.$$

Assume that a solution of eqn (3.66) is of the form

$$u_{i,j} = f_i g_j. \tag{3.67}$$

Substitution of (3.67) into (3.66) leads to

$$\frac{g_{j+1}}{g_j} = \frac{rf_{i-1} + (1-2r)f_i + rf_{i+1}}{f_i}. \tag{3.68}$$

Since the left-hand side of (3.68) is independent of i and the right-hand side is independent of j it follows that both sides must equal a parameter c which is independent of i and j. This gives two homogeneous difference equations for f_i and g_j, namely,

$$g_{j+1} - cg_j = 0 \tag{3.69}$$

and

$$f_{i+1} + \frac{(1-2r-c)}{r} f_i + f_{i-1} = 0. \tag{3.70}$$

The solution of eqn (3.69) is

$$g_j = Ac^j. \tag{3.71}$$

As the solution of the partial differential equation is periodic in x it is reasonable to assume that the solution of (3.70) is periodic in i, so that

$$f_i = B \cos i\theta + D \sin i\theta, \tag{3.72}$$

where

$$\cos \theta = (2r + c - 1)/2r. \tag{3.73}$$

Then, by (3.72), the boundary condition $u_{0,j} = 0$ for all j gives that

$$f_0 = 0 = B.$$

Similarly, the condition $u_{N,j} = 0$ for all j gives that

$$f_N = 0 = D \sin N\theta,$$

showing that

$$N\theta = s\pi, \quad s \text{ an integer.}$$

Therefore

$$f_i = D \sin \frac{is\pi}{N}. \tag{3.74}$$

By eqn (3.73)

$$c = 1-2r(1-\cos\theta) = 1-4r\sin^2\frac{s\pi}{2N}.$$

Hence eqns (3.67), (3.71), and (3.74) give that

$$u_{i,j} = E\left(1-4r\sin^2\frac{s\pi}{2N}\right)^j \sin\frac{s\pi i}{N},$$

where E replaces AD and s is an integer. But eqn (3.66) is linear in $u_{i,j}$ so the sum of different solutions is a solution. It follows therefore that a more general solution that will satisfy fairly general initial conditions is

$$u_{i,j} = \sum_{s=1}^{\infty} E_s\left(1-4r\sin^2\frac{s\pi}{2N}\right)^j \sin\frac{s\pi i}{N}. \tag{3.75}$$

Case (*i*) If the initial function $u_{i,0}$ is known only at the $(N+1)$ mesh points $(i, 0)$, $i = 0(1)N$, only the first $(N+1)$ values of E_s can be found by solving the $(N+1)$ linear equations

$$u_{i,0} = \sum_{s=1}^{N+1} E_s \sin\frac{s\pi i}{N}, \quad i = 0(1)N.$$

Case (*ii*) If $u_{i,0} = \phi(x)$, say, is a continuous function of x in $0 < x < 1$, it follows from eqn (3.75) that

$$\phi(x) = \sum_{s=1}^{\infty} E_s \sin\frac{s\pi i}{N} = \sum_{s=1}^{\infty} E_s \sin\frac{s\pi ih}{Nh}$$
$$= \sum_{s=1}^{\infty} E_s \sin s\pi x.$$

In this case E_s are the coefficients in the Fourier sine series for $\phi(x)$ and their values will be given by

$$E_s = 2\int_0^1 \phi(x)\sin s\pi x\,dx, \quad s = 1, 2, 3, \ldots.$$

Exercises and solutions

1. (a) Define the (S, T) Padé approximant. Calculate the $(2, 0)$ Padé approximant to $\exp\theta$ and show that its leading error term is $\theta^3/6$.

(b) The equation $\partial U/\partial t = \partial^2 U/\partial x^2$, $0<x<1$, $t>0$, for which U is zero at $x=0$ and 1, $t>0$, is approximated at the mesh points $x_i = ih$, $i = 1(1)N-1$, $Nh = 1$, along time-level $t_j = jk$ by the set of difference equations

$$\mathbf{u}_{j+1} = R_{2,0}(r\mathbf{T}_{N-1})\mathbf{u}_j,$$

where $R_{2,0}(\theta)$ is the $(2,0)$ Padé approximant to $\exp\theta$, $r = k/h^2$ and $\mathbf{T}_{N-1}$ is the matrix of order $(N-1)$ defined on p. 55. Show:

(i) That this difference scheme is L_0-stable.

(ii) That its extrapolated form is L_0-stable.

Could discontinuities between boundary and initial values possibly induce finite oscillations in the numerical solution?

Solution

(a) $\exp\theta = \{1/(1+q_1\theta+q_2\theta^2)\} + c_3\theta^3 + \dots$. The Maclaurin expansion of $\exp\theta$ leads to $(1+q_1)\theta + (q_1+q_2+\frac{1}{2})\theta^2 + (\frac{1}{2}q_1+q_2-c_3)\theta^3 + \dots \equiv 0$ for all θ. Hence $q_1 = -1$, $q_2 = \frac{1}{2}$, $c_3 = \frac{1}{6}$.

(b) $R_{2,0}(-z) = 1/(1+z+\frac{1}{2}z^2)$, $z>0$. Clearly, $|R_{2,0}(-z)|<1$ for all $z>0$ and $R_{2,0}(-z)$ tends to zero *monotonically* as $z\to\infty$ and discontinuities cannot induce finite oscillations.

(ii) $\mathbf{u}^{(E)} = \alpha\mathbf{u}^{(2)} - (\alpha-1)\mathbf{u}^{(1)}$, where $\alpha = 2^{S+T}/(2^{S+T}-1) = \frac{4}{3}$.

$$\begin{aligned} S_{2,0}(-z) &= \tfrac{4}{3}[R_{2,0}(-z)]^2 - \tfrac{1}{3}R_{2,0}(-2z) \\ &= 4/3(1+z+\tfrac{1}{2}z^2)^2 - 1/3(1+2z+2z^2) \\ &= \frac{1+2z+2z^2-\frac{1}{3}z^3-\frac{1}{12}z^4}{1+4z+8z^2+9z^3+\frac{25}{4}z^4+\frac{5}{2}z^5+\frac{1}{2}z^6}. \end{aligned}$$

Therefore, $|S_{2,0}(-z)|<1$ for all $z>0$ and $S_{2,0}(-z)\to 0$ as $z\to\infty$. The graph of $S_{2,0}(-z)$ is as in Fig. 3.6 and shows that $S_{2,0}$

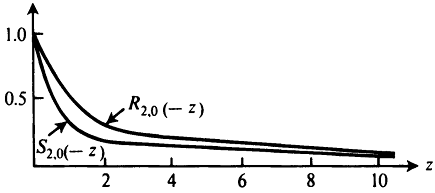


Fig. 3.6

decreases monotonically to zero. Hence the scheme is suitable for problems with discontinuities.

2. Show that the extrapolated form of the classical explicit-difference approximation to $\partial U/\partial t = \partial^2 U/\partial x^2$ is stable for $0 < r \leqslant \frac{1}{4}$, $r = k/h^2$.

Solution

$\mathbf{u}_{j+1} = R_{0,1}(k\mathbf{A})\mathbf{u}_j = (I + k\mathbf{A})\mathbf{u}_j$. $\mathbf{u}^{(1)}(t_j + 2k) = (\mathbf{I} + 2k\mathbf{A})\mathbf{u}_j$ and $\mathbf{u}^{(2)}(t_j + 2k) = (\mathbf{I} + k\mathbf{A})^2\mathbf{u}_j = (\mathbf{I} + 2k\mathbf{A} + k^2\mathbf{A}^2)\mathbf{u}_j$. $\mathbf{V}(t_j + 2k) = \{\exp(2k\mathbf{A})\}\mathbf{V}_j = (\mathbf{I} + 2k\mathbf{A} + 2k^2\mathbf{A}^2)\mathbf{V}_j + O(k^3)$. Hence,

$$\mathbf{u}^{(E)}(t_j + 2k) = 2\mathbf{u}^{(2)} - \mathbf{u}^{(1)} = (\mathbf{I} + 2k\mathbf{A} + 2k^2\mathbf{A}^2)\mathbf{u}_j = S_{0,1}(k\mathbf{A})\mathbf{u}_j.$$

Therefore, $S_{0,1}(-z) = 1 - 2z + 2z^2$, where $z = 4r\sin^2 s\pi/2N$, $s = 1(1)N-1$. For stability, $-1 < 1 - 2z + 2z^2 < 1$. The lower inequality implies that $0 < (z - \frac{1}{2})^2 + \frac{3}{4}$, which is satisfied by all real z. The upper inequality gives $z(z-1) < 0$. Hence, $0 < z < 1$, i.e. $0 < r < 1/4\sin^2 s\pi/2N > 1/4$, which is satisfied by $0 < r \leqslant 1/4$.

3. The function U satisfies the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < X, \quad t > 0,$$

and the boundary conditions $U(0, t) = U(X, t) = 0$, $t > 0$. If the space derivative is replaced by

$$\frac{1}{h^2}\{U(x-h, t) - 2U(x, t) + U(x+h, t)\} + O(h^2),$$

where $Nh = X$, show that the partial differential equation, when applied at each of the $(N-1)$ internal mesh points (ih, t), $i = 1(1)N-1$, can be approximated by a system of first-order ordinary differential equations $\mathrm{d}\mathbf{V}(t)/\mathrm{d}t = \mathbf{A}\mathbf{V}(t)$, and define vector $\mathbf{V}(t)$ and matrix $\mathbf{A}$.

Show that the solution of this system satisfies the time-step relation

$$\mathbf{V}(t+k) = \{\exp(k\mathbf{A})\}\mathbf{V}(t).$$

This relation is approximated by

$$\mathbf{u}(t+k) = \{\mathbf{I} - k(1-\theta)\mathbf{A}\}^{-1}(\mathbf{I} + k\theta\mathbf{A})\mathbf{u}(t) = L_k\mathbf{u}(t),$$

where $\mathbf{I}$ is the unit matrix of order $(N-1)$ and θ a scalar parameter. If the vectors $\mathbf{u}^{(1)}(t+2k)$ and $\mathbf{u}^{(2)}(t+2k)$ are defined by the computational procedures

$$\mathbf{u}^{(1)}(t+2k)=\mathbf{L}_{2k}\mathbf{u}(t) \quad \text{and} \quad \mathbf{u}^{(2)}(t+2k)=\mathbf{L}_k^2\mathbf{u}(t),$$

show that the vector of extrapolated values $\mathbf{u}^{(E)}(t+2k)$, defined by

$$\mathbf{u}^{(E)}(t+2k)=\alpha\mathbf{u}^{(2)}+(1-\alpha)\mathbf{u}^{(1)},$$

where α is a scalar parameter, is second-order accurate in t if $\alpha=2$ for all θ, or $\theta=\frac{1}{2}$ for all α.

Show that the method is L_0-stable for $\alpha=2$, $\theta=-1$.

Solution

$$\begin{aligned}\mathbf{u}^{(1)}(t+2k)&=\{\mathbf{I}-2k(1-\theta)\mathbf{A}\}^{-1}(\mathbf{I}+2k\theta\mathbf{A})\mathbf{u}(t)\\&=\{\mathbf{I}+2k\mathbf{A}+4(1-\theta)k^2\mathbf{A}^2+8(1-\theta)^2k^3\mathbf{A}^3+\ldots\}\mathbf{u}(t).\end{aligned}$$

$$\begin{aligned}\mathbf{u}^{(2)}(t+2k)&=\{\mathbf{I}-k(1-\theta)\mathbf{A}\}^{-2}(\mathbf{I}+k\theta\mathbf{A})^2\mathbf{u}(t)\\&=\{\mathbf{I}+2k\mathbf{A}+(3-2\theta)k^2\mathbf{A}^2+2(1-\theta)(2-\theta)k^3\mathbf{A}^3+\ldots\}\mathbf{u}(t).\end{aligned}$$

$$\begin{aligned}\mathbf{u}^{(E)}(t+2k)=[\mathbf{I}+2k\mathbf{A}+\{(3-2\theta)\alpha+4(1-\alpha)(1-\theta)k^2\mathbf{A}^2\\+\{2\alpha(1-\theta)(2-\theta)+8(1-\alpha)(1-\theta)^2\}k^3\mathbf{A}^3+\ldots]\mathbf{u}(t).\end{aligned}$$

$$\mathbf{V}(t+2k)=\exp(2k\mathbf{A})\mathbf{V}(t)=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2+\tfrac{4}{3}k^3\mathbf{A}^3+\ldots)\mathbf{V}(t).$$

$\mathbf{u}^{(E)}(t+2k)$ and $\mathbf{V}(t+2k)$ will be equal to terms in k^2 if $(3-2\theta)\alpha+4(1-\alpha)(1-\theta)=2$, i.e. $(\alpha-2)(2\theta-1)=0$. Hence the result.

Putting $\mathbf{u}^{(E)}(t+2k)=S(k\mathbf{A})\mathbf{u}(t)$, it follows that

$$\begin{aligned}S(k\mathbf{A})=\alpha\{\mathbf{I}-(1-\theta)k\mathbf{A}\}^{-2}(\mathbf{I}+\theta k\mathbf{A})^2\\+(1-\alpha)\{\mathbf{I}-2(1-\theta)k\mathbf{A}\}^{-1}(\mathbf{I}+2\theta k\mathbf{A}).\end{aligned}$$

Hence

$$\begin{aligned}S(-z)=\alpha(1-\theta z)^2/\{1+(1-\theta)z\}^2\\+(1-\alpha)(1-2\theta z)/\{1+2(1-\theta)z\}.\end{aligned}$$

For $\alpha=2$, $\theta=-1$,

$$\begin{aligned}S(-z)&=(1+6z+6z^2)/(1+8z+20z^2+16z^3)\\&=1-[(2z+14z^2+16z^3)/(1+8z+20z^2+16z^3)]\end{aligned}$$

Hence $0<S(-z)<1$, $z>0$, and $S(-z)\rightarrow 0$ as $z\rightarrow\infty$.

Comments

Gourlay and Morris, reference 10, give numerical results for various values of θ and α and also develop very accurate schemes of third- and fourth-order accuracy in t.

4. The equation $\partial U/\partial t-\partial^2U/\partial x^2=0$, $0<x<1$, $t>0$, for which U is zero at $x=0$ and 1, $t>0$, is approximated at the mesh points $x_i=ih$, $i=1(1)N-1$, along time-level $t_j=jk$ by the set of difference equations

$$\frac{1}{k}\{\mathbf{u}_{j+1}-R_{1,0}(r\mathbf{T}_{N-1})\mathbf{u}_j\}=0,$$

where $R_{1,0}(\theta)$ is the $(1,0)$ Padé approximant to $\exp\theta$, $r=k/h^2$ and matrix $\mathbf{T}_{N-1}$ is as previously defined.

Show that the principal part of the local truncation error at the point (ih, jk) is

$$\left[-\tfrac{1}{12}h^2\frac{\partial^4U}{\partial x^4}-\tfrac{1}{2}k\frac{\partial^2U}{\partial t^2}\right]_{i,j}, \quad i=2(1)N-2.$$

Solution

$$T_{i,j}=i\text{th row of }\frac{1}{k}\{(\mathbf{I}-r\mathbf{T}_{N-1})\mathbf{U}_{j+1}-\mathbf{U}_j\}$$

$$=\frac{1}{k}\{-rU_{i-1,j+1}+(1+2r)U_{i,j+1}-rU_{i+1,j+1}-U_{i,j}\},$$

where

$$U_{i-1,j+1}=\left[U+\left(-h\frac{\partial}{\partial x}+k\frac{\partial}{\partial t}\right)U+\frac{1}{2}\left(-h\frac{\partial}{\partial x}+k\frac{\partial}{\partial t}\right)^2U+\ldots\right]_{i,j},$$

$$U_{i,j+1}=\left[U+k\frac{\partial U}{\partial t}+\tfrac{1}{2}k^2\frac{\partial^2U}{\partial t^2}+\ldots\right]_{i,j}$$

$$U_{i+1,j+1}=\left[U+\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial t}\right)U+\frac{1}{2}\left(h\frac{\partial}{\partial x}+k\frac{\partial}{\partial t}\right)^2U+\ldots\right]_{i,j}.$$

First- and second-order derivative terms $=\left[\frac{\partial U}{\partial t}-\frac{\partial^2U}{\partial x^2}+\tfrac{1}{2}k\frac{\partial^2U}{\partial t^2}\right]_{i,j}.$

$$\text{Third-order terms}=\left[-k\frac{\partial^3U}{\partial x^2\,\partial t}+\tfrac{1}{6}k^2\frac{\partial^3U}{\partial t^3}\right]_{i,j}.$$

$$\text{Fourth-order terms} = \left[-\tfrac{1}{12}h^2\frac{\partial^4 U}{\partial x^4}\right]_{i,j}.$$

Hence the result since $\dfrac{\partial}{\partial t} \equiv \dfrac{\partial^2}{\partial x^2}$ so $\dfrac{\partial^3 U}{\partial x^2\,\partial t} = \dfrac{\partial^2 U}{\partial t^2}$.

5. As in Exercise 3, show that the solution $\mathbf{V}(t)$ of the system of ordinary differential equations approximating the solution of $\partial U/\partial t = \partial^2 U/\partial x^2$, $0<x<X$, $t>0$, $U(0,t)=U(X,t)=0$, at the mesh points $x_i = ih$, $i=1(1)N-1$, satisfies the recurrence relation

$$\mathbf{V}(t+k) = \{\exp(k\mathbf{A})\}\mathbf{V}(t).$$

If the exponential of $k\mathbf{A}$ is approximated by its $(2,0)$ Padé approximant, show that the vector $\mathbf{u}(t)$ approximating $\mathbf{V}$ is the solution of the difference equations

$$\mathbf{L}_k\mathbf{u}(t_j+k) = \mathbf{u}(t_j),$$

where

$$\mathbf{L}_k = \mathbf{I}_{N-1} - k\mathbf{A} + \tfrac{1}{2}k^2\mathbf{A}^2 = \begin{bmatrix} g & d & a & & & & \\ d & c & d & a & & & \\ a & d & c & d & a & & \\ & & \cdot & \cdot & \cdot & \cdot & \cdot \\ & & a & d & c & d & a \\ & & & a & d & c & d \\ & & & & a & d & g \end{bmatrix},$$

$g = 1+2r+\tfrac{5}{2}r^2$, $c = 1+2r+3r^2$, $d = -r-2r^2$, $a = \tfrac{1}{2}r^2$, $r = k/h^2$, $\mathbf{u}(t_j) = [u_{1,j}, u_{2,j}, \ldots, u_{N-1,j}]^T$ and $t_j = jk$.

Defining $\mathbf{u}^{(1)}(t+2k) = \mathbf{L}_k^{-2}\mathbf{u}(t)$ and $\mathbf{u}^{(2)}(t+2k) = \mathbf{L}_{2k}^{-1}\mathbf{u}(t)$, show that both $\mathbf{u}^{(1)}$ and $\mathbf{u}^{(2)}$ are second-order accurate in t, i.e. with a leading error term $O(k^3)$ in t.

Find the value of α if

$$\mathbf{u}^{(E)}(t+2k) = \alpha\mathbf{u}^{(1)} - (\alpha-1)\mathbf{u}^{(2)}$$

is third-order accurate in t.

Comments

In Exercise 1, the extrapolated scheme is shown to be L_0-stable and to exclude spurious oscillations. Twizell and Khaliq, reference 28, show that the method is highly accurate, with maximum errors for $U(x, 0)=1$, $X=2$ and $t=1.2$, of 0.74×10^{-4}, $(r=10)$, 0.41×10^{-3}, $(r=40)$, and 0.36×10^{-3}, $(r=160)$.

Solution

$R_{2,0}(k\mathbf{A})=(\mathbf{I}-k\mathbf{A}+\frac{1}{2}k^2\mathbf{A}^2)^{-1}=\mathbf{L}_k^{-1}$. Matrix $\mathbf{A}^2$ is given in the section on extrapolation, so $\mathbf{L}_k$ is easily written down.

$$\begin{aligned}\mathbf{u}^{(1)}(t+2k)&=(\mathbf{I}-k\mathbf{A}+\tfrac{1}{2}k^2\mathbf{A}^2)^{-2}\mathbf{u}(t)\\&=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2+k^3\mathbf{A}^3-\tfrac{1}{4}k^4\mathbf{A}^4)\mathbf{u}(t)+O(k^5).\\\mathbf{u}^{(2)}(t+2k)&=(\mathbf{I}-2k\mathbf{A}+2k^2\mathbf{A}^2)^{-1}\mathbf{u}(t)\\&=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2-4k^4\mathbf{A}^4)\mathbf{u}(t)+O(k^5).\\\mathbf{u}(t+2k)&=(\exp 2k\mathbf{A})\mathbf{u}(t)\\&=(\mathbf{I}+2k\mathbf{A}+2k^2\mathbf{A}^2+\tfrac{4}{3}k^3\mathbf{A}^3+\tfrac{2}{3}k^4\mathbf{A}^4)\mathbf{u}(t)+O(k^5).\end{aligned}$$

$$\alpha=\tfrac{4}{3},\text{ so }\mathbf{u}^{(E)}(t+2k)=\tfrac{4}{3}\mathbf{u}^{(1)}-\tfrac{1}{3}\mathbf{u}^{(2)}+O(k^4)+O(h^2).$$

6. The Douglas equations approximating $\partial U/\partial t=\partial^2 U/\partial x^2$, $0<x<1$, $t>0$, are given by

$$\begin{aligned}&(1-6r)u_{i-1,j+1}+(10+12r)u_{i,j+1}+(1-6r)u_{i+1,j+1}\\&\quad=(1+6r)u_{i-1,j}+(10-12r)u_{i,j}+(1+6r)u_{i+1,j},\\&\qquad\qquad i=1(1)N-1,\quad Nh=1.\end{aligned}$$

Given that the boundary values are known for all j, prove that the equations are unconditionally stable in the Lax–Richtmyer sense for all $r=k/h^2>0$. (See Exercise 8(c)).

Solution

The equations in matrix form are

$$\{(10+12r)\mathbf{I}+(1-6r)\mathbf{A}\}\mathbf{u}_{j+1}=\{(10-12r)\mathbf{I}+(1+6r)\mathbf{A}\}\mathbf{u}_j+\mathbf{b}_j,$$

where $\mathbf{b}_j$ is known, and matrix $\mathbf{A}$ is

$$\begin{bmatrix}0&1&&&\\1&0&1&&\\&1&0&1&\\&&&\ddots&\\&&&1&0\end{bmatrix}$$

of order $N-1$. Since the coefficient matrices of $\mathbf{u}_{j+1}$ and $\mathbf{u}_j$ are symmetric and commute, it follows that

$$\mathbf{B}=\{(10+12r)\mathbf{I}+(1-6r)\mathbf{A}\}^{-1}\{(10-12r)\mathbf{I}+(1+6r)\mathbf{A}\}$$

is real and symmetric. Hence, $\|\mathbf{B}\|_2=\rho(\mathbf{B})$. If λ is an eigenvalue of $\mathbf{A}$, the equations will be stable when

$$-1\leqslant\frac{(10-12r)+(1+6r)\lambda}{(10+12r)+(1-6r)\lambda}\leqslant 1,$$

for all λ, i.e.,

$$\begin{aligned}-(10+12r)-(1-6r)\lambda\leqslant(10-12r)+(1+6r)\lambda\\ \leqslant(10+12r)+(1-6r)\lambda,\end{aligned}$$

provided $(10+12r)+(1-6r)\lambda>0$. Hence $-10\leqslant\lambda\leqslant 2$. The eigenvalues of $\mathbf{A}$ are $\lambda_s=2\cos s\pi/N$. This value satisfies all the necessary inequalities for $r>0$.

7. Show that $U_{i,j+1}=\{\exp(k(\partial/\partial t))\}U_{i,j}$, where $U_{i,j}=U(x_i,t_j)$, $x_i=ih$, $i=0,\pm1,\pm2,\ldots$, and $t_j=jk$, $j=0,1,2,\ldots$. Using the result

$$D_x\equiv\frac{2}{h}\sinh^{-1}(\tfrac{1}{2}\delta_x)=\frac{1}{h}(\delta_x-\tfrac{1}{24}\delta_x^3+\tfrac{3}{640}\delta_x^5+\ldots),$$

where $D_x\equiv(\partial/\partial x)$ and $\delta_x U_{i,j}=U_{i+\frac{1}{2},j}-U_{i-\frac{1}{2},j}$, show that the *exact* difference replacement of the equation $(\partial U/\partial t)=(\partial^2U/\partial x^2)$ is given by

$$U_{i,j+1}=\{1+r\,\delta_x^2+\tfrac{1}{2}r(r-\tfrac{1}{6})\,\delta_x^4+\tfrac{1}{6}r(r^2-\tfrac{1}{2}r+\tfrac{1}{15})\,\delta_x^6+\ldots\}U_{i,j},$$

where $r=k/h^2$.

If only second-order central-differences are retained show that the function $u_{i,j}$ approximating $U_{i,j}$ is the solution of the classical explicit equation

$$u_{i,j+1}=ru_{i-1,j}+(1-2r)u_{i,j}+ru_{i+1,j}.$$

If only second- and fourth-order differences are retained show that $u_{i,j}$ is the solution of the explicit equation

$$\begin{aligned}u_{i,j+1}=\tfrac{1}{2}(2-5r+6r^2)u_{i,j}+\tfrac{2}{3}r(2-3r)(u_{i+1,j}+u_{i-1,j})\\ -\tfrac{1}{12}r(1-6r)(u_{i+2,j}+u_{i-2,j}).\end{aligned}$$

(This equation is stable for $r\leqslant\frac{2}{3}$.)

Solution

By Taylor's series,

$$U_{i,j+1}=U(x_i,\, t_j+k)=U_{i,j}+k\frac{\partial U_{i,j}}{\partial t}+\tfrac{1}{2}k^2\frac{\partial^2 U_{i,j}}{\partial x^2}+\dots$$

$$=\left(1+k\frac{\partial}{\partial t}+\frac{1}{2!}k^2\frac{\partial^2}{\partial t^2}+\frac{1}{3!}k^3\frac{\partial^3}{\partial t^3}+\dots\right)U_{i,j}=\left\{\exp\left(k\frac{\partial}{\partial t}\right)\right\}U_{i,j}.$$

For the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad \frac{\partial}{\partial t}\equiv\frac{\partial^2}{\partial x^2}.$$

Hence

$$U_{i,j+1}=\left\{\exp\left(k\frac{\partial^2}{\partial x^2}\right)\right\}U_{i,j}=\{\exp(kD_x^2)\}U_{i,j}$$

$$=(1+kD_x^2+\tfrac{1}{2}k^2D_x^4+\tfrac{1}{6}k^3D_x^6+\dots)U_{i,j}$$

where

$$D_x\equiv\frac{1}{h}(\delta_x-\tfrac{1}{24}\delta_x^3+\tfrac{3}{640}\delta_x^5+\dots)$$

Evaluate D_x^2, D_x^4, and D_x^6 to terms in δ_x^6 and substitute into the expression for $u_{i,j+1}$ etc. The results stated follow from

$$\delta_x^2 u_{i,j}=u_{i-1,j}-2u_{i,j}+u_{i+1,j}$$

and

$$\delta_x^4 u_{i,j}=\delta_x^2(u_{i-1,j}-2u_{i,j}+u_{i+1,j})$$

$$=u_{i-2,j}-4u_{i-1,j}+6u_{i,j}-4u_{i+1,j}+u_{i+2,j}.$$

8. (a) The equation

$$\frac{\partial U}{\partial t}=\frac{\partial}{\partial x}\left\{a(x)\frac{\partial U}{\partial x}\right\},\quad a(x)>0,$$

is approximated at the point (ih, jk) by the difference equation

$$\frac{1}{k}\Delta_t u_{i,j}=\frac{1}{h}\delta_x\left(a_i\frac{1}{h}\delta_x u_{i,j}\right).$$

Show that this gives the explicit difference equation

$$u_{i,j+1}=ra_{i-\frac{1}{2}}u_{i-1,j}+\{1-r(a_{i-\frac{1}{2}}+a_{i+\frac{1}{2}})\}u_{i,j}+ra_{i+\frac{1}{2}}u_{i+1,j},$$

where $r=k/h^2$.

(b) Using the result

$$U(x, t+k) = \left\{\exp\left(k\frac{\partial}{\partial t}\right)\right\}U(x, t),$$

show that the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}$$

may be replaced exactly by the equation

$$\exp(-\tfrac{1}{2}kD_x^2)\,U_{i,j+1} = \exp(\tfrac{1}{2}kD_x^2)\,U_{i,j},$$

where

$$D_x^2 U_{i,j} = \frac{\partial^2 U_{i,j}}{\partial x^2} = \frac{1}{h^2}(\delta_x^2 - \tfrac{1}{12}\delta_x^4 + \tfrac{1}{90}\delta_x^6 + \ldots)U_{i,j}.$$

Hence show that to second-order differences the differential equation may be approximated by the (Crank–Nicolson) equation

$$-ru_{i-1,j+1} + (2+2r)u_{i,j+1} - ru_{i+1,j+1} = ru_{i-1,j} + (2-2r)u_{i,j} + ru_{i+1,j}.$$

(c) Eliminate the fourth-order term in

$$D_x^2 \equiv \frac{1}{h^2}(\delta_x^2 - \tfrac{1}{12}\delta_x^4 + \tfrac{1}{90}\delta_x^6 + \ldots).$$

Hence use the first part of (b) to show that $\partial U/\partial t = \partial^2 U/\partial x^2$ may be approximated by the Douglas difference equation

$$(1-6r)u_{i+1,j+1} + (10+12r)u_{i,j+1} + (1-6r)u_{i+1,j+1}$$
$$= (1+6r)u_{i-1,j} + (10-12r)u_{i,j} + (1+6r)u_{i+1,j},$$

where $r = k/h^2$. (As mentioned previously this uses the same six grid points as the Crank–Nicolson equation but gives a more accurate solution.)

Solution

(a) As $\delta_x u_{i,j} = u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j}$, the approximation leads to

$$\frac{1}{k}(u_{i,j+1} - u_{i,j}) = \frac{1}{h^2}\delta_x\{a_i(u_{i+\frac{1}{2},j} - u_{i-\frac{1}{2},j})\}$$

$$= \frac{1}{h^2}\{a_{i+\frac{1}{2}}(u_{i+1,j} - u_{i,j}) - a_{i-\frac{1}{2},j}(u_{i,j} - u_{i-1,j})\} \quad \text{etc.}$$

(b) $U_{i,j+1}=\{\exp(k(\partial/\partial t))\}U_{i,j}$. Operate on both sides with $\exp(-\frac{1}{2}k(\partial/\partial t))$. Hence $\{\exp(-\frac{1}{2}k(\partial/\partial t))\}U_{i,j+1}=\{\exp(\frac{1}{2}k(\partial/\partial t))\}U_{i,j}$. By the differential equation $(\partial/\partial t)\equiv D_x^2$, therefore $\{\exp(-\frac{1}{2}kD_x^2)\}U_{i,j+1}=\{\exp(\frac{1}{2}kD_x^2)\}U_{i,j}$. Substituting for D_x^2 in terms of δ_x^2 and expanding to terms in δ_x^2 shows that

$$(1-\tfrac{1}{2}r\delta_x^2)u_{i,j+1}=(1+\tfrac{1}{2}r\delta_x^2)u_{i,j}$$

approximates the partial differential equation to this order of accuracy. Hence the result.

(c) Operate on both sides of

$$D_x^2\equiv\frac{1}{h^2}(\delta_x^2-\tfrac{1}{12}\delta_x^4+\ldots)$$

with $(1+\frac{1}{12}\delta_x^2)$ to give that

$$(1+\tfrac{1}{12}\delta_x^2)D_x^2\equiv\frac{1}{h_2}\delta_x^2+O(\delta_x^6).$$

By part (b), $(1-\frac{1}{2}kD_x^2+\ldots)U_{i,j+1}=(1+\frac{1}{2}kD_x^2+\ldots)U_{i,j}$. Operate on both sides with $(1+\frac{1}{12}\delta_x^2)$ and use the previous identity. Neglecting terms of order δ^6 and above yields the approximation equation

$$(1+\tfrac{1}{12}\delta_x^2-\tfrac{1}{2}r\delta_x^2)u_{i,j+1}=(1+\tfrac{1}{12}\delta_x^2+\tfrac{1}{2}r\delta_x^2)u_{i,j}.$$

Hence the result.

9. The Crank–Nicolson method approximates the equation

$$\left(\frac{\partial U}{\partial t}\right)_{i,j-\frac{1}{2}}=\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j-\frac{1}{2}} \quad\text{by}\quad \left(\frac{\partial U}{\partial t}\right)_{i,j-\frac{1}{2}}=\frac{1}{2}\left\{\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j}+\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j-1}\right\}.$$

Assuming the following central-difference formulae for the derivatives f' and f'' for a mesh length h, namely,

$$hf'_{\frac{1}{2}}=(\delta-\tfrac{1}{24}\delta^3+\tfrac{3}{640}\delta^5-\ldots)f_{\frac{1}{2}},$$

and

$$h^2f''_0=(\delta^2-\tfrac{1}{12}\delta^4+\tfrac{1}{90}\delta^6-\ldots)f_0,$$

show that the Crank–Nicolson equation leads to

$$(u_{i,j}-u_{i,j-1})/k$$
$$=(u_{i+1,j}-2u_{i,j}+u_{i-1,j}+u_{i+1,j-1}-2u_{i,j-1}+u_{i-1,j-1})/2h^2+C,$$

where the correction term C is given by

$$C=\frac{1}{k}\left(\tfrac{1}{24}\delta^3 u_{i,j-\frac{1}{2}}-\tfrac{3}{640}\delta^5 u_{i,j-\frac{1}{2}}+\dots\right)_t+\frac{1}{2h^2}\{(-\tfrac{1}{12}\delta^4 u_{i,j}+\tfrac{1}{90}\delta^6 u_{i,j}-\dots)_x$$
$$+(-\tfrac{1}{12}\delta^4 u_{i,j-1}+\tfrac{1}{90}\delta^6 u_{i,j-1}-\dots)_x\}.$$

Comment

The correction term can be used to improve the accuracy of the finite-difference solution obtained initially by neglecting it. The function values of this first approximation are differenced in the t direction to give $\delta^3 u_{i,j-\frac{1}{2}}$ and in the x-direction to give

$$\delta^4 u_{i,j},\quad \delta^4 u_{i,j-1},\quad \delta^6 u_{i,j}\quad \text{and}\quad \delta^6 u_{i,j-1}.$$

The correction term for each equation is then calculated from these differences, which of course are numbers, and the corrected finite-difference equations re-solved for the $u_{i,j}$.

10. The function U satisfies the non-linear equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U^2}{\partial x^2},\quad 0<x<1,\quad t>0,$$

the initial condition $U=1$, $0\leqslant x\leqslant 1$, $t=0$, and the boundary conditions

$$\frac{\partial U}{\partial x}=0 \text{ at } x=0,\ t>0,\quad \frac{\partial U}{\partial x}=-U \text{ at } x=1,\ t>0.$$

If the equation is approximated by the difference equation

$$\frac{1}{k}\delta_t u_{i,j+\frac{1}{2}}=\frac{1}{2h^2}(\delta_x^2 u_{i,j+1}^2+\delta_x^2 u_{i,j}^2),$$

and the derivative boundary conditions are approximated by the usual central-difference formulae, show that the corresponding non-linear approximation equations are

$$u_0^2+pu_0-u_1^2+\{u_{0,j}^2-pu_{0,j}-u_{1,j}^2\}=0,$$

$$u_{i-1}^2-2(u_i^2+pu_i)+u_{i+1}^2+\{u_{i-1,j}^2-2(u_{i,j}^2-pu_{i,j})+u_{i+1,j}^2\}=0,$$

$$i=1(1)(N-1),$$

and

$$\begin{aligned} &u_{N-1}^2 - 2hu_{N-1}u_N - (1-2h^2)u_N^2 - pu_N \\ &\qquad + \{u_{N-1,j}^2 - 2hu_{N-1,j}u_{N,j} - (1-2h^2)u_{N,j}^2 + pu_{N,j}\} = 0, \end{aligned}$$

where $p = h^2/k$, $Nh = 1$ and u_i denotes the unknown $u_{i,j+1}$, $i = 0(1)N$.

Briefly describe Newton's method for deriving a set of linear equations giving an improved value $(V_i + \varepsilon_i)$ to the approximate solution values V_i, $i = 1(1)N$, of the N non-linear equations

$$f_i(u_1, u_2, \ldots, u_N) = 0, \quad i = 1(1)N.$$

Apply this method to the approximation equations previously obtained and write out the linear equations giving the first iteration values at mesh points along the time-level $t = k$, taking V_i, $i = 0(1)N$, equal to the initial value 1.

Solution

$$\begin{aligned} &\frac{1}{k}(u_{i,j+1} - u_{i,j}) \\ &\qquad = \frac{1}{2h^2}(u_{i-1,j+1}^2 - 2u_{i,j+1}^2 + u_{i+1,j+1}^2 + u_{i-1,j}^2 - 2u_{i,j}^2 + u_{i+1,j}^2). \end{aligned}$$

Put $p = h^2/k$ and denote $u_{i,j+1}$ by u_i. Then

$$u_{i-1}^2 - 2(u_i^2 + pu_i) + u_{i+1}^2 + \{u_{i-1,j}^2 - 2(u_{i,j}^2 - pu_{i,j}) + u_{i+1,j}^2\} = 0.$$

By the B.C.,

$$\left(\frac{\partial U}{\partial x}\right)_{0,j} \simeq \frac{u_{1,j} - u_{-1,j}}{2h} = 0 \Rightarrow u_{-1,j} = u_{1,j}.$$

Put $i = 0$ in the approximation equation and eliminate $u_{-1,j}$ to give

$$u_0^2 + pu_0 - u_1^2 + (u_{0,j}^2 - pu_{0,j} - u_{1,j}^2) = 0.$$

By the B.C. at

$$i = N, \quad \frac{1}{2h}(u_{N+1,j+1} - u_{N-1,j+1}) = -u_{N,j} \quad \text{etc.}$$

By Newton's method,

$$\left(\frac{\partial f_i}{\partial u_1}\varepsilon_1 + \ldots + \frac{\partial f_i}{\partial u_N}\varepsilon_N\right)_{u_i = V_i} + f_i(V_1, V_2, \ldots, V_N) = 0.$$

Hence

$$
\begin{aligned}
&(2V_0+p)\varepsilon_0-2V_1\varepsilon_1+\{V_0^2+pV_0-V_1^2+(u_{0,j}^2-pu_{0,j}-u_{1,j}^2)\}=0,\\
&2V_{i-1}\varepsilon_{i-1}-2(2V_i+p)\varepsilon_i+2V_{i+1}\varepsilon_{i+1}\\
&\qquad+\{V_{i-1}^2-2(V_i^2+pV_i)+V_{i-1}^2\\
&\qquad+(u_{i-1,j}^2-2[u_{i,j}^2-pu_{i,j}]+u_{i+1,j}^2)\}=0,
\end{aligned}
$$

and

$$
\begin{aligned}
&(2V_{N-1}-2hV_N)\varepsilon_{N-1}+\{-2hV_{N-1}-2(1-2h^2)V_N-p\}\varepsilon_N\\
&\quad+\{V_{N-1}^2-2hV_{N-1}V_N-(1-2h^2)V_N^2-pV_N\\
&\quad+u_{N-1,j}^2-2hu_{N-1,j}u_{N,j}-(1-2h^2)u_{N,j}^2+pu_{N,j}\}=0.
\end{aligned}
$$

Put $V_i=1$, $i=0(1)N$, $j=0$, and $u_{i,0}=1$, $i=0(1)N$ etc.

11. The equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad 0<x<1,\quad t>0,$$

is approximated at the point (ih, jk) by the difference equation

$$\frac{3}{2}\left(\frac{u_{i,j+1}-u_{i,j}}{k}\right)-\frac{1}{2}\left(\frac{u_{i,j}-u_{i,j-1}}{k}\right)=\frac{1}{h^2}\,\delta_x^2 u_{i,j+1},$$

where $x=ih$, $t=jk$, and $Nh=1$. Investigate the stability of this system of equations by the matrix method, for known boundary values and a fixed mesh size.

Solution

Equations are $-2ru_{i-1,j+1}+(3+4r)u_{i,j+1}-2ru_{i+1,j+1}=4u_{i,j}-u_{i,j-1}$, $i=1(1)(N-1)$. For known boundary values they can be written as

$$\mathbf{u}_{j+1}=4\mathbf{A}^{-1}\mathbf{u}_j-\mathbf{A}^{-1}\mathbf{u}_{j-1}+\mathbf{A}^{-1}\mathbf{c}_{j+1},$$

where $\mathbf{c}_{j+1}$ is a vector of known boundary values and

$$\mathbf{A}=\begin{bmatrix}(3+4r) & -2r & & \\ -2r & (3+4r) & -2r & \\ & & & \\ & & -2r & (3+4r)\end{bmatrix}$$

is of order $(N-1)$. This equation and $\mathbf{u}_j = \mathbf{u}_j$ can be written as

$$\left[\begin{array}{c}\mathbf{u}_{j+1}\\ \hline \mathbf{u}_j\end{array}\right]=\left[\begin{array}{c|c}4\mathbf{A}^{-1} & -\mathbf{A}^{-1}\\ \hline \mathbf{I} & \mathbf{O}\end{array}\right]\left[\begin{array}{c}\mathbf{u}_j\\ \hline \mathbf{u}_{j-1}\end{array}\right]+\left[\begin{array}{c}\mathbf{A}^{-1}\mathbf{c}\\ \hline \mathbf{O}\end{array}\right],$$

i.e. $\mathbf{v}_{j+1}=\mathbf{P}\mathbf{v}_j+\mathbf{c}'$. $\mathbf{A}$ has distinct eigenvalues and all the submatrices of $\mathbf{P}$ commute with each other so the eigenvalues λ of $\mathbf{P}$ are the eigenvalues of

$$\begin{bmatrix}\dfrac{4}{\mu_k} & -\dfrac{1}{\mu_k}\\ 1 & 0\end{bmatrix}$$

where μ_k is the kth eigenvalue of $\mathbf{A}$. As $\det(\mathbf{P}-\lambda\mathbf{I})=0=\lambda^2-(4/\mu_k)\lambda+(1/\mu_k)$, $\lambda=\{2\pm\sqrt{(4-\mu_k)}\}/\mu_k$ where

$$\mu_k=3+8r\sin^2(k\pi/2N),\quad k=1(1)(N-1)$$

by p. 59. Hence

$$\lambda=\left\{2\pm\left(1-8r\sin^2\frac{k\pi}{2N}\right)^{\frac{1}{2}}\right\}\bigg/\left(3+8r\sin^2\frac{k\pi}{2N}\right).$$

When the roots are real, $|\lambda|<(2+1)/(3+\delta)$, $\delta>0$. Hence $|\lambda|<1$. When the roots are complex, $|\lambda|=\mu_k^{-\frac{1}{2}}<1$. Therefore the equations are unconditionally stable.

12. Show that the components of the eigenvector $\mathbf{v}$, corresponding to the eigenvalue λ of the matrix $\mathbf{A}$ of order N defined by

$$\mathbf{A}=\begin{bmatrix}-2 & 2 & & & & & \\ 1 & -2 & 1 & & & & \\ & 1 & -2 & 1 & & & \\ & & & \cdot & & & \\ & & & & \cdot & & \\ & & & & 1 & -2 & 1\\ & & & & & 1 & -2\end{bmatrix}$$

are given by the solution of the difference equations

$$v_{j-1}-(2+\lambda)v_j+v_{j+1}=0,\quad j=2(1)N,$$

satisfying the conditions

$$-(2+\lambda)v_1+2v_2=0=v_{N+1}.$$

Hence prove that the jth component of $\mathbf{v}$ can be expressed as

$$v_j=\mathrm{B}\cos j\theta+\mathrm{C}\sin j\theta,$$

where B and C are constants. Deduce that the rth eigenvalue of $\mathbf{A}$ is $-4\sin^2((2r-1)\pi/4N)$.

Solution

Expand $\mathbf{Av}=\lambda\mathbf{v}$. The solution of $v_{j-1}-(2+\lambda)v_j+v_{j+1}=0$ is $v_j=\mathrm{D}m_1^j+\mathrm{E}m_2^j$ where m_1, m_2 are the roots of $m^2-(2+\lambda)m+1=0$. Therefore $m_1m_2=1$, $m_1+m_2=2+\lambda$, λ real by part (a). Put $m_1=re^{i\theta}$. Then $r=1$, giving $\lambda=2(-1+\cos\theta)$, $v_j=\mathrm{B}\cos j\theta+\mathrm{C}\sin j\theta$. As $v_{N+1}=0$, $\mathrm{B}/\mathrm{C}=-\tan(N+1)\theta$. Substitute for v_1, v_2, and λ in terms of θ into $-(2+\lambda)v_1+2v_2=0$ to get $\mathrm{C}\sin\theta\cos N\theta=0$. Hence $\theta=(2r-1)\pi/2N$, r an integer.

13. Verify that $U=e^{-\pi^2t}\sin\pi x$ is a solution of the equation

$$\frac{\partial U}{\partial t}=\frac{\partial^2 U}{\partial x^2},\quad 0<x<1,\quad t>0,$$

which satisfies the boundary conditions $U=0$ at $x=0$ and 1, $t>0$, and the initial condition $U=\sin\pi x$ when $t=0$, $0\leqslant x\leqslant 1$. The differential equation is approximated at the point (ih, jk) by the explicit equation

$$\frac{1}{k}\Delta_t u_{i,j}=\frac{1}{h^2}\delta_x^2 u_{i,j}.$$

Show that the analytical solution of the difference equation satisfying the same boundary and initial values is

$$u_{i,j}=\left(1-4r\sin^2\frac{\pi h}{2}\right)^j\sin\pi x_i,$$

where $r=k/h^2$.

Given that $0<r\leqslant\frac{1}{2}$, deduce that $u_{i,j}$ converges to $U_{i,j}$ as h tends to zero, for finite values of t.

Solution

As in the text. Because of the initial function, only the first term of the series solution is needed.

$$|u_{i,j}-U_{i,j}|=\left|\left(1-4r\sin^2\frac{\pi h}{2}\right)^j-e^{-\pi^2kj}\right||\sin\pi x|=|a^j-b^j|\,|\sin\pi x|,$$

where

$$|a| = \left|1 - 4r\sin^2\frac{\pi h}{2}\right| \leqslant 1$$

since $r \leqslant \frac{1}{2}$ and $|b| = |e^{-\pi^2 k}| < 1$. Hence

$$|u_{i,j} - U_{i,j}| = |a - b|\,|a^{j-1} + a^{j-2}b + \ldots + b^{j-1}|\,|\sin \pi x| < j\,|a - b|,$$

since there are j terms in the second series, $= j\,|1 - 2r + 2r\cos\pi h - e^{-\pi^2 k}|$. Replace $\cos \pi h$ and $e^{-\pi^2 k}$ by their Maclaurin expansions to get that

$$|u_{i,j} - U_{i,j}| < t\pi^4 h^2 A,$$

where A is bounded because it cannot exceed $|a - b|$. Hence the result.

4 Hyperbolic equations

First-order quasi-linear equations and characteristics

Consider the equation

$$a\frac{\partial U}{\partial x}+b\frac{\partial U}{\partial y}=c,$$

where a, b, and c are, in general, functions of x, y, and U but not of $\partial U/\partial x$ and $\partial U/\partial y$, i.e. the first-order derivatives occur only to the first degree although the equation need not be linear in U. Such an equation is said to be quasi-linear. It is customary to put $\partial U/\partial x=p$ and $\partial U/\partial y=q$ and to write the equation as

$$ap+bq=c. \tag{4.1}$$

The following analysis shows that at each point of the solution domain of such an equation there is a direction along which the integration of eqn (4.1) transforms to the integration of an ordinary differential equation. In other words, *in this direction* the expression to be integrated will be independent of partial derivatives in other directions, such as p and q.

Assume we know the solution values U of eqn (4.1) at every point on a curve C in the x–y plane, where C does *not* coincide with the curve Γ on which initial values of U are specified. The question to be asked at this stage is this. Can we determine values for p and q on C from the values of U on C so that they satisfy eqn (4.1)?

If we can, then in directions tangential to C from points on C we shall automatically satisfy the differential relationship

$$\mathrm{d}U=\frac{\partial U}{\partial x}\mathrm{d}x+\frac{\partial U}{\partial y}\mathrm{d}y=p\,\mathrm{d}x+q\,\mathrm{d}y, \tag{4.2}$$

where $\mathrm{d}y/\mathrm{d}x$ is the slope of the tangent to C at $P(x, y)$ on C.

Equations (4.1) and (4.2) are two equations for p and q.

Elimination of p between them gives that

$$\mathrm{d}U=\frac{c-bq}{a}\,\mathrm{d}x+q\,\mathrm{d}y,$$

which can be written as

$$q(a\,\mathrm{d}y-b\,\mathrm{d}x)+(c\,\mathrm{d}x-a\,\mathrm{d}U)=0. \tag{4.3}$$

This equation is explicitly independent of p because the coefficients a, b, and c are functions of x, y, and U only. It can also be made independent of q by choosing the curve C so that its slope $\mathrm{d}y/\mathrm{d}x$ satisfies the equation

$$a\,\mathrm{d}y-b\,\mathrm{d}x=0. \tag{4.4}$$

By eqns (4.4) and (4.3) it then follows that

$$c\,\mathrm{d}x-a\,\mathrm{d}U=0. \tag{4.5}$$

Equation (4.4) is a differential equation for the curve C and (4.5) is a differential equation for the solution values of U along C. The curve C is called a characteristic curve or characteristic. These equations are easy to remember because they can be written as

$$\frac{\mathrm{d}x}{a}=\frac{\mathrm{d}y}{b}=\frac{\mathrm{d}U}{c}.$$

This also shows that U may be found from either the equation $\mathrm{d}U=(c/a)\,\mathrm{d}x$ or the equation $\mathrm{d}U=(c/b)\,\mathrm{d}y$.

Example 4.1

Consider the equation

$$y\frac{\partial U}{\partial x}+\frac{\partial U}{\partial y}=2, \tag{4.6}$$

where U is known along the initial segment Γ defined by $y=0$, $0\leqslant x\leqslant 1$.

The differential equation of the family of characteristic curves is

$$\frac{\mathrm{d}x}{y}=\frac{\mathrm{d}y}{1}.$$

Hence the equation of this family is $x=\frac{1}{2}y^2+A$, where the parameter A is a constant for each characteristic. For the characteristic through $R(x_R, 0)$, $A=x_R$, so the equation of this particular characteristic is $y^2=2(x-x_R)$.

The solution along a characteristic curve is given by

$$\frac{dy}{1}=\frac{dU}{2},$$

which integrates to $U=2y+B$, where B is constant along a particular characteristic. If $U=U_R$ at $R(x_R, 0)$ then $B=U_R$ and the solution along the characteristic $y^2=2(x-x_R)$ is $U=2y+U_R$.

Since initial values for U are known only on the line segment *OF*, Fig. 4.1, where $0\leqslant x_R\leqslant 1$, it follows that the solution is defined only in the region bounded by, and including, the terminal characteristics $y^2=2x$ and $y^2=2(x-1)$. In this region the solution is clearly unique. Outside this region the solution is undefined.

If the initial curve Γ coincides with a characteristic, say, for example, the characteristic $y^2=2x$ through $O(0,0)$, then the solution along this characteristic is $U=2y+U_0$, where U_0 is the specified initial value of U at O. In other words, the initial values

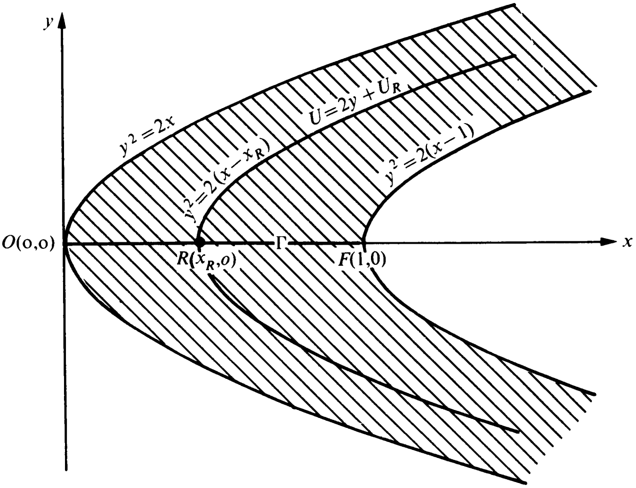


Fig. 4.1

for U on the initial curve $y^2=2x$ cannot now be arbitrarily prescribed as is obviously possible when Γ does not coincide with a characteristic curve. In this case it is also easily shown that the solution is not unique at points off $y^2=2x$. Consider, for example,

$$U=2y+U_0+A(y^2-2x),$$

where A is an arbitrary constant. This is clearly the solution along $y^2=2x$, whatever the value of A. It is also a solution of eqn (4.6) when $y^2\neq 2x$ as can easily be verified by direct differentiation. Since A is an arbitrary constant there is an infinite number of different solutions at points off the characteristic $y^2=2x$.

A method for numerical integration along a characteristic

Let U be specified on the initial curve Γ which must not be a characteristic curve.

Let $R(x_R, y_R)$ be a point on Γ and $P(x_P, y_P)$ be a point on the characteristic curve C through R such that x_P-x_R is small, Fig. 4.2. The differential equation for the characteristic is

$$a\,\mathrm{d}y=b\,\mathrm{d}x,$$

which gives either $\mathrm{d}y$ or $\mathrm{d}x$ when the other quantities are known.

The differential equation for the solution along a characteristic is either

$$a\,\mathrm{d}U=c\,\mathrm{d}x \quad \text{or} \quad b\,\mathrm{d}U=c\,\mathrm{d}y,$$

which gives $\mathrm{d}U$ for known $\mathrm{d}x$ or $\mathrm{d}y$ and known a, b, and c.

Denote a first approximation to U by $u^{(1)}$, a second approximation by $u^{(2)}$, etc.

First approximations

Assume that x_P is known. Then by the equations (4.4), (4.5)

$$a_R\{y_P^{(1)}-y_R\}=b_R(x_P-x_R)$$

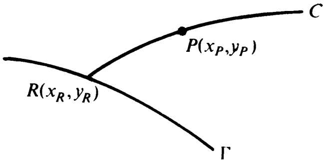


Fig. 4.2

gives a first approximation $y_P^{(1)}$ to y_P and

$$a_R\{u_P^{(1)} - u_R\} = c_R(x_P - x_R) \text{ gives } u_P^{(1)}.$$

Second and subsequent approximations

Replace the coefficients a, b, and c by known mean values over the arc RP. Then

$$\tfrac{1}{2}(a_R + a_P^{(1)})(y_P^{(2)} - y_R) = \tfrac{1}{2}(b_R + b_P^{(1)})(x_P - x_R) \text{ gives } y_P^{(2)},$$

and

$$\tfrac{1}{2}(a_R + a_P^{(1)})(u_P^{(2)} - u_R) = \tfrac{1}{2}(c_R + c_P^{(1)})(x_P - x_R) \text{ gives } u_P^{(2)}.$$

This second procedure can be repeated iteratively until successive iterates agree to a specified number of decimal places.

Example 4.2

The function U satisfies the equation

$$\sqrt{x}\frac{\partial U}{\partial x} + U\frac{\partial U}{\partial y} = -U^2$$

and the condition $U = 1$ on $y = 0$, $0 < x < \infty$.

Show that the Cartesian equation of the characteristic through the point $R(x_R, 0)$, $x_R > 0$, is $y = \log(2\sqrt{x} + 1 - 2\sqrt{x_R})$. Use a finite-difference method to calculate a first approximation to the solution and to the value of y at the point $P(1.1, y)$, $y > 0$, on the characteristic through the point $R(1, 0)$.

Calculate a second approximation to these values by an iterative method. Compare the results with those given by the analytical formulae for y and U. We have that

$$\frac{dx}{\sqrt{x}} = \frac{dy}{U} = \frac{dU}{-U^2}.$$

Hence $y = -\log AU$. As $U = 1$ at $(x_R, 0)$, $A = 1$, so

$$y = \log\frac{1}{U}. \tag{4.7}$$

Similarly, $2\sqrt{x} = 1/U + B$. As $U = 1$ at $(x_R, 0)$, $B = 2\sqrt{x_R} - 1$. Therefore,

$$\frac{1}{U} = 2\sqrt{x} + 1 - 2\sqrt{x_R}. \tag{4.8}$$

Elimination of the parameter U between eqns (4.7) and (4.8) shows that the Cartesian equation of the characteristic through $(x_R, 0)$ is

$$y = \log(2\sqrt{x} + 1 - 2\sqrt{x_R}). \tag{4.9}$$

The solution along the characteristic is given either by

$$U = e^{-y} \quad \text{or} \quad U = \frac{1}{(2\sqrt{x} + 1 - 2\sqrt{x_R})}. \tag{4.10}$$

First approximations at (1.1, y). (Fig. 4.2)

We have that $\sqrt{x}\,dy = U\,dx$ and $\sqrt{x}\,dU = -U^2\,dx$.
Hence

$$\sqrt{x_R}(y_P^{(1)} - 0) = U_R\,dx \text{ giving } y_P^{(1)} = \frac{1}{\sqrt{1}}(0.1) = 0.1,$$

and

$$\sqrt{x_R}(u_P^{(1)} - 1) = -U_R^2\,dx \text{ giving } u_P^{(1)} = 1 - 0.1 = 0.9.$$

Second approximations

Using average values for the coefficients,

$$\tfrac{1}{2}(u_R + u_P^{(1)})\,dx = \tfrac{1}{2}(\sqrt{x_R} + \sqrt{x_P})(y_P^{(2)} - y_R),$$

giving

$$(1 + 0.9)(0.1) = (1 + 1.0488)(y_P^{(2)} - 0),$$

from which $y_P^{(2)} = 0.0927$. Also

$$\tfrac{1}{2}(\sqrt{x_R} + \sqrt{x_P})(u_P^{(2)} - u_R) = -\tfrac{1}{2}\{u_R^2 + (u_P^{(1)})^2\}\,dx,$$

giving

$$(1 + 1.0488)(u_P^{(2)} - 1) = -(1 + 0.81)(0.1),$$

from which $U_P^{(2)} = 0.9117$.

Note: The differential equations could have been written as

$$dy = \frac{U}{\sqrt{x}}\,dx \quad \text{and} \quad dU = -\frac{U^2}{\sqrt{x}}\,dx$$

and the second approximations obtained from

$$y_P^{(2)} = \frac{1}{2}\left\{\left(\frac{u}{\sqrt{x}}\right)_R + \left(\frac{u}{\sqrt{x}}\right)_P\right\}dx$$

and

$$u_P^{(2)} - u_R = \frac{1}{2}\left\{-\left(\frac{u^2}{\sqrt{x}}\right)_R - \left(\frac{u^2}{\sqrt{x}}\right)_P\right\} \mathrm{d}x.$$

These approximations yield $y_P^{(2)} = 0.0929$ and $u_P^{(2)} = 0.9114$.

Analytical values

By eqn (4.9),

$$y_P = \log_e \{2(1.0488) + 1 - 2\} = 0.0934.$$

By eqn (4.10),

$$U_P = \frac{1}{1.0976} = 0.9111.$$

Finite-difference methods on a rectangular mesh for first-order equations

Lax–Wendroff explicit method

In the theory of fluid flow the equations of motion, of continuity, and of energy can be combined into one conservation equation of the form

$$\frac{\partial \mathbf{U}}{\partial t} + \frac{\partial \mathbf{F}(\mathbf{U})}{\partial x} = 0, \tag{4.11}$$

where **U** and **F** are each column vectors with three components. The Lax–Wendroff method, as illustrated below for a single dependent variable, can be used to approximate eqn (4.11) by an explicit difference equation of second-order accuracy.

Consider

$$\frac{\partial U}{\partial t} + a\frac{\partial U}{\partial x} = 0,$$

a a positive constant. By Taylor's expansion,

$$U_{i,j+1} = U(x_i, t_j + k) = U_{i,j} + k\left(\frac{\partial U}{\partial t}\right)_{i,j} + \tfrac{1}{2}k^2\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j} + \ldots,$$

where $x_i = ih$ and $t_j = jk$, $i = 0, \pm 1, \pm 2, \ldots$, $j = 0, 1, 2, \ldots$.

The differential equation can now be used to eliminate the

t-derivatives because it gives that

$$\frac{\partial}{\partial t} \equiv -a\frac{\partial}{\partial x},$$

so

$$U_{i,j+1} = U_{i,j} - ka\left(\frac{\partial U}{\partial x}\right)_{i,j} + \tfrac{1}{2}k^2a^2\left(\frac{\partial^2 U}{\partial x^2}\right)_{i,j} + \dots.$$

Finally, the replacement of the x-derivatives by central-difference approximations gives, to terms in k^2, the explicit difference equation

$$\begin{aligned} u_{i,j+1} &= u_{i,j} - \frac{ka}{2h}(u_{i+1,j} - u_{i-1,j}) + \frac{k^2a^2}{2h^2}(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}) \\ &= \tfrac{1}{2}ap(1+ap)u_{i-1,j} + (1-a^2p^2)u_{i,j} - \tfrac{1}{2}ap(1-ap)u_{i+1,j}, \end{aligned} \tag{4.12}$$

where $p = k/h$. This may be used for both initial-value and initial-value boundary-value problems. It is often used to obtain numerical solutions to differential equations in fluid-flow problems when the dependent variables change rapidly with time. In such problems k must be kept small so the advantages of implicit methods are lost, namely stable equations and accurate results for fairly large values of k. As shown in Exercise 3, eqn (4.12) is stable for $0 < ap \leqslant 1$ and its local truncation error is

$$\tfrac{1}{6}k^2\frac{\partial^3 U}{\partial t^3} + \tfrac{1}{6}ah^2\frac{\partial^3 U}{\partial x^3} + \dots.$$

Example 4.3

The function U satisfies the equation

$$\frac{\partial U}{\partial t} + \frac{\partial U}{\partial x} = 0, \quad 0 < x < \infty, \quad t > 0,$$

the boundary condition

$$U(0, t) = 2t, \quad t > 0,$$

and the initial conditions

$$\begin{aligned} U(x, 0) &= x(x-2), \quad 0 \leqslant x \leqslant 2, \\ U(x, 0) &= 2(x-2), \quad 2 \leqslant x. \end{aligned}$$

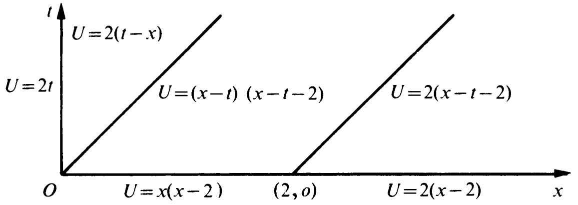


Fig. 4.3

Calculate (i) the analytical solution and (ii) a numerical solution using the explicit Lax–Wendroff equation.

(i) $\mathrm{d}t = \mathrm{d}x = \mathrm{d}U/0$. Hence U is constant along the straight line characteristics $t = x + \text{constant}$. Therefore, the equation of the characteristic from $R(x_R, 0)$ is $t = x - x_R$ and if $U(x, 0)$ is $\phi(x)$, say, then the solution along this characteristic is $U(x, t) = \phi(x_R) = \phi(x - t)$. Similarly, if $U(0, t) = \psi(t)$, the solution along the characteristic $t - t_S = x$ from $S(0, t_s)$ is $U(x, t) = \psi(t - x)$. Hence the conditions of the problem give the solution shown in Fig. 4.3.

(ii) Equation (4.12) for $h = \frac{1}{4}$ and $k = \frac{1}{8}$ is

$$u_{i,j+1} = \tfrac{3}{8}u_{i-1,j} + \tfrac{3}{4}u_{i,j} - \tfrac{1}{8}u_{i+1,j}.$$

Table 4.1 displays the analytical solution of the differential equation below each mesh point and the corresponding Lax–Wendroff solution to 3*D* above each mesh point for $x = 0(1)7$ and $t = 0(\frac{1}{2})3$. The boundary and initial conditions make $U(x, t)$ continuous at $(0, 0)$ and $(2, 0)$, $\partial U/\partial x$ continuous at $(2, 0)$, and ensures that $\partial U/\partial x$ and $\partial U/\partial t$ satisfy the differential equation at $(0, 0)$.

The Lax–Wendroff method for a set of simultaneous equations

The Lax–Wendroff approximation is easily extended to the set of simultaneous equations

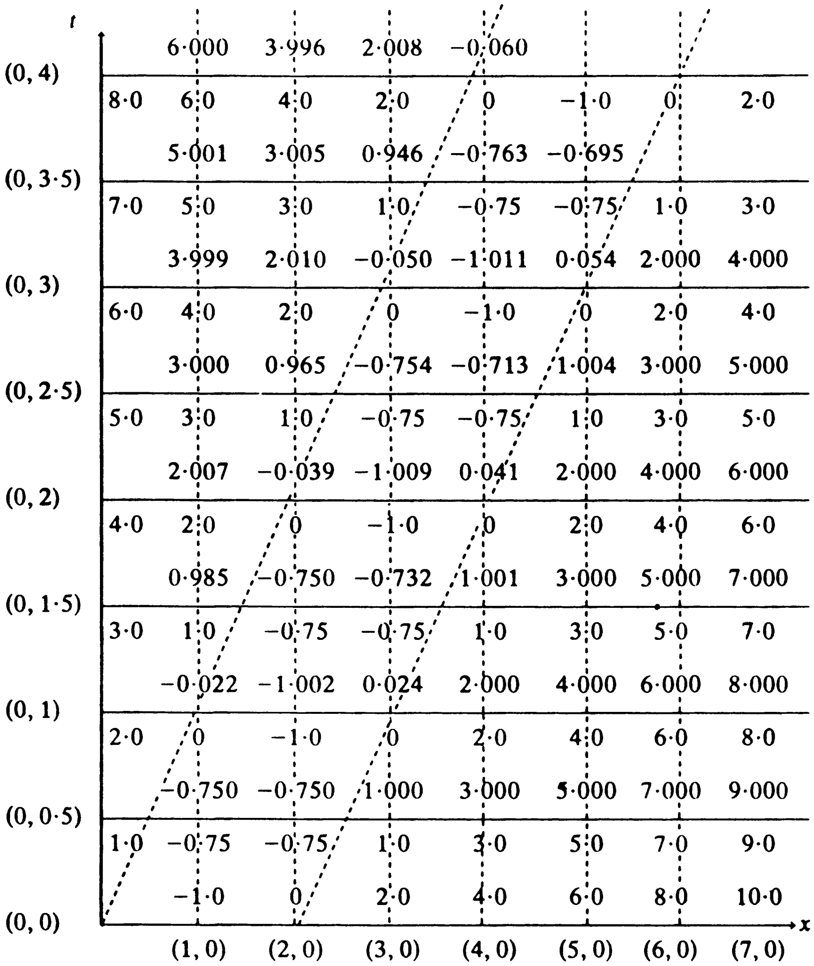
$$\frac{\partial \mathbf{U}}{\partial t} + \mathbf{A}\frac{\partial \mathbf{U}}{\partial x} = 0, \tag{4.13}$$

where

$$\mathbf{U} = [U_1, U_2, \ldots, U_N]^{\mathrm{T}}, \quad \frac{\partial \mathbf{U}}{\partial t} = \left[\frac{\partial U_1}{\partial t}, \frac{\partial U_2}{\partial t}, \ldots, \frac{\partial U_N}{\partial t}\right]^{\mathrm{T}},$$

etc., and $\mathbf{A}$ is an $N \times N$ matrix with real constant elements.

TABLE 4.1



t		(1, 0)	(2, 0)	(3, 0)	(4, 0)	(5, 0)	(6, 0)	(7, 0)
		6·000	3·996	2·008	−0·060			
(0, 4)								
	8·0	6·0	4·0	2·0	0	−1·0	0	2·0
		5·001	3·005	0·946	−0·763	−0·695		
(0, 3·5)								
	7·0	5·0	3·0	1·0	−0·75	−0·75	1·0	3·0
		3·999	2·010	−0·050	−1·011	0·054	2·000	4·000
(0, 3)								
	6·0	4·0	2·0	0	−1·0	0	2·0	4·0
		3·000	0·965	−0·754	−0·713	1·004	3·000	5·000
(0, 2·5)								
	5·0	3·0	1·0	−0·75	−0·75	1·0	3·0	5·0
		2·007	−0·039	−1·009	0·041	2·000	4·000	6·000
(0, 2)								
	4·0	2·0	0	−1·0	0	2·0	4·0	6·0
		0·985	−0·750	−0·732	1·001	3·000	5·000	7·000
(0, 1·5)								
	3·0	1·0	−0·75	−0·75	1·0	3·0	5·0	7·0
		−0·022	−1·002	0·024	2·000	4·000	6·000	8·000
(0, 1)								
	2·0	0	−1·0	0	2·0	4·0	6·0	8·0
		−0·750	−0·750	1·000	3·000	5·000	7·000	9·000
(0, 0·5)								
	1·0	−0·75	−0·75	1·0	3·0	5·0	7·0	9·0
		−1·0	0	2·0	4·0	6·0	8·0	10·0
(0, 0)								x

By Taylor's expansion,

$$\mathbf{U}_{i,j+1}=\mathbf{U}_{i,j}+k\frac{\partial \mathbf{U}_{i,j}}{\partial t}+\tfrac{1}{2}k^2\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{i,j}+\ldots .$$

By eqn (4.13),

$$\frac{\partial}{\partial t}\equiv -\mathbf{A}\frac{\partial}{\partial x}.$$

Hence,

$$\mathbf{U}_{i,j+1}=\mathbf{U}_{i,j}-k\mathbf{A}\left(\frac{\partial \mathbf{U}}{\partial x}\right)_{i,j}+\tfrac{1}{2}k^2\mathbf{A}^2\left(\frac{\partial^2\mathbf{U}}{\partial x^2}\right)_{i,j}+\ldots\ldots$$

The standard central-difference approximations for $\partial\mathbf{U}/\partial x$ and $\partial^2\mathbf{U}/\partial x^2$ then give the approximating difference equations

$$\mathbf{u}_{i,j+1}=\mathbf{u}_{i,j}-\tfrac{1}{2}p\mathbf{A}(\mathbf{u}_{i+1,j}-\mathbf{u}_{i-1,j})+\tfrac{1}{2}p^2\mathbf{A}^2(\mathbf{u}_{i-1,j}-2\mathbf{u}_{i,j}+\mathbf{u}_{i+1,j}).$$

(See Exercise 4.)

The formal development of an approximation to (4.11) is as follows. By Taylor's series,

$$\mathbf{U}_{i,j+1}=\mathbf{U}_{i,j}+k\frac{\partial \mathbf{U}}{\partial t}+\tfrac{1}{2}k^2\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{U}}{\partial t}\right)_{i,j}+\ldots.$$

In virtue of eqn (4.11) it follows that

$$\mathbf{U}_{i,j+1}=\mathbf{U}_{i,j}-k\left(\frac{\partial \mathbf{F}}{\partial x}\right)_{i,j}-\tfrac{1}{2}k^2\frac{\partial}{\partial t}\left(\frac{\partial \mathbf{F}}{\partial x}\right)_{i,j}+\ldots. \tag{4.14}$$

But

$$\frac{\partial}{\partial t}\frac{\partial \mathbf{F}}{\partial x}=\frac{\partial}{\partial x}\frac{\partial \mathbf{F}}{\partial t}=\frac{\partial}{\partial x}\left\{\frac{\partial \mathbf{F}}{\partial \mathbf{U}}\frac{\partial \mathbf{U}}{\partial t}\right\}=-\frac{\partial}{\partial x}\left\{\frac{\partial \mathbf{F}}{\partial \mathbf{U}}\frac{\partial \mathbf{F}}{\partial x}\right\},$$

where

$$\frac{\partial \mathbf{F}}{\partial \mathbf{U}}=A(\mathbf{U}),$$

the Jacobian matrix of **F** with respect to **U**, is defined by

$$A_{m,n}=\frac{\partial F_m}{\partial U_n}.$$

For example, if

$$\mathbf{U}=\begin{bmatrix}U_1\\U_2\end{bmatrix}\quad\text{and}\quad\mathbf{F}(\mathbf{U})=\begin{bmatrix}F_1(U)\\F_2(U)\end{bmatrix}$$

$$\text{then}\quad A(\mathbf{U})=\begin{bmatrix}\partial F_1/\partial U_1 & \partial F_1/\partial U_2\\ \partial F_2/\partial U_1 & \partial F_2/\partial U_2\end{bmatrix}.$$

Therefore eqn (4.14) can be written as

$$\mathbf{U}_{i,j+1}=\mathbf{U}_{i,j}-k\left(\frac{\partial \mathbf{F}}{\partial x}\right)_{i,j}+\tfrac{1}{2}k^2\frac{\partial}{\partial x}\left\{A(\mathbf{U})\frac{\partial \mathbf{F}}{\partial x}\right\}_{i,j}+\ldots,$$

and using the central-difference approximation for the last term as shown in Chapter 3, Exercise 8(a), we obtain the Lax–Wendroff approximation

$$\mathbf{u}_{i,j+1}=\mathbf{u}_{i,j}-\tfrac{1}{2}p(\mathbf{F}_{i+1,j}-\mathbf{F}_{i-1,j})$$
$$+\tfrac{1}{2}p^2\{A_{i+\frac{1}{2},j}(\mathbf{F}_{i+1,j}-\mathbf{F}_{i,j})-A_{i-\frac{1}{2},j}(\mathbf{F}_{i,j}-\mathbf{F}_{i-1,j})\}.$$

To avoid midpoint evaluations it is usual to approximate $A_{i+\frac{1}{2},j}$ by $\frac{1}{2}(A_{i,j}+A_{i+1,j})$ and $A_{i-\frac{1}{2},j}$ by $\frac{1}{2}(A_{i-1,j}+A_{i,j})$.

Further details concerning stability, well posedness, and the solution of eqn (4.11) are given in references 25 and 18.

The Courant–Friedrichs–Lewy (C.F.L.) condition for first-order equations

Assume that a first-order hyperbolic differential equation has been approximated by a difference equation of the form

$$u_{i,j+1}=au_{i-1,j}+bu_{i,j}+cu_{i+1,j}.$$

Then u_P, Fig. 4.4, depends on the values of u at the mesh points A, B, and C. Assume now that the characteristic curve through P of the hyperbolic equation meets the line AC at D and consider AC as an initial line segment. If the initial values along AC are altered then the solution value at P of the finite-difference equation will change, but these alterations will not affect the solution value at P of the differential equation which depends on the initial value at D. In this case u_P cannot converge to U_P as $h\to 0$, $k\to 0$. For convergence D must lie between A and C. (The C.F.L. condition.) Consider, for example, the Lax–Wendroff approximation (4.12). The slope dt/dx of the characteristic of the corresponding differential equation is given by $dt/1=dx/a$. For convergence of the difference equation, slope of $PD\geqslant$ slope of

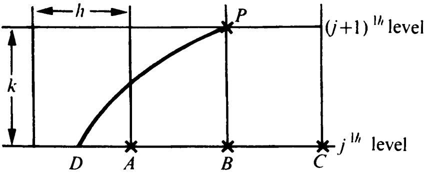


Fig. 4.4

PA, i.e. $1/a \geqslant k/h$, giving $ap \leqslant 1$, which coincides with the condition for stability, namely, $0 < ap \leqslant 1$, since $a > 0$, $p > 0$.

Wendroff's implicit approximation

An implicit approximation of second-order accuracy to the equation

$$a\frac{\partial U}{\partial x}+b\frac{\partial U}{\partial y}=c$$

at the point P, Fig. 4.5, is given by approximating $(\partial U/\partial x)_P$ and $(\partial U/\partial y)_P$ by

$$\frac{1}{2}\left\{\left(\frac{\partial U}{\partial x}\right)_H+\left(\frac{\partial U}{\partial x}\right)_F\right\} \quad \text{and} \quad \frac{1}{2}\left\{\left(\frac{\partial U}{\partial y}\right)_G+\left(\frac{\partial U}{\partial y}\right)_E\right\}$$

respectively, then approximating these derivatives by central-difference formulae to give the Wendroff equation

$$\frac{a}{2}\left\{\frac{u_B-u_A}{h}+\frac{u_C-u_D}{h}\right\}+\frac{b}{2}\left\{\frac{u_D-u_A}{k}+\frac{u_C-u_B}{k}\right\}=c,$$

which can be written as

$$(b-ap)u_D+(b+ap)u_C=(b+ap)u_A+(b-ap)u_B+2kc.$$

This is unconditionally stable. (See Exercise 5.) It cannot be used for pure initial-value problems, i.e. conditions on $t=0$ only, because it would give an infinite number of simultaneous equations. If, however, initial values are known on Ox, $x \geqslant 0$, and boundary values on Ot, $t \geqslant 0$, the equation can be used explicitly by writing it as

$$u_C=u_A+\frac{b-ap}{b+ap}(u_B-u_D)+\frac{2kc}{b+ap}, \tag{4.15}$$

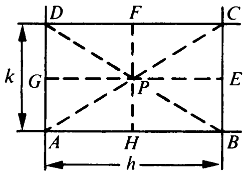


Fig. 4.5

and will give the solution in the quarter plane $x>0$, $t>0$. (See p. 191 for a numerical example.) It will also give approximation values in the strip $0<x\leqslant 1$, $t>0$, when initial values are known on *Ox*, $0\leqslant x\leqslant 1$, and boundary values are known on *Ot*, $t>0$.

Propagation of discontinuities in first-order equations

Consider the equation

$$\frac{\partial U}{\partial x}+\frac{\partial U}{\partial y}=1, \quad y\geqslant 0, \quad -\infty<x<\infty,$$

where U is known at points $P(x_{\mathrm{P}}, 0)$ on the x-axis. The characteristic direction is given by $\mathrm{d}x=\mathrm{d}y$ and along the characteristics, $\mathrm{d}U=\mathrm{d}y$. Hence the characteristic through P is $y=x-x_{\mathrm{P}}$ and the solution along it is $U=U_{\mathrm{P}}+y$.

Discontinuous initial values

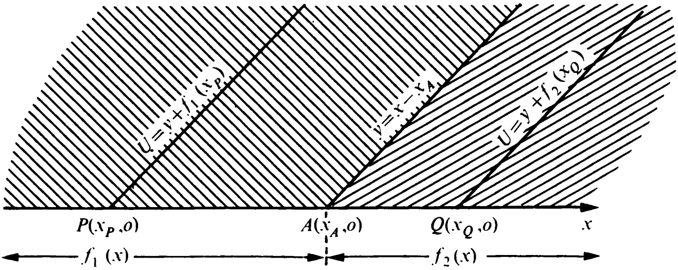


Fig. 4.6

Let

$$U(x,0)=f_1(x), \quad -\infty<x<x_{\mathrm{A}},$$

and

$$U(x,0)=f_2(x), \quad x_{\mathrm{A}}<x<\infty.$$

To the left of the characteristic $y=x-x_{\mathrm{A}}$ the solution $U_{(\mathrm{L})}$ is $U_{(\mathrm{L})}=f_1(x_{\mathrm{P}})+y$ along $y=x-x_{\mathrm{P}}$. To the right of the straight line $y=x-x_{\mathrm{A}}$ the solution $U_{(\mathrm{R})}$ is $U_{(\mathrm{R})}=f_2(x_Q)+y$ along $y=x-x_Q$.

Hence, for the same value of y in both solutions,

$$U_{(\mathrm{L})}-U_{(\mathrm{R})}=f_1(x_{\mathrm{P}})-f_2(x_Q).$$

Clearly, as x_P and x_Q both tend to x_A, $U_{(L)} - U_{(R)}$ is discontinuous along $y = x - x_A$ where

$$\lim_{x_P \to x_A} f_1(x_P) \neq \lim_{x_Q \to x_A} f_2(x_Q).$$

This shows that when the initial values are discontinuous at a particular point *A*, say, then the solution is discontinuous along the characteristic curve *C* from *A*. Moreover, the effect of this initial discontinuity does not diminish as we move away from *A* along *C*. With parabolic and elliptic differential equations the effect of an initial discontinuity is quite different as it tends to be localized and to diminish fairly rapidly with distance from the point of discontinuity.

Discontinuous initial derivatives

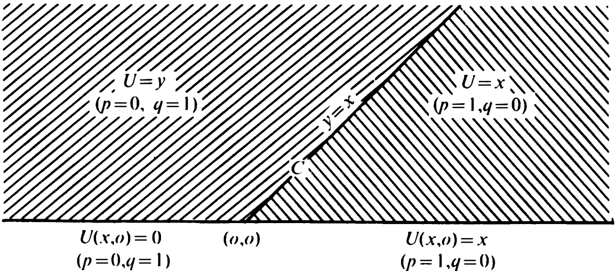


Fig. 4.7

Let

$$U(x, 0) = 0, \quad -\infty < x \leqslant 0,$$

and

$$U(x, 0) = x, \quad 0 < x < \infty.$$

This makes the initial derivative $p(x, 0) = \partial U(x, 0)/\partial x$ discontinuous at $(0, 0)$. The initial derivative $q(x, 0) = [\partial U(x, y)/\partial y]_{(x,0)}$ is also discontinuous at $(0, 0)$ because $p + q = 1$ from the differential equation. $U(x, 0)$ itself is clearly continuous at $(0, 0)$. As before, $U - U_P = y = x - x_P$ along the characteristic $y = x - x_P$ from $P(x_P, 0)$. Therefore the solution $U_{(L)}$ to the left of $y = x$ is

$$U_{(L)} = y, \quad -\infty < x \leqslant 0, \quad y \geqslant 0,$$

and the solution $U_{(R)}$ to the right of $y=x$ is

$$U_{(R)}=x, \quad 0<x<\infty, \quad y\geqslant 0$$

It is seen from this that $U_{(L)}=U_{(R)}$ along $y=x$ but that $p_{(L)}=\partial U_{(L)}/\partial x=0$ and $p_{(R)}=\partial U_{(R)}/\partial x=1$, i.e. the solution is continuous along the characteristic C from the point of discontinuity but the initial discontinuities in the partial derivatives are propagated undiminished across the solution domain along this characteristic. In many cases discontinuous partial derivatives arise from sudden changes in the direction of the boundary-initial curve as illustrated in the next section.

Discontinuities and finite difference methods

As shown above, a discontinuity in the initial data of a first-order equation is propagated across the solution domain along the characteristic from the point of discontinuity. In such a case one would expect the 'method of characteristics' on p. 178 to give a more accurate numerical solution than finite-difference methods because the solution corresponding to a particular initial value is developed along a characteristic that does not intersect the characteristic from a point of discontinuity. But the programming of the method of characteristics, especially for problems involving a set of simultaneous first-order equations, is much more difficult than the programming of difference methods. For this reason a great deal of research has been devoted to the formulation of finite-difference schemes that simulate the propagation of discontinuities along characteristics in the sense that rapid changes are confined to narrow regions. (See references 15 and 25.)

The 'blurring' of discontinuities that occurs with difference methods is illustrated in the following example. Consider the equation

$$\frac{\partial U}{\partial x}+\frac{\partial U}{\partial y}=1, \quad 0<x<\infty, \quad y>0, \tag{4.16}$$

where $U(0, y)=0, \quad 0<y<\infty,$
$U(x, 0)=0, \quad 0\leqslant x\leqslant 3,$
$U(x, 0)=x-3, \quad 3\leqslant x<\infty.$

In general, if $U=U_R$ at the initial point $R(x_R, y_R)$ then the characteristic through R and the solution along it are given by

$U-U_R=x-x_R=y-y_R$. Therefore:

(i) $U=x$ along the characteristic from $(0, y_R)$, $y_R>0$,
(ii) $U=y$ along the characteristic from $(x_R, 0)$ $0\leqslant x_R\leqslant 3$, and
(iii) $U=x-3$ along the characteristic from $(x_R, 0)$, $x_R\geqslant 3$.

The numerical values for this analytical solution are shown in Table 4.2 for $x=0(1)8$ and $y=0(1)5$. Although the boundary values and initial values are continuous at $(0, 0)$ and $(3, 0)$ it is seen that $p=\partial U/\partial x$ and $q=\partial U/\partial y$ are discontinuous at these points. The discontinuities at $(0, 0)$ arise from the sudden change in the direction of the boundary curve yOx at O. (N.B. As $p+q=1$ from the differential equation, $\delta p=-\delta q$.) The table also illustrates numerically how the discontinuities in p and q in the solution domain occur on the characteristics from $(0, 0)$ and $(3, 0)$, i.e. the discontinuities are propagated cleanly along these characteristics.

Let us now approximate eqn (4.16) by Wendroff's equation (4.15). By Fig. 4.5,

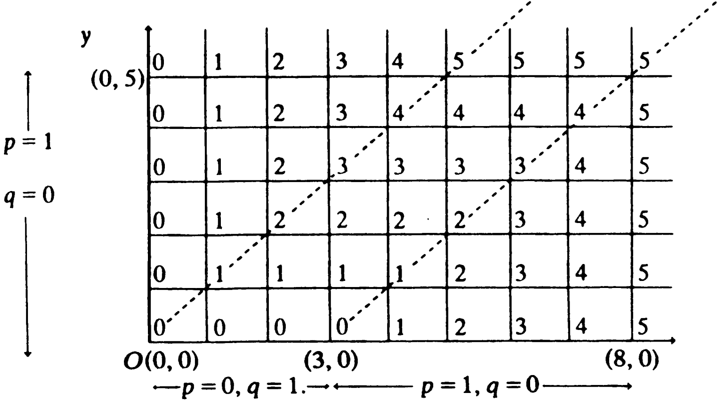
$$u_C=u_A+\frac{h-k}{h+k}(u_B-u_D)+\frac{2hk}{h+k}.$$

Case (*i*) Take $h=k=1$. Then

$$u_C=u_A+1, \quad \text{i.e.} \quad u_{i+1,j+1}=u_{i,j}+1,$$

and the application of this equation to the boundary and initial values of Table 4.2 shows that the numerical solution coincides

TABLE 4.2



0	1	2	3	4	5	5	5	5
0	1	2	3	4	4	4	4	5
0	1	2	3	3	3	3	4	5
0	1	2	2	2	2	3	4	5
0	1	1	1	1	2	3	4	5
0	0	0	0	1	2	3	4	5

with the differential equation solution. This happens because $u_{i+1,j+1}=u_{i,j}+1$ propagates the finite-difference solution forward along the characteristics from the initial points and boundary points and because the local truncation error at the point $(i+\frac{1}{2}, j+\frac{1}{2})$ is zero in this case.

Case (*ii*) Take $h=1$, $k=\frac{1}{2}$. Then

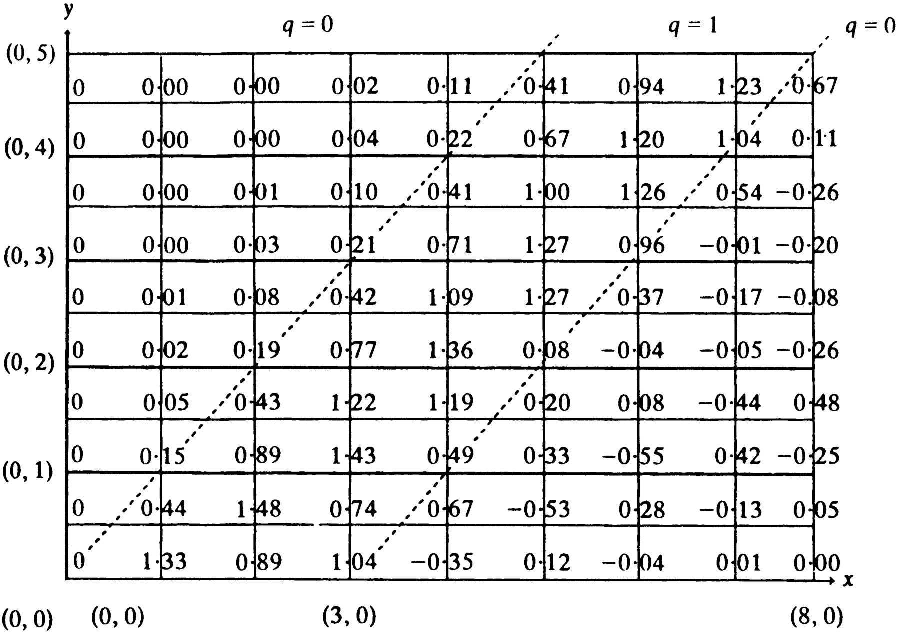
$$u_C = u_A + \tfrac{1}{3}(u_B - u_D) + \tfrac{2}{3}.$$

The numerical solution to $2D$ is displayed in Table 4.3. Every second row corresponds to a row of Table 4.2. The values of $q_{i,j}=(u_{i,j+1}-u_{i,j})/k$ are shown in Table 4.4. The analytical values for $q_{i,j}$ for the three different solution domains are indicated at the top of Table 4.4. The values for u clearly improve in accuracy as one moves away from the characteristics through (0, 0) and (3, 0) and Table 4.4 shows that the discontinuity in q along these characteristics is 'diffused' over an area each side of the characteristics.

TABLE 4.3

y									
(0, 5)	0	1·00	2·00	2·99	3·95	4·77	5·23	5·22	5·13
	0	1·00	2·00	2·98	3·90	4·57	4·76	4·61	4·80
(0, 4)	0	1·00	1·00	2·96	3·79	4·23	4·17	4·09	4·74
	0	1·00	1·99	2·91	3·58	3·73	3·54	3·82	4·87
(0, 3)	0	1·00	1·98	2·81	3·23	3·09	3·05	3·82	4·97
	0	1·00	1·94	2·60	2·68	2·46	2·87	3·91	5·01
(0, 2)	0	0·99	1·84	2·22	2·00	2·06	2·89	3·93	5·14
	0	0·96	1·63	1·60	1·41	1·96	2·85	4·15	4·90
(0, 1)	0	0·89	1·19	0·89	1·16	1·79	3·12	3·94	5·02
	0	0·67	0·44	0·52	0·83	2·06	2·98	4·01	5·00
	0	0	0	0	1	2	3	4	5 → x
	(0, 0)			(3, 0)					(8, 0)

TABLE 4.4



(0, 5)	0	0·00	0·00	0·02	0·11	0·41	0·94	1·23	0·67
(0, 4)	0	0·00	0·00	0·04	0·22	0·67	1·20	1·04	0·11
	0	0·00	0·01	0·10	0·41	1·00	1·26	0·54	−0·26
(0, 3)	0	0·00	0·03	0·21	0·71	1·27	0·96	−0·01	−0·20
	0	0·01	0·08	0·42	1·09	1·27	0·37	−0·17	−0·08
(0, 2)	0	0·02	0·19	0·77	1·36	0·08	−0·04	−0·05	−0·26
	0	0·05	0·43	1·22	1·19	0·20	0·08	−0·44	0·48
(0, 1)	0	0·15	0·89	1·43	0·49	0·33	−0·55	0·42	−0·25
	0	0·44	1·48	0·74	0·67	−0·53	0·28	−0·13	0·05
	0	1·33	0·89	1·04	−0·35	0·12	−0·04	0·01	0·00
(0, 0)	(0, 0)			(3, 0)					(8, 0)

Reduction of a first-order equation to a system of ordinary differential equations

This section can be omitted if the reader wishes to study immediately the numerical solution of second-order hyperbolic equations.

Consider the first-order hyperbolic equation

$$\frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}=0, \quad x>0, \quad t>0, \tag{4.17}$$

where a is a real positive constant and U satisfies the initial condition

$$U(x, 0)=g(x), \quad x\geqslant 0,$$

and the boundary condition

$$U(0, t)=b(t), \quad t>0.$$

Assume that the solution is needed in that strip R of the first quadrant $x>0$, $t>0$, defined by $R=[0<x<X]\times[t>0]$. (Read as the set of all points (x, t) such that $0<x<X$ and $t>0$.) Subdivide the interval $0\leqslant x\leqslant X$ into N parts each of width h, so that $Nh=X$, and discretize t in steps of length k. A typical mesh point (x_i, t_j) is then defined by

$$(x_i, t_j)=(ih, jk), \quad 0\leqslant i\leqslant N, \quad j=0, 1, 2, \ldots .$$

If the x-derivative at (x, t) is replaced by the backward-difference formula

$$\frac{\partial U}{\partial x}=\frac{1}{h}\{U(x, t)-U(x-h, t)\}+O(h), \tag{4.18}$$

and x is considered as a constant, eqn (4.17) can be written as the ordinary differential equation

$$\frac{\mathrm{d}U(t)}{\mathrm{d}t}=-\frac{a}{h}\{U(x, t)-U(x-h, t)\}+O(h).$$

On writing down this ordinary equation at the N mesh points $x_i=ih$, $i=1(1)N$, along time-level t, it follows that the values $V_i(t)$ approximating $U_i(t)$ are the exact solution values of the N equations

$$\begin{aligned}
\frac{\mathrm{d}V_1}{\mathrm{d}t}&=-\frac{a}{h}\{V_1-b(t)\}\\
\frac{\mathrm{d}V_2}{\mathrm{d}t}&=-\frac{a}{h}\{V_2-V_1\}\\
&\vdots\\
\frac{\mathrm{d}V_N}{\mathrm{d}t}&=-\frac{a}{h}\{V_N-V_{N-1}\}.
\end{aligned}$$

These can be expressed as

$$\frac{\mathrm{d}\mathbf{V}(t)}{\mathrm{d}t}=-a\mathbf{C}\mathbf{V}(t)+a\mathbf{b}(t), \tag{4.19}$$

where $\mathbf{V}(t)=[V_1(t), V_2(t), \ldots, V_N(t)]^T$, $\mathbf{C}$ is the $N\times N$ lower bidiagonal matrix

$$\mathbf{C}=\frac{1}{h}\begin{bmatrix} 1 & & & & \\ -1 & 1 & & & \\ & -1 & 1 & & \\ & & \ddots & \ddots & \\ & & & -1 & 1 \end{bmatrix}$$

and $\mathbf{b}(t)=[b(t)/h, 0, 0, \ldots, 0]^T$ is known from the boundary condition. When $\mathbf{b}(t)$ is a constant vector the solution of (4.19) is, by eqn (3.5),

$$\mathbf{V}(t)=\mathbf{C}^{-1}\mathbf{b}(t)+\{\exp(-at\mathbf{C})\}\{\mathbf{g}-\mathbf{C}^{-1}\mathbf{b}(t)\}, \qquad (4.20)$$

and satisfies the recurrence relationship

$$\mathbf{V}(t+k)=\mathbf{C}^{-1}\mathbf{b}(t)+\{\exp(-ak\mathbf{C})\}\{\mathbf{V}(t)-\mathbf{C}^{-1}\mathbf{b}(t)\}. \quad (4.21)$$

The approximating finite-difference equations

Replacement of the exponential in eqn (4.21) by an (S, T) Padé approximant, $S \geqslant T$, leads to an unconditionally stable implicit difference scheme that can always be put into an explicit form because the initial conditions and boundary conditions are known along the two axes of the first quadrant.

For example, the (1, 0) Padé approximant approximates (4.21) by the difference equations

$$\mathbf{u}_{j+1}=\mathbf{C}^{-1}\mathbf{b}_j+(\mathbf{I}+ak\mathbf{C})^{-1}\{\mathbf{u}_j-\mathbf{C}^{-1}\mathbf{b}_j\},$$

giving the implicit scheme

$$(\mathbf{I}+ak\mathbf{C})\mathbf{u}_{j+1}=\mathbf{u}_j+ak\mathbf{b}_j.$$

In detail, the equations are

$$(1+ap)u_{1,j+1}=u_{1,j}+apb_j$$

and

$$-apu_{i-1,j+1}+(1+ap)u_{i,j+1}=u_{i,j}, \quad i=2(1)N,$$

where $p=k/h$. A small modification enables us to write them more compactly. The boundary value has been treated as a constant over $t_j \leqslant t \leqslant t_{j+1}$ and could be given its value at t_j or t_{j+1}, or even averaged. For implicit schemes it is usually assigned its value at t_{j+1}. All N equations of the $(1, 0)$ approximant can then be written as

$$-apu_{i-1,j+1} + (1+ap)u_{i,j+1} = u_{i,j}, \quad i = 1(1)N. \tag{4.22}$$

They can obviously be expressed explicitly as

$$u_{i,j+1} = (apu_{i-1,j+1} + u_{i,j})/(1+ap), \tag{4.23}$$

and are shown to be unconditionally stable in Exercise 7. The principal part of the local truncation error of (4.22) at (ih, jk) is

$$\left[-\tfrac{1}{2}ah\frac{\partial^2 U}{\partial x^2} - \tfrac{1}{2}k\frac{\partial^2 U}{\partial t^2}\right]_{i,j}.$$

(See Exercise 6.)

A comment on the non-stiffness of the equations

Although matrix $\mathbf{C}$ has N eigenvalues λ_s each equal to $1/h$, it has N linearly independent eigenvectors $\mathbf{v}_s$, so the vector of initial values $\mathbf{g}$, with its N components $g_1, g_2, \ldots, g_N$, can be expressed as $\mathbf{g} = \sum_{s=1}^{N} c_s \mathbf{v}_s$, c_s constants. Hence the solution (4.20) can be written as

$$\mathbf{V}(t) = \mathbf{C}^{-1}\mathbf{b}(t) + \left\{\sum_{s=1}^{N} c_s[\exp(-at\lambda_s)]\mathbf{v}_s\right\} - \{\exp(-at\mathbf{C})\}\mathbf{C}^{-1}\mathbf{b}(t).$$

This proves that the system of ordinary differential equations is non-stiff because the components $\{\exp(-at\lambda_s)\}\mathbf{v}_s$ of $\mathbf{V}(t)$ do not decay at different rates, as occurs with parabolic equations. The components are also non-oscillatory because a, t and $\lambda_s = 1/h$ are real.

The Crank–Nicolson scheme

It is easily shown that the $(1, 1)$ Padé approximant to $\exp(-at\mathbf{C})$ in eqn (4.20) leads to the explicit equations,

$$u_{1,j+1} = [(1-\tfrac{1}{2}ap)u_{1,j} + \tfrac{1}{2}ap(u_{0,j+1} + u_{0,j})]/(1+\tfrac{1}{2}ap),$$

and

$$u_{i,j+1}=[\tfrac{1}{2}apu_{i-1,j+1}+(1-\tfrac{1}{2}ap)u_{i,j}+\tfrac{1}{2}apu_{i-1,j}]/(1+\tfrac{1}{2}ap),$$
$$i=2(1)N. \quad (4.24)$$

These are A_0-stable and the principal part of the local truncation error at (ih, jk) is

$$\left[-\tfrac{1}{2}ah\frac{\partial^2 U}{\partial x^2}-\tfrac{1}{12}k^2\frac{\partial^3 U}{\partial t^3}\right]_{i,j}.$$

The k-components of the local truncation errors of eqns (4.23) and (4.24) can be improved by at least one power of k by extrapolating in t, as described in Chapter 3, but this may not be worth the computational effort involved if the h-component is not improved to at least $O(h^2)$, as described in the next section.

An improved approximation to $\partial U/\partial x$

By Maclaurin's expansion,

$$U(x-h, t)=U(x, t)-h\frac{\partial U(x, t)}{\partial x}+\tfrac{1}{2}h^2\frac{\partial^2 U(x, t)}{\partial x^2}+O(h^3),$$

$$U(x-2h, t)=U(x, t)-2h\frac{\partial U(x, t)}{\partial x}+2h^2\frac{\partial^2 U(x, t)}{\partial x^2}+O(h^3).$$

Elimination of $h^2\,\partial^2 U/\partial x^2$ leads to

$$\frac{\partial U(x, t)}{\partial x}=\frac{1}{2h}\{3U(x, t)-4U(x-h, t)+U(x-2h, t)\}+O(h^2). \quad (4.25)$$

As this involves three points along any one time-row it can be used only for $i=2, 3, \ldots, N$. At $i=1$ we shall use the previous backward difference replacement (4.18). The equation

$$\frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}=0, \quad 0<x<X, \quad t>0,$$

is then approximated at the mesh points ih, $i=1(1)N$, along time-level t, by the system of ordinary equations

$$\frac{\mathrm{d}V_1}{\mathrm{d}t} = -\frac{2a}{2h}\{V_1 - b(t)\}$$

$$\frac{\mathrm{d}V_2}{\mathrm{d}t} = -\frac{a}{2h}\{3V_2 - 4V_1 + b(t)\}$$

$$\frac{\mathrm{d}V_3}{\mathrm{d}t} = -\frac{a}{2h}\{3V_3 - 4V_2 + V_1\}$$

$$\vdots$$

$$\frac{\mathrm{d}V_N}{\mathrm{d}t} = -\frac{a}{2h}\{3V_N - 4V_{N-1} + V_{N-2}\}.$$

These can be written in matrix form as

$$\frac{\mathrm{d}\mathbf{V}(t)}{\mathrm{d}t} = -\tfrac{1}{2}a\mathbf{D}\mathbf{V}(t) + \tfrac{1}{2}a\mathbf{b}(t), \tag{4.26}$$

where the matrix $\mathbf{D}$ of order N is

$$\mathbf{D} = \frac{1}{h}\begin{bmatrix} 2 & & & & & \\ -4 & 3 & & & & \\ 1 & -4 & 3 & & & \\ & 1 & -4 & 3 & & \\ & & \ddots & \ddots & \ddots & \\ & & & 1 & -4 & 3 \end{bmatrix} \quad \text{and} \quad \mathbf{b}(t) = \frac{1}{h}\begin{bmatrix} 2b(t) \\ -b(t) \\ 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

The solution of eqn (4.26) satisfying the initial condition $\mathbf{V}(0) = \mathbf{g}$ is, by eqn (3.5), for $\mathbf{b}(t)$ constant

$$\mathbf{V}(t) = \mathbf{D}^{-1}\mathbf{b}(t) + \{\exp(-\tfrac{1}{2}at\mathbf{D})\}\{\mathbf{g} - \mathbf{D}^{-1}\mathbf{b}(t)\}. \tag{4.27}$$

Gustafsson's theorems, reference 11, justify this mixture of approximations to $\partial U/\partial x$ provided the Padé approximant chosen to approximate the exponential is such that when applied to the set of ordinary differential equations given by the lower-order approximation (4.18), the resulting difference equations are unconditionally stable. The difference equations given by (4.26) will also be unconditionally stable and converge at the same rate as the more accurate approximation.

Equations (4.26) are non-stiff because matrix $\mathbf{D}$ has one eigenvalue $2/h$ and $(N-1)$ eigenvalues $3/h$, so the range of eigenvalues is small.

The solution (4.27) satisfies the recurrence relationship

$$\mathbf{V}(t+k)=\mathbf{D}^{-1}\mathbf{b}(t)+\{\exp(-\tfrac{1}{2}ak\mathbf{D})\}\{\mathbf{V}(t)-\mathbf{D}^{-1}\mathbf{b}(t)\}, \quad (4.28)$$

on treating $\mathbf{b}(t)$ as constant in $t_j \leqslant t \leqslant t_{j+1}$.

The (1, 0) Padé approximant

Approximating the exponential in (4.28) by its (1, 0) Padé approximant $(\mathbf{I}+\frac{1}{2}ak\mathbf{D})^{-1}$ and replacing $\mathbf{b}(t)$ by $\mathbf{b}(t+k)$ gives the L_0-stable system

$$(\mathbf{I}+\tfrac{1}{2}ak\mathbf{D})\mathbf{u}_{j+1}-\tfrac{1}{2}ak\mathbf{b}_{j+1}=\mathbf{u}_j.$$

In detail, the equations are

$$\begin{aligned}(1+ap)u_{1,j+1}-apb_{j+1}&=u_{1,j}\\ -2apu_{1,j+1}+(1+\tfrac{3}{2}ap)u_{2,j+1}+\tfrac{1}{2}apb_{j+1}&=u_{2,j}\\ \tfrac{1}{2}apu_{i-2,j+1}-2apu_{i-1,j+1}+(1+\tfrac{3}{2}ap)u_{i,j+1}&=u_{i,j}, \quad i=3(1)N.\end{aligned}$$

Although they look implicit they can be solved explicitly for $u_{i,j+1}$, $i=1(1)N$.

The principal part of the local truncation error is

$$\left[-\tfrac{1}{2}ah\frac{\partial^2 U}{\partial x^2}-\tfrac{1}{2}k\frac{\partial^2 U}{\partial t^2}\right] \quad \text{at } (h, jk)$$

and

$$\left[-\tfrac{1}{3}ah^2\frac{\partial^3 U}{\partial x^3}-\tfrac{1}{2}k\frac{\partial^2 U}{\partial t^2}\right] \quad \text{at } (ih, jk), \quad i=2(1)N.$$

This scheme is worth extrapolating in t because the local truncation error is $O(h^2)$ in x and the equations L_0-stable. As in Chapter 3 we calculate $(\mathbf{I}-2k\mathbf{A})\mathbf{u}^{(1)}=\mathbf{u}(t)$, $(\mathbf{I}-k\mathbf{A})^2\mathbf{u}^{(2)}=\mathbf{u}(t)$, where $\mathbf{A}=-\frac{1}{2}a\mathbf{D}$, then $\mathbf{u}^{(E)}(t+2k)=2\mathbf{u}^{(2)}-\mathbf{u}^{(1)}$.

All the equations can be solved explicitly and the principal part of the local truncation error is

$$\left[-\tfrac{1}{2}ah\frac{\partial^2 U}{\partial x^2}+\tfrac{4}{3}k^2\frac{\partial^3 U}{\partial t^2}\right] \quad \text{at } (1, j)$$

and
$$\left[-\tfrac{1}{3}ah^2\frac{\partial^3 U}{\partial x^3}+\tfrac{4}{3}k^2\frac{\partial^3 U}{\partial t^3}\right] \quad \text{at } (i,j), \quad i=2(1)N.$$

Twizell and Khaliq, references 35 and 13 did numerical experiments with the (1, 1), (2, 0), (2, 1), and (2, 2) Padé approximants and observed that the (1, 1) and (2, 0) approximants with the low-order (4.18) replacement of the space derivative, and the (2, 1) and (2, 2) approximants with the higher-order (4.25) replacement of the space derivative, gave better results than any other method in the literature existing to 1983.

A word of caution on the central-difference approximation to $\partial U/\partial x$

If $\partial U/\partial x$ is approximated at (ih, jk) by
$$[U(x+h,t)-U(x-h,t)]/2h,$$
then the system of differential equations approximating
$$\frac{\partial U}{\partial t}+a\frac{\partial U}{\partial x}=0, \quad a>0, \quad 0<x<X, \quad t>0, \tag{4.29}$$
$U(x,0)=g(x)$, $U(0,t)=b(t)$, at the mesh points ih, $i=1(1)N$, along time-level t is
$$\frac{\mathrm{d}\mathbf{V}(t)}{\mathrm{d}t}=-\tfrac{1}{2}a\mathbf{B}\mathbf{V}(t)+\tfrac{1}{2}a\mathbf{b}(t), \tag{4.29a}$$
where $\mathbf{V}(t)=[V_1(t), V_2(t),\ldots, V_N(t)]^{\mathrm{T}}$,
$$\mathbf{B}=\frac{1}{h}\begin{bmatrix} 0 & 1 & & & & \\ -1 & 0 & 1 & & & \\ & -1 & 0 & 1 & & \\ & & & \cdot & & \\ & & & & \cdot & \\ & & & -1 & 0 & 1 \\ & & & & -1 & 0 \end{bmatrix} \quad \text{of order } N,$$
and $h\mathbf{b}(t)=[b(t),0,0,\ldots,-V_{N+1}(t)]^{\mathrm{T}}$. V_{N+1} is the solution at $\{(N+1)h, t\}$ and would be known only for periodic boundary conditions. That part of the solution corresponding to the term
$$-\tfrac{1}{2}a\mathbf{B}\mathbf{V}(t) \quad \text{is} \quad \{\exp(-\tfrac{1}{2}at\mathbf{B})\}\mathbf{g}. \tag{4.30}$$

The eigenvalues λ_s of $\mathbf{B}$ are $\lambda_s = 2(\sqrt{-1})\cos s\pi/N + 1$, $s = 1(1)N$, and are all different. Hence the N eigenvectors $\mathbf{v}_s$ of $\mathbf{B}$ are linearly independent so the initial vector $\mathbf{g} = [g_1, g_2, \ldots, g_N]^T$ can be written as $\mathbf{g} = \sum_1^N c_s\mathbf{v}_s$. Equation (4.30) then gives that the complementary function of the differential equation is

$$\sum_1^N c_s\{\exp(-\tfrac{1}{2}at\lambda_s)\}\mathbf{v}_s.$$

But λ_s is complex and $\exp(\sqrt{-1}\,\theta) = \cos\theta + \sqrt{-1}\sin\theta$, showing that the solution $\mathbf{V}(t)$ contains oscillatory terms. These are due entirely to the central-difference approximation to $\partial U/\partial x$ and would not occur with backward-difference approximations.

The numerical solution shown in Fig. 4.8 was calculated by Khaliq, reference 13, for eqn (4.29*a*), at $t = 1$ with $a = 1$, $g(x) = \sin 4\pi x$, $0 \leqslant x \leqslant 1$, $b(t) = -\sin 4\pi t$, $h = 1/80$ and $p = k/h = 4$, using the (1, 0) Padé approximant. The solution of the partial differential equation is $U(x, t) = \sin 4\pi(x - t)$. The diagram clearly shows

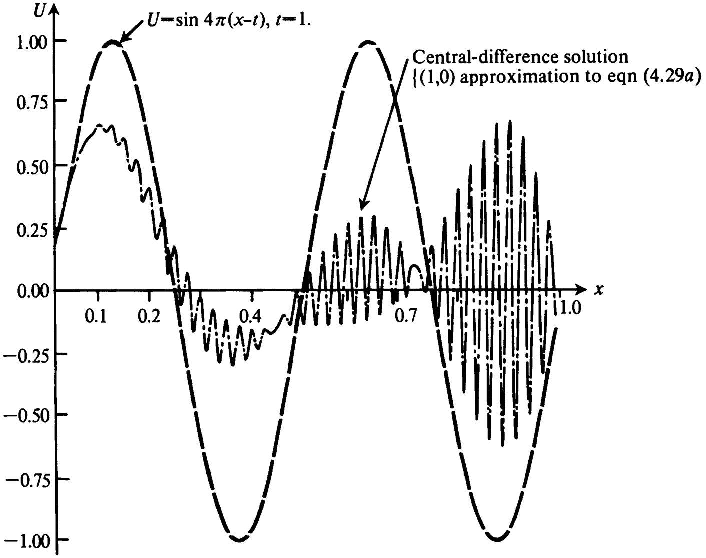


Fig. 4.8

that the effect of the oscillatory terms in the finite-difference solution increases with increasing x.

Second-order quasi-linear hyperbolic equations

Consider the second-order partial differential equation

$$a\frac{\partial^2 U}{\partial x^2}+b\frac{\partial^2 U}{\partial x\,\partial y}+c\frac{\partial^2 U}{\partial y^2}+e=0, \tag{4.31}$$

where a, b, c, and e may be functions of x, y, U, $\partial U/\partial x$, and $\partial U/\partial y$ but *not* of $\partial^2 U/\partial x^2$, $\partial^2 U/\partial x\,\partial y$, and $\partial^2 U/\partial y^2$, i.e. the second-order derivatives occur only to the first degree. Such an equation is said to be quasi-linear.

It will be shown that at each point of the solution domain there are two directions along which the integration of the partial differential equation transforms to the integration of an equation involving total differentials only. In other words, in these directions the equation to be integrated is not complicated by the presence of partial derivatives in other directions. Furthermore, it will be seen that this leads to a natural classification of partial differential equations.

Denote the partial derivatives by

$$\frac{\partial U}{\partial x}=p,\quad \frac{\partial U}{\partial y}=q,\quad \frac{\partial^2 U}{\partial x^2}=r,\quad \frac{\partial^2 U}{\partial x\,\partial y}=s \quad\text{and}\quad \frac{\partial^2 U}{\partial y^2}=t.$$

Let C be a curve in the x–y plane on which it is assumed that we know the solution values U of eqn (4.31), together with values for p and q related to U through the equation

$$\mathrm{d}U=\frac{\partial U}{\partial x}\,\mathrm{d}x+\frac{\partial U}{\partial y}\,\mathrm{d}y=p\,\mathrm{d}x+q\,\mathrm{d}y,$$

where $\mathrm{d}y/\mathrm{d}x$ is the slope of the tangent to C. (Curve C is *not* an initial curve on which initial values for U, p, and q are specified. The reason for this will be apparent later.) The question to ask now is this. Is it possible to find values for r, s, and t on C that will satisfy the partial differential equation, namely,

$$ar+bs+ct+e=0. \tag{4.32}$$

As r, s, and t on C must satisfy the equations

$$\mathrm{d}p=\frac{\partial p}{\partial x}\,\mathrm{d}x+\frac{\partial p}{\partial y}\,\mathrm{d}y=r\,\mathrm{d}x+s\,\mathrm{d}y \tag{4.33}$$

and

$$\mathrm{d}q=\frac{\partial q}{\partial x}\,\mathrm{d}x+\frac{\partial q}{\partial y}\,\mathrm{d}y=s\,\mathrm{d}x+t\,\mathrm{d}y, \tag{4.34}$$

it is seen that eqns (4.32), (4.33), and (4.34) are three equations for r, s, and t. Elimination of r and t from eqn (4.32) by means of eqns (4.33) and (4.34) leads to

$$\frac{a}{\mathrm{d}x}(\mathrm{d}p-s\,\mathrm{d}y)+bs+\frac{c}{\mathrm{d}y}(\mathrm{d}q-s\,\mathrm{d}x)+e=0,$$

i.e.

$$s\left\{a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-b\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)+c\right\}-\left\{a\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}\right\}=0, \tag{4.35}$$

where $\mathrm{d}y/\mathrm{d}x$ is the slope of the tangent to C.

By hypothesis, eqn (4.32) is quasi-linear so a, b, c, and e are independent of r, s, and t. Hence eqn (4.35) is independent of r and t. It can also be made independent of s by choosing the curve C so that the slope of the tangent at each point on C is a root of the equation

$$a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-b\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)+c=0. \tag{4.36}$$

By eqns (4.36) and (4.35) it follows that along directions tangential to C from points on C,

$$a\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}=0. \tag{4.37}$$

Hence we have shown that at each point of the solution domain there are two directions, given by the roots of eqn (4.36), along which there are relationships between the total differentials $\mathrm{d}p$ and $\mathrm{d}q$, given by eqn (4.37), that are independent of partial derivatives in other directions. As will be seen later this relationship can be used to solve the original differential equation numerically by a series of step-by-step integrations.

The directions given by the roots of eqn (4.36) are called the *characteristic directions* and the partial differential equation is said to be *hyperbolic, parabolic,* or *elliptic* according to whether these roots are real and distinct, equal, or complex, respectively, i.e. according to whether $b^2-4ac \gtreqless 0$. The best-known examples in these classes are the hyperbolic 'wave-equation' $\partial^2U/\partial t^2 = \partial^2U/\partial x^2$, the parabolic 'heat-conduction' or 'diffusion' equation $\partial U/\partial t = \partial^2U/\partial x^2$, and the elliptic 'Laplace equation' $\partial^2U/\partial x^2 + \partial^2U/\partial y^2 = 0$.

Assume eqn (4.31) is hyperbolic and that the roots of eqn (4.36) are $dy/dx = f$ and $dy/dx = g$. Then the curve through the point $P(x, y)$ whose slope at every point is f is said to be an f characteristic. Clearly there are two different characteristic curves through every point of the solution domain of a second-order hyperbolic equation.

It should be noted that the classification of a partial differential equation, and consequently its method of solution, may depend on the region in which the solution is to be found. For example, the characteristic directions of the equation

$$y\frac{\partial^2U}{\partial x^2} + x\frac{\partial^2U}{\partial x\,\partial y} + y\frac{\partial^2U}{\partial y^2} = F(x, y, U, p, q)$$

are given by the roots m_1, m_2 of the quadratic

$$ym^2 - xm + y = 0, \quad m = dy/dx,$$

which are real, equal or complex according to whether $x^2 \gtreqless 4y^2$. Thus, the equation is hyperbolic when $|x| > 2\,|y|$, parabolic along $|x| = 2\,|y|$, and elliptic for $|x| < 2\,|y|$.

Solution of hyperbolic equations by the method of characteristics

Summarizing the previous work, the slopes of the characteristic directions associated with the equation

$$a\frac{\partial^2U}{\partial x^2} + b\frac{\partial^2U}{\partial x\,\partial y} + c\frac{\partial^2U}{\partial y^2} + e = 0 \tag{4.38}$$

are given by the roots of the quadratic equation

$$a\left(\frac{dy}{dx}\right)^2 - b\left(\frac{dy}{dx}\right) + c = 0, \tag{4.39}$$

and along these characteristic directions the differentials $\mathrm{d}p$ and $\mathrm{d}q$ are related by the equation

$$a\frac{\mathrm{d}y}{\mathrm{d}x}\frac{\mathrm{d}p}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}=0,$$

which can be written as

$$a\frac{\mathrm{d}y}{\mathrm{d}x}\mathrm{d}p+c\,\mathrm{d}q+e\,\mathrm{d}y=0. \tag{4.40}$$

Assuming that eqn (4.38) is hyperbolic the roots of eqn (4.39) will be real and distinct. Let them be

$$\frac{\mathrm{d}y}{\mathrm{d}x}=f \quad \text{and} \quad \frac{\mathrm{d}y}{\mathrm{d}x}=g. \tag{4.41}$$

Let Γ be a *non-characteristic curve* along which initial values for U, p, and q are known. Let P and Q be points on Γ that are close together and let the f characteristic through P intersect the g characteristic through Q at the point $R(x_R, y_R)$, Fig. 4.9.

As a first approximation we may regard the arcs PR and QR as straight lines of slopes f_P and g_Q respectively. Then eqns (4.41) can be approximated by

$$y_R-y_P=f_P(x_R-x_P) \tag{4.42}$$

and

$$y_R-y_Q=g_Q(x_R-x_Q), \tag{4.43}$$

giving two equations for the two unknowns x_R, y_R.

By eqn (4.40) the differential relationships along the charac-

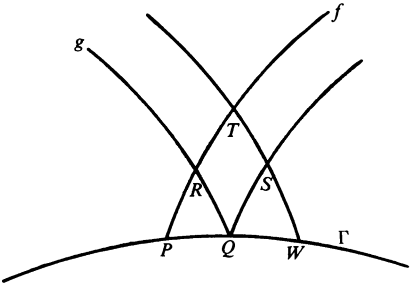


Fig. 4.9

teristics are

$$af\,\mathrm{d}p + c\,\mathrm{d}q + e\,\mathrm{d}y = 0 \tag{4.44}$$

and

$$ag\,\mathrm{d}p + c\,\mathrm{d}q + e\,\mathrm{d}y = 0. \tag{4.45}$$

The first one can be approximated along *PR* by the equation

$$a_P f_P(p_R - p_P) + c_P(q_R - q_P) + e_P(y_R - y_P) = 0, \tag{4.46}$$

and the second along *QR* by the equation

$$a_Q g_Q(p_R - p_Q) + c_Q(q_R - q_Q) + e_Q(y_R - y_Q) = 0. \tag{4.47}$$

These are two equations for the two unknowns p_R, q_R, as soon as x_R, y_R have been calculated from (4.42) and (4.43). The value of U at R can then be obtained from

$$\mathrm{d}U = \frac{\partial U}{\partial x}\,\mathrm{d}x + \frac{\partial U}{\partial y}\,\mathrm{d}y = p\,\mathrm{d}x + q\,\mathrm{d}y,$$

by replacing the values of p and q along *PR* by their average values and approximating the last equation by

$$u_R - u_P = \tfrac{1}{2}(p_P + p_R)(x_R - x_P) + \tfrac{1}{2}(q_P + q_R)(y_R - y_P). \tag{4.48}$$

This first approximation for u_R can now be improved by replacing the pivotal values of the various coefficients by average values. Equations (4.42) and (4.43) for improved values of x_R and y_R then become

$$y_R - y_P = \tfrac{1}{2}(f_P + f_R)(x_R - x_P) \tag{4.49}$$

and

$$y_R - y_Q = \tfrac{1}{2}(g_Q + g_R)(x_R - x_Q), \tag{4.50}$$

and eqns (4.46), (4.47) for improved values of p_R, q_R become

$$\tfrac{1}{2}(a_P + a_R)\tfrac{1}{2}(f_P + f_R)(p_R - p_P) + \tfrac{1}{2}(c_P + c_R)(q_R - q_P) + \tfrac{1}{2}(e_P + e_R)(y_R - y_P) = 0 \tag{4.51}$$

and

$$\tfrac{1}{2}(a_Q + a_R)\tfrac{1}{2}(g_Q + g_R)(p_R - p_Q) + \tfrac{1}{2}(c_Q + c_R)(q_R - q_Q) + \tfrac{1}{2}(e_Q + e_R)(y_R - y_Q) = 0. \tag{4.52}$$

An improved value for u_R can then be found from eqn (4.48). Repetition of this last cycle of operations will eventually yield u_R

to the accuracy warranted by these finite-difference approximations. Provided Q is close to P the number of iterations will usually be small.

In this way we can calculate solution values at the grid points R and S, Fig. 4.9, and thence proceed to the grid-point T, and so on.

Example 4.4

Use the method of characteristics to derive a solution of the quasi-linear equation

$$\frac{\partial^2 U}{\partial x^2} - U^2\frac{\partial^2 U}{\partial y^2} = 0,$$

at the first characteristic grid point between $x = 0.2$ and 0.3, $y > 0$, where U satisfies the conditions

$$U = 0.2 + 5x^2 \quad \text{and} \quad \frac{\partial U}{\partial y} = 3x,$$

along the initial line $y = 0$, for $0 \leqslant x \leqslant 1$.

Since U is given as a continuous function of x along Ox the initial value of $p = \partial U/\partial x$ is $10x$. The slopes of the characteristics are the roots of the equation $m^2 - U^2 = 0$. Hence

$$f = U = -g.$$

In this example the characteristics depend on the solution so the network of characteristics can be built-up only as the solution unfolds.

Initially,
$$U = 0.2 + 5x^2 = f = -g,$$
$$p = 10x \quad \text{and} \quad q = 3x.$$

Also $a = 1$, $b = e = 0$, $c = -U^2$, therefore, Fig. 4.10

$$f_P = 0.4, \quad g_Q = -0.65, \quad p_P = 2.0, \quad p_Q = 3.0, \quad U_P = 0.4,$$
$$U_Q = 0.65, \quad q_P = 0.6, \quad q_Q = 0.9, \quad c_P = -0.16, \quad c_Q = -0.4225.$$

By eqns (4.42) and (4.43),

$$y_R = 0.4(x_R - 0.2)$$

and

$$y_R = -0.65(x_R - 0.3),$$

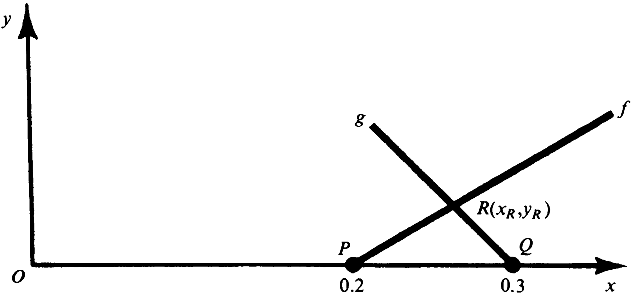


Fig. 4.10

giving, as a first approximation,

$$x_R = 0.26190, \quad y_R = 0.024762,$$

to five significant figures.

The differential relationships along the characteristics are, by eqns (4.46) and (4.47),

$$0.4(p_R - 2.0) - 0.16(q_R - 0.6) = 0,$$

and

$$-0.65(p_R - 3.0) - 0.4225(q_R - 0.9) = 0.$$

Their solution is

$$p_R = 2.45524; \quad q_R = 1.73810.$$

By eqn (4.48),

$$u_R = 0.4 + \tfrac{1}{2}(2.0 + 2.45524)(0.0619) + \tfrac{1}{2}(1.73810 + 0.6)(0.024762)$$
$$= 0.56684.$$

For the second approximation,

$$f_R = -g_R = u_R = 0.56684; \quad c_R = -u_R^2 = -0.32131.$$

By eqns (4.49) and (4.50), more accurate values for x_R, y_R are given by

$$y_R = \tfrac{1}{2}(0.4 + 0.56684)(x_R - 0.2)$$

and

$$y_R = -\tfrac{1}{2}(0.65 + 0.56684)(x_R - 0.3),$$

from which,

$$x_R = 0.25572, \quad y_R = 0.026938.$$

By eqns (4.51) and (4.52),

$$\tfrac{1}{2}(0.4+0.56684)(p_R-2.0)-\tfrac{1}{2}(0.16+0.32131)(q_R-0.6)=0,$$
$$-\tfrac{1}{2}(0.65+0.56684)(p_R-3.0)-\tfrac{1}{2}(0.4225+0.32131)(q_R-0.9)=0.$$

These equations give the improved values

$$p_R = 2.53117; \quad q_R = 1.66700.$$

Hence the second approximation to u_R, by eqn (4.48) is

$$u_R = 0.4+\tfrac{1}{2}\{(2+2.53117)(0.05572)+(0.6+1.6670)(0.026938)\}$$
$$= 0.55677.$$

It is left to the reader to show that the next iteration gives

$$x_R = 0.25578, \quad y_R = 0.02668,$$
$$p_R = 2.52876, \quad q_R = 1.67637,$$

and

$$u_R = 0.55667.$$

Since, to four decimal places,

$$u_R^{(1)} = 0.5668, \quad u_R^{(2)} = 0.5568 \quad \text{and} \quad u_R^{(3)} = 0.5567,$$

it is obvious that the solution of the finite-difference equations for u_R is 0.5567, to this degree of accuracy. A fourth iteration does, in fact, give $u_R = 0.55666$ to five decimal places.

Additional comments on characteristics

A characteristic as an initial curve

When the curve on which initial values are given is itself a characteristic the equation can have no solution unless the initial conditions satisfy the necessary differential relationship for this characteristic. If they do, the solution will be unique along the initial curve but nowhere else, as is illustrated in the example below. It is also impossible in this case to use the method of characteristics to extend the solution from points on the initial curve to points off it, because the locus of all points such as R in Fig. 4.9 is the initial curve itself.

Consider the equation

$$\frac{\partial^2 U}{\partial x^2}-\frac{\partial^2 U}{\partial x\,\partial y}-6\frac{\partial^2 U}{\partial y^2}=0.$$

The characteristic directions are given by

$$\left(\frac{dy}{dx}\right)^2+\left(\frac{dy}{dx}\right)-6=0=\left(\frac{dy}{dx}+3\right)\left(\frac{dy}{dx}-2\right),$$

so the characteristics are the straight lines $y+3x=\text{constant}$, and $y-2x=\text{constant}$. Let the initial curve be the characteristic $y-2x=0$. The differential relationship along this line by eqn (4.40) is $2\,dp-6\,dq=0$, i.e. $p-3q=\text{constant}$, and is obviously satisfied by the initial conditions $U=2$, $p=-2$, $q=1$. It is easily verified that one solution satisfying these conditions is

$$U=2+(y-2x)+A(y-2x)^2,$$

where A is an arbitrary constant. This is unique along $y-2x=0$ but nowhere else in the x–y plane.

Propagation of discontinuities

It can be proved that the solutions of elliptic and parabolic equations are analytic even when the boundary or initial conditions are discontinuous. Hyperbolic equations however are different in that discontinuities in initial conditions are propagated as discontinuities into the solution domain along the characteristics.

Let Γ, Fig. 4.11, be a non-characteristic curve along which initial values for U, p, and q are known. Let P and Q be two distinct points on Γ and let the f characteristic through P meet the g characteristic through Q at R. Then the solution at R can be calculated in terms of the initial conditions at P and Q. Assuming no two characteristics of the same family intersect it follows that the solution at every point such as S inside the curvilinear triangle PQR is determined by the initial conditions between the points P and Q. Similarly, the solution at each point U inside the curvilinear strip $PRVT$ is determined by the initial conditions at a point along the arc TP (propagating along an f characteristic) and the initial conditions at a point along the arc PQ (propagated along a g characteristic). When the initial conditions along TP are analytically different from the initial condi-

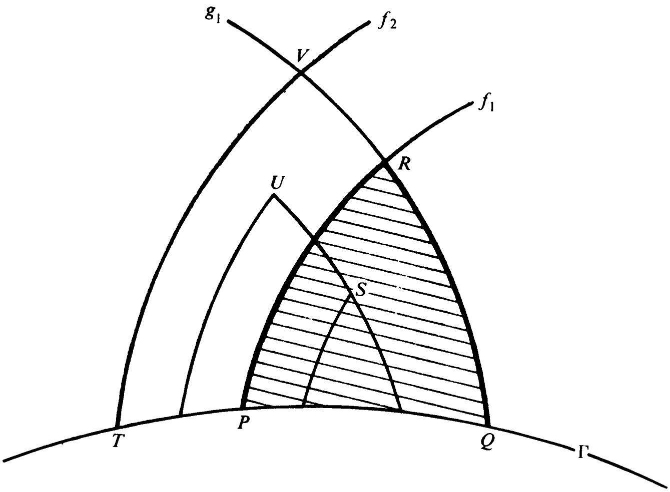


Fig. 4.11

tions along *PQ* then the solution inside the strip *PRVT* will be analytically different from the solution inside the curvilinear triangle *PQR*. As *T* tends to *P* the strip tends to the characteristic *PR* proving that the discontinuity in the initial conditions at *P* is propagated along a characteristic.

The argument above shows that a characteristic can separate two different solutions. *An extremely important feature of second-order equations is that these two solutions together with their first-order derivatives can be continuous across the dividing characteristic, but their second- and higher-order derivatives discontinuous across the same characteristic.* This is easily seen by recalling that the characteristic directions and the differential relationships in these directions were originally defined from eqn (4.35), namely,

$$s\left\{a\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)^2-b\left(\frac{\mathrm{d}y}{\mathrm{d}x}\right)+c\right\}-\left\{a\frac{\mathrm{d}p}{\mathrm{d}x}\frac{\mathrm{d}y}{\mathrm{d}x}+c\frac{\mathrm{d}q}{\mathrm{d}x}+e\frac{\mathrm{d}y}{\mathrm{d}x}\right\}=0,$$

by making the expressions inside both pairs of braces zero. Then $s=\partial^2U/\partial x\,\partial y$ is indeterminate and can be given an arbitrary value along a characteristic. Once, however, a value has been assigned to s the values of r and t are not indeterminate but can be calculated uniquely from any pair of eqns (4.32), (4.33), and