Sudan University of Science and Technology

College of Graduate Studies

Finite Difference Methods for Parabolic Equations

طرق الفروق المنتهية للمعادلات المكافئية

A thesis Submitted in Partial Fulfillment for the Degree of M.Sc. in Mathematics

By

Manal Salah Ahmed Elamein

Supervised by

Dr. Mohamed Hassan Mohamed Khabir

May 2015

الآية

قَالَ تَعَالَىٰ:

بسم الله الرحمن الرحيم

﴿ ٱقۡرَأۡ بِٱسۡمِ رَبِّكَ ٱلَّذِي خَلَقَ ١ خَلَقَ ٱلۡإِنسَٰنَ مِنۡ عَلَقٍ ٢ ٱقۡرَأۡ وَرَبُّكَ ٱلۡأَكۡرَمُ ٣
ٱلَّذِي عَلَّمَ بِٱلۡقَلَمِ ٤ عَلَّمَ ٱلۡإِنسَٰنَ مَا لَمۡ يَعۡلَمۡ ٥ ﴾

صدق الله العظيم

سورة العلق: ١ – ٥

Dedication

To whom we shared the stations in a career and I am live within odean fingerprints do not forget my friends and companions have trained, to my family, faithful friends and everyone who help me.

Acknowledgments

Praise God for your first and foremost to be with me to complete this research scientists praise you all thanks. Gratitude and appreciation to those who carried the message the most sacred in life especially **Dr. Mohamed Hassan Mohamed Khabir** the supervisor of this research, for his supervision, unlimited helps and remarks.

Also, I would like to send my thanks to fellows and friends for their assistance. My grateful thanks are for my family for enthusiastic and great help that they have offered me during this thesis.

Abstract

We present in this thesis the numerical solution to the partial differential equations of parabolic type using the finite-difference methods, namely explicit and Crank-Nicolson methods. We account the local truncation error of the two schemes by using Taylor series and discuss the consistency or compatibility, convergence and stability of these schemes for the parabolic equations. We present vector and matrix norms, also a necessary and sufficient condition for stability. Finally study the stability by the Fourier series method (von Neumann's method) and use lax's equivalence theorem to determine the consistency, stability and convergence.

الخلاصة

نقدم فى هذا البحث الحل العددي للمعادلة التفاضلية الجزئية من النوع المكافئ باستخدام طرق الفروق المنتهية، تحديداً طريقتي الصريحة وكرانك–نكولسون. ثم نحسب الخطأ لهاتين الطريقتين باستخدام متسلسلة تايلر كما نناقش التوائُم والتقارب والاستقرارية لكل طريقة للمعادلات المكافئة. بعد ذلك نقدم المعيار للمتجه والمصفوفة، وأيضاً الشرط الكافي والضروري للاستقرارية. أخيراً ندرس الاستقرارية بطريقة متسلسلة فورير (طريقة فون نيومان) ونستخدم نظرية لاكس للتكافؤ لتحديد التوائُم والاستقرارية والتقارب.

The contents

الآية.....I
Dedication.....II
Aknowledgments III
Abstract.....IV
الخلاصة.....V
The contents VI

Chapter One
Finite deference methods

1.1 Transformation to non-dimensional form 1
1.2 An explicit method of solution 2
1.3 Crank-Nicolson implicit method 9
1.4 Derivative boundary conditions 15

Chapter Two
The local truncation error and consistency

2.1The local truncation error 27
2.2 Consistency or Compatibility 29
2.3 Convergence and stability 32

Chapter Three
Stability Analysis

3.1 Introduction 37
3.2 A necessary and sufficient condition for stability (constants coefficients) 40
3.3 Stability by the Fourier series method (von Neuman's methods)..... 57
3.4 Lax's equivalence theorem 61
References 65

Chapter One

Finite difference methods

1.1 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 deal 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 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

$$\frac{\partial U}{\partial T} = \kappa \frac{\partial^2 U}{\partial X^2}, \quad \kappa \text{ constant}, \tag{1.1}$$

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

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

so equation (1.1) 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

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \tag{1.2}$$

as the non-dimensional form of (1.1).

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

1.2 An explicit method of solution

one finite-difference approximation to

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2} \tag{1.3}$$

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

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

and

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

this can be written as

$$u_{i,j+1} = ru_{i-1,j} + (1 - 2r)u_{i,j} + ru_{i+1,j}, \tag{1.4}$$

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 (Figure. (1.1)). 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 (1.1):

As a numerical example let us solve (1.4) given that the ends of the rod are kept in contact with blocks of melting ice and that the initial temperature distribution in non-dimensional form is

$$\begin{aligned} &a)\ U = 2x && 0 \le x \le \frac{1}{2}, \\ &b)\ U = 2(1 - x), && \frac{1}{2} \le x \le 1. \end{aligned} \tag{1.5}$$

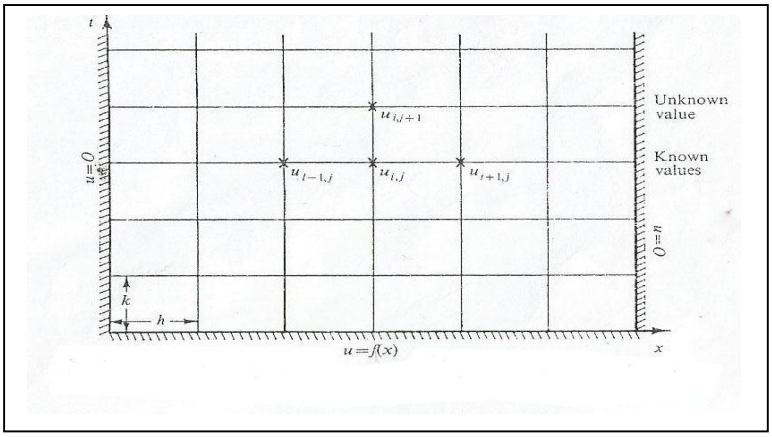


Figure (1.1)

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) $U(x, t = 0) = \begin{cases} 2x & for \;\; 0 \le x \le \frac{1}{2} \\ 2(1-x) & for \;\; \frac{1}{2} \le x \le 1 \end{cases}$

(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 (1.1). The problem is symmetric with respect to $x = \frac{1}{2}$ so we need the solution only for $0 \le x \le \frac{1}{2}$.

Case (1):

Take $\delta x = h = \frac{1}{10}$, $\delta t = k = \frac{1}{1000}$, so $r = k/h^2 = \frac{1}{10}$. Equation (1.4) then reads as

$$u_{i,j+1} = \frac{1}{10}\left(u_{i-1,j} + 8u_{i,j} + u_{i+1,j}\right). \tag{1.6}$$

For pencil and paper calculations the relationship between these four function values is represented very conveniently by the 'molecule' in figure (1.2). The numbers in the 'atoms' are the multipliers of the function values at the corresponding mesh points.

Application of equation (1.6) is shown in Table (1.1), remembering that the values of U at $x = \frac{4}{10}$ and $\frac{6}{10}$ are equal

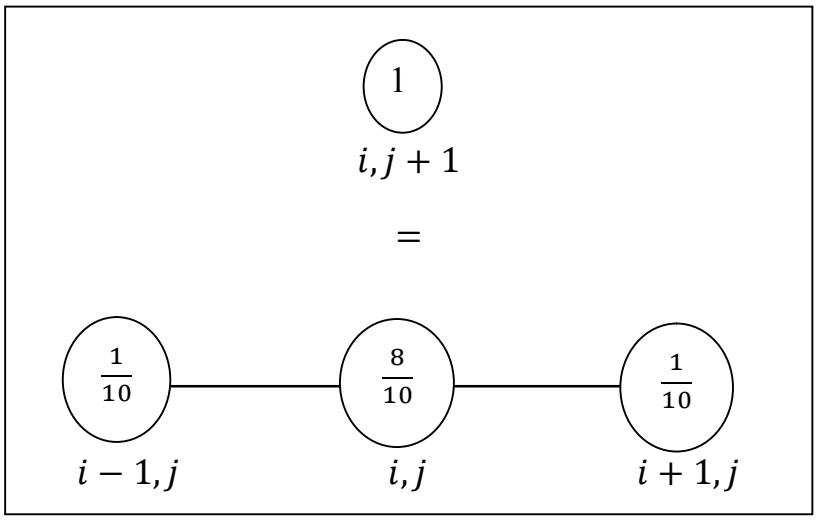


Figure (1.2)

because of symmetry. (Increasing values of t, i.e. of j, are shown moving downwards for convenience of calculation.) As examples,

$$u_{5,1} = \frac{1}{10}\{0.8 + (8 \times 1) + 0.8\} = 0.9600.$$
$$u_{4,2} = \frac{1}{10}\{0.6 + (8 \times 0.8) + 0.96\} = 0.7960.$$

The analytical solution of the partial differential equation satisfying these conditions is [4]

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

Table (1.1)

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

Table (1.2)

	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 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 (equation (1.5)). Inspection of Table (1.3) 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.

Richtmyer, [3] 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 ordinarily discontinuous.

Table (1.3)

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 (1.4)

	$i = 0$	1	2	3	4	5	6
	$x = 0$	0.1	0.2	0.3	0.4	0.5	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

(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 [4].

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 estimate indicates. 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 $r = \frac{k}{h^2} = 0.5$. Then equation (1.4) gives

$$u_{i,j+1} = \frac{1}{2}\left(u_{i-1,j} + u_{i+1,j}\right), \tag{1.7}$$

and the solution obtained by applying this finite-difference equation to the boundary and initial values is recorded in Table (1.4).

Table (1.5)

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 equation (1.4) gives

$$u_{i,j+1} = u_{i-1,j} - u_{i,j} + u_{i+1,j}, \tag{1.8}$$

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

Table (1.6)

	$i=0$	1	2	3	4	5	6
	$x=0$	0.1	0.2	0.3	0.4	0.5	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 equation (1.8) with respect to the initial values and boundary values given.

These three cases clearly indicate that the values of r is important and it will be proved later that this explicit method is valid only for $0 < r \leq \frac{1}{2}$.

1.3 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 $0 < k/h^2 \leq \frac{1}{2}$, i.e. $k \leq \frac{1}{2}h^2$, 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 replace $\partial^2 U/\partial x^2$ 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

$$-ru_{i-1,j+1} + (2+2r)u_{i,j+1} - ru_{i+1,j+1}$$
$$= ru_{i-1,j} + (2-2r)u_{i,j} + ru_{i+1,j}, \qquad (1.9)$$

where $r = k/h^2$.

In general, the left side of equation (1.9) contains three unknown, and the right side three known pivotal values of u figure (1.3).

If there are N internal mesh points along each time row then for $j = 0$ and $i = 1,2, \dots, N$, equation (1.9) gives N simultaneous equations for the

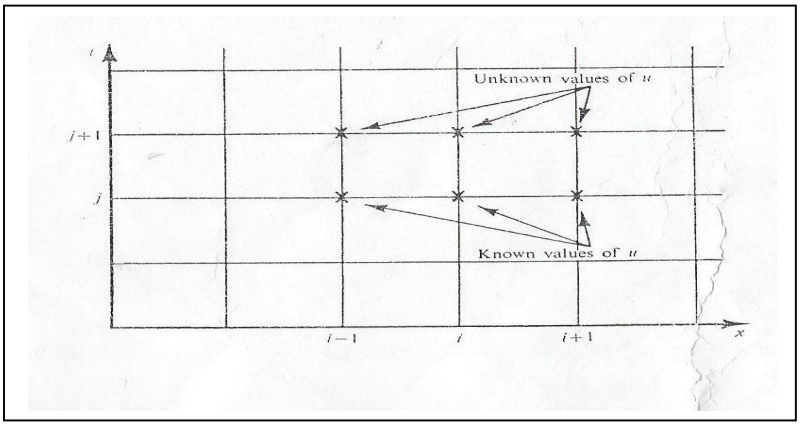


Figure (1.3)

N unknown pivotal values along the first 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 unknown pivotal value necessitates the solution of a set of simultaneous equations is described as an implicit one.

The Crank-Nicolson method approximates the partial differential equation at the point $\{ih, (j+\frac{1}{2})k\}$ by

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

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 (1.2):

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

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

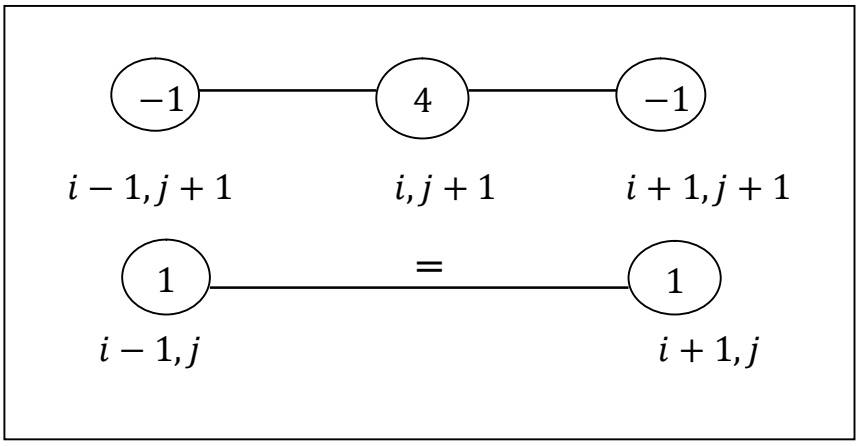
where

i) $U = 0, x = 0$ and $1, t \geq 0$,
ii) $U = 2x, 0 \leq x \leq \frac{1}{2}, t = 0$,
iii) $U = 2(1 - x), \frac{1}{2} \leq x \leq 1, t = 0$,

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 (1.9). Then $k = \frac{1}{100}$ and (1.9) read as

$$-u_{i-1,j+1} + 4u_{i,j+1} - u_{i+1,j+1} = u_{i-1,j} + u_{i+1,j} \tag{1.10}$$

The computational molecule corresponding to equation (1.10) is shown in figure (1.4). 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. (figure (1.5)).



The values of u for the first time step then satisfy

$$-0 + 4u_1 - u_2 = 0 + 0.4,$$

$$-u_1 + 4u_2 - u_3 = 0.2 + 0.6,$$

$$-u_2 + 4u_3 - u_4 = 0.4 + 0.8,$$

$$-u_3 + 4u_4 - u_5 = 0.6 + 1.0,$$

$$-2u_4 + 4u_5 \quad = 0.8 + 0.8.$$

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

$$-0 + 4u_1 - u_2 = 0 + 0.3956,$$

$$-u_1 + 4u_2 - u_3 = 0.1989 + 0.5834,$$

$$-u_2 + 4u_3 - u_4 = 0.3956 + 0.7381,$$

$$-u_3 + 4u_4 - u_5 = 0.5834 + 0.7691,$$

$$-2u_4 + 4u_5 = 2 \times 0.7381$$

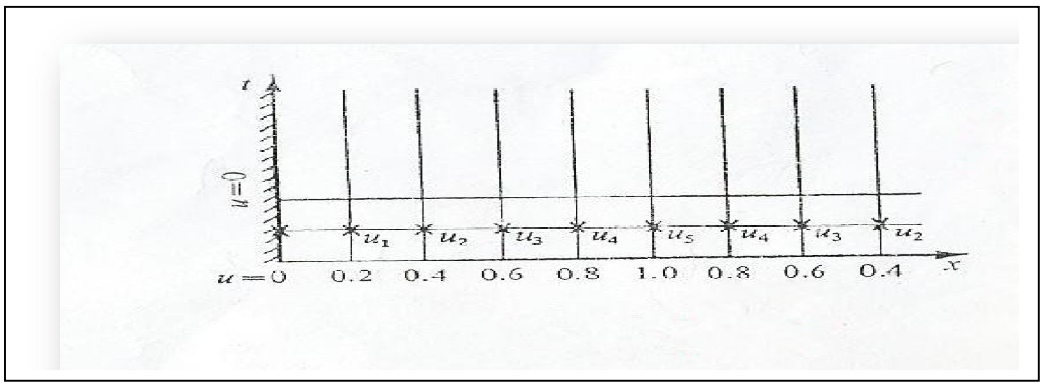


Figure (1.5)

Table (1.7)

	$x = 0$	0.1	0.2	0.3	0.4	0.5
$t = 0.00$	0	0.2	0.4	0.6	0.8	1.0
0.01	0	0.1989	0.3956	0.5834	0.7381	0.7691
0.02	0	0.1936	0.3789	0.5400	0.6461	0.6921
	0					
$\vdots$	$\vdots$					
$t = 0.10$	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 (1.7) .

Table (1.8) below displays both solutions at $x = 0.5$ for various values of t. A glance at Table (1.3) 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.

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 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.

Table (1.8)

	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

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

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

where, in practice, $0 \le \theta \le 1$.This replacement approximates the partial differential equation at the point $\left\{i\delta x, \left(j + \frac{1}{2}\right)\delta t\right\}$ 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 $\frac{1}{2} \le \theta \le 1$, but for $0 \le \theta < \frac{1}{2}$ we must have

$$r = \frac{\delta t}{(\delta x)^2} \le \frac{1}{2(1-2\theta)}. \quad (See[\,4\,], Chapter\ 2.)$$

1.4 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

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

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, 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 $\left(-\,\delta x, j\delta t\right)$ (Figure 1.6), by imagining the rod to be extended a distance δx at this end. The boundary condition can then be represented by

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

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

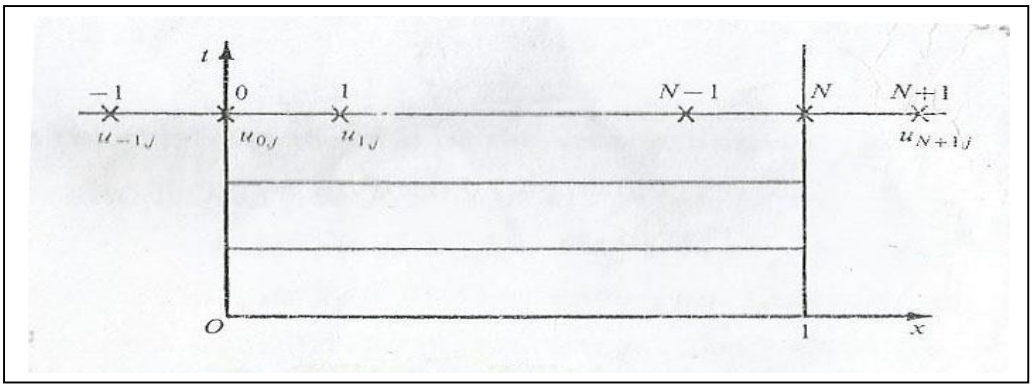


Figure (1.6)

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 (1.3):

Solve the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \tag{1.11}$$

Satisfying the initial condition,

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

and the boundary conditions,

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

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

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

One explicit finite-difference representation of equation (1.11) 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.

$$u_{i,j+1} = u_{i,j} + r\left(u_{i-1,j} - 2u_{i,j} + u_{i+1,j}\right), \tag{1.12}$$

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

At $x = 0$,

$$u_{0,j+1} = u_{0,j} + r\left(u_{-1,j} - 2u_{0,j} + u_{1,j}\right). \tag{1.13}$$

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

$$\frac{u_{1,j} - u_{-1,j}}{2\delta x} = u_{0,j}. \tag{1.14}$$

Eliminating $u_{-1,j}$ between (1.13) and (1.14) gives

$$u_{0,j+1} = u_{0,j} + 2r\{u_{1,j} - (1 + \delta x)u_{0,j}\}. \tag{1.15}$$

Let $\delta x = 0.1$. Then at $x = 1$, equation (1.12) becomes

$$u_{10,j+1} = u_{10,j} + r(u_{9,j} - 2u_{10,j} + u_{11,j}), \tag{1.16}$$

and the boundary condition is

$$\frac{u_{11,j} - u_{9,j}}{2\delta x} = -u_{10,j}. \tag{1.17}$$

Elimination of the 'fictitious' value $u_{11,j}$ between (1.16) and (1.17) yields

$$u_{10,j+1} = u_{10,j} + 2r\{u_{9,j} - (1 + \delta x)u_{10,j}\}. \tag{1.18}$$

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 $r \leq 1/(2 + h\delta x)$, i.e. $r \leq 1/2.1$ in this example.

Choose $r = \frac{1}{4}$. The difference equations (1.15), (1.12), then become

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

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

and the use of symmetry rather than equation (1.18) gives

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

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

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

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

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

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

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

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

Similarly for subsequent time-steps. The values for several steps are recorded in Table (1.9).

The analytical solution of the partial differential equation satisfying these boundary and initial conditions is [4]

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

where α_n are the positive roots of

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

Table (1.9)

	$i =$	1	2	3	4	5
	$x = 0$	0.1	0.2	0.3	0.4	0.5
$t = 0.000$	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

Values of U calculated from this analytical solution are recorded in Table (1.10).

The two solutions are compared at $x = 0.2$ in Table (1.11).

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 (1.4):

Re-solve the Worked Example (1.3) using an explicit method and employing a forward-difference for the boundary condition at $x = 0$.

Table (1.10)

$x =$	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 (1.11)

	Finite-difference Solution $(x = 0.2)$	Analytical solution $(x = 0.2)$	Percentage error
$t = 0.005$	1.0000	0.9984	0.16
0.0050	0.9126	0.9120	0.07
0.0100	0.8345	0.8342	0.04
0.0250	0.6452	0.6454	−0.03
0.0500	0.4205	0.4212	−0.16
1.000	0.1786	0.1794	−0.45

By equation (1.12), one explicit finite-difference representation of the partial differential equation is

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

Hence, for $i = 1$,

$$u_{1,j+1} = u_{1,j} + r\left(u_{0,j} - 2u_{1,j} + u_{2,j}\right). \tag{1.19}$$

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

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

or

$$u_{0,j} = \frac{u_{1,j}}{1 + \delta x}. \tag{1.20}$$

Eliminating $u_{0,j}$ between (1.19) and (1.20) gives

$$u_{1,j+1} = \left(1 - 2r + \frac{r}{1 + \delta x}\right) u_{1,j} + r u_{2,j}. \tag{1.21}$$

This scheme is valid for $0 \le r \le \frac{1}{2}$ ([4],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

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

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

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

and

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

The solution of these equations for an initial value of $U = 1$ is shown in Table (1.12). A comparison with the analytical solution at $x = 0.2$ is given in Table (1.13).

Table (1.12)

$x =$	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 (1.13)

	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 (1.5):

Solve Example (1.3) 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

$$-ru_{i-1,j+1} + (2+2r)u_{i,j+1} - ru_{i+1,j+1} = ru_{i-1,j} + (2-2r)u_{i,j} + ru_{i+1,j}. \quad (1.22)$$

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 (1.22).

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

shows that the equations for the unknown pivotal values $u_{0,j+1}, u_{1,j+1}, \dots, u_{5,j+1}$ 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

$$2.1u_0 - u_1 = 0.9,$$

$$-u_0 + 4u_1 - u_2 = 2.0,$$

$$-u_1 + 4u_2 - u_3 = 2.0,$$

$$-u_2 + 4u_3 - u_4 = 2.0,$$

$$-u_3 + 4u_4 - u_5 = 2.0,$$

$$-u_4 + 2u_5 = 1.0.$$

These can be solved by the direct elimination method . Table (1.14) records the solution for several time-steps and Table (1.15) compares it with the analytical solution at $x = 0.2$.

Table (1.14)

$i =$	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 (1.15)

	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

Chapter Two

The local truncation error and consistency

2.1 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 (2.1):

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

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

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

$$U_{i,j+1} = U(x_i, t_j + k) = U_{i,j} + k\left(\frac{\partial U}{\partial t}\right)_{i,j} + \frac{1}{2}k^2\left(\frac{\partial^2 U}{\partial t^2}\right)_{i,j} + \frac{1}{6}k^3\left(\frac{\partial^3 U}{\partial t^3}\right)_{i,j} + \cdots$$

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

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

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

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

Hence

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

When $k = rh^2, 0 < r \leq \frac{1}{2}, T_{i,j}$ is $O(k)$ or $O(h^2)$.

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

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

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.

2.2 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 next example (2.2).

The real importance of the concept of consistency lies in a theorem by Lax [3], 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 definitions $T_{i,j}$ 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 (2.1) 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 (2.2):

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

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

We also 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, 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

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

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

Case (i): $k = rh$

As $h \to 0$,

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

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): $k = rh^2$

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

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) in [4]).

2.3 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 unboundedness 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.

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.

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-differential equation by u. Then $e = U - u$.

Consider the equation

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1, t > 0, \tag{2.1}$$

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

The simplest explicit finite-difference approximation to (2.1) is

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2}. \tag{2.2}$$

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 (2.2) leads to

$$\begin{aligned} e_{i,j+1} = re_{i-1,j} + (1-2r)e_{i,j}, + re_{i+1,j} + U_{i,j+1} - U_{i,j} \\ + r\left(2U_{i,j} - U_{i-1,j} - U_{i+1,j}\right). \end{aligned} \tag{2.3}$$

By Taylor's theorem,

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

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

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

where $0 < \theta_1 < 1,\ 0 < \theta_2 < 1$ and $0 < \theta_3 < 1$. Substitution into equation (2.3) gives

$$e_{i,j+1} = re_{i-1,j} + (1-2r)e_{i,j}, + re_{i+1,j} + k\left\{\frac{\partial U}{\partial t}(x_i, t_j + \theta_3 k) - \frac{\partial^2 U}{\partial x^2}(x_i + \theta_4 h, t_j)\right\}. \tag{2.4}$$

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 modules of the expression in the braces for all i and j. When $r \leq \frac{1}{2}$, all the coefficients of e in equation (2.4) are positive or zero, so

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

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

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

etc, from which it follows that

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

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 equation (2.1) the limiting value of M and therefore of E_j is zero. As $|U_{i,j} - u_{i,j}| \leq E_j$, this proves that u converges to U as h tends to zero when $r \leq \frac{1}{2}$ and t is finite.

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 $k, 2k, \dots, Jk = T$, 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 $(i, j), i = 0(1)N, 0 < j \leq J$.

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 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 $k \to 0$, i.e. as $J \to \infty$, it being assumed that $\delta x = h$ is related to k in such a way that $h \to 0$ as $k \to 0$.

Assume that the vector of solution values $\mathbf{u}_{j+1} = \left[u_{1,j+1}, u_{2,j+1}, \dots, u_{N-1,j+1}\right]^T$ 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

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

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

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

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

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

when the solution of the partial differential equation 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 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 eignvalues of matrix $\mathbf{A}$, i.e. the spectral radius $\rho(\mathbf{A})$ of $\boldsymbol{A}$, satisfies

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

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 eignvalues of $\mathbf{A}$ are restricted to satisfy $\rho(\mathbf{A}) \leq \|\mathbf{A}\| \leq 1$, as $N \to \infty$. 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 [1].

Chapter Three

Stability Analysis

3.1 Introduction

Vector and matrix norms:

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

Vector norms:

The norm of the vector $\mathbf{x}$ 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) $\|\mathbf{x} + \mathbf{y}\| \leq \|\mathbf{x}\| + \|\mathbf{y}\|$.

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

The 1-norm of $\mathbf{x}$ is the sum of the moduli of the components of $\mathbf{x}$, i.e.

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

The infinity norm of $\mathbf{x}$ is the maximum of the moduli of the components of $\mathbf{x}$, i.e.

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

The 2-norm of $\mathbf{x}$ is the square root of the sum of the squares of the moduli of the components of $\mathbf{x}$, 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) $\|\mathbf{A} + \mathbf{B}\| \leq \|\mathbf{A}\| + \|\mathbf{B}\|$.
(iv) $\|\mathbf{AB}\| \leq \|\mathbf{A}\|\|\mathbf{B}\|$

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

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

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

$$\|\mathbf{A}\| = \|\mathbf{Ax_0}\| = \max_{\|\mathbf{x}\|=1} \|\mathbf{Ax}\|$$

The 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, $\|\mathbf{Ax}_1\| \leq \|\mathbf{Ax_0}\| = \|\mathbf{A}\| = \|\mathbf{A}\|\|\mathbf{x}_1\|$ since $\|\mathbf{x}_1\| = 1$. It also follows that for all subordinate matrix norms,

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

The definitions of the 1,2, and ∞ norms with $\|\mathbf{x}\| = 1$ lead to the following results which are provided in most linear algebra books.

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 $\mathbf{A}^H = (\overline{\mathbf{A}})^T$, the transpose of the conjugate complex of $\mathbf{A}$. For complex, 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

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

Therefore,

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

Hence,

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

3.2 A necessary and sufficient condition for stability (Constants coefficients):

Let the solution domain of the partial differential equation be the finite rectangle $0 \le x \le 1, 0 \le t \le T$, 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

$$b_{i-1}u_{i-1,j+1} + b_i u_{i,j+1} + b_{i+1}u_{i+1,j+1} = c_{i-1}u_{i-1,j} + c_i u_{i,j} + c_{i+1}u_{i+1,j}$$

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

$$\begin{bmatrix} b_1 & b_2 & & & & \\ b_1 & b_2 & b_3 & & & \\ \cdot & & \cdot & \cdot & & \\ & \cdot & \cdot & \cdot & & \\ & & \cdot & b_{N-3} & b_{N-2} & b_{N-1} \\ & & & & b_{N-2} & b_{N-1} \end{bmatrix} \begin{bmatrix} u_{1,j+1} \\ u_{2,j+1} \\ \vdots \\ u_{N-2,j+1} \\ u_{N-1,j+1} \end{bmatrix} = \begin{bmatrix} c_1 & c_2 & & & \\ c_1 & c_2 & c_3 & & \\ \cdot & \cdot & \cdot & \cdot & \\ & \cdot & c_{N-3} & c_{N-2} & c_{N-1} \\ & & & c_{N-2} & c_{N-1} \end{bmatrix} \begin{bmatrix} u_{1,j} \\ u_{2,j} \\ \vdots \\ u_{N-2,j} \\ u_{N-1,j} \end{bmatrix} + \begin{bmatrix} c_0 u_{0,j} - b_0 u_{0,j+1} \\ 0 \\ \vdots \\ 0 \\ c_N u_{N,j} - b_N u_{N,j+1} \end{bmatrix}$$

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

$$
\begin{aligned}
u_j &= \mathbf{A}u_{j-1} + f_{j-1} = \mathbf{A}\left(\mathbf{A}\mathbf{u}_{j-2} + \mathbf{f}_{j-2}\right) + \mathbf{f}_{j-1} \\
&= \mathbf{A}^2\mathbf{u}_{j-2} + \mathbf{A}\mathbf{f}_{j-2} + \mathbf{f}_{j-1} \\
&= \cdots \\
&= \mathbf{A}^j\mathbf{u}_0 + \mathbf{A}^{j-1}\mathbf{f}_0 + \mathbf{A}^{j-2}\mathbf{f}_1 + \cdots + \mathbf{f}_{j-1}
\end{aligned} \tag{3.1}
$$

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 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

$$
\mathbf{u}_j^* = \mathbf{A}^j\mathbf{u}_0^* + \mathbf{A}^{j-1}\mathbf{f}_0 + \mathbf{A}^{j-2}\mathbf{f}_1 + \cdots + \mathbf{f}_{j-1}. \tag{3.2}
$$

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

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

it follows by equation (3.1) and (3.2) that

$$
\mathbf{e}_j = \mathbf{u}_j^* - \mathbf{u}_j = \mathbf{A}^j(\mathbf{u}_0^* - \mathbf{u}_0) = \mathbf{A}^j\mathbf{e}_0, \quad j = 1(1)J. \tag{3.3}
$$

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

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

Hence, for compatible matrix and vector norms,

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

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

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

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

$$\|\mathbf{e}_j\| \leq M\|\mathbf{e}_0\|.$$

In reference [3], this definition of stability is related to convergence via Lax's equivalence theorem .

Since

$$\|\mathbf{A}^j\| = \|\mathbf{A}\mathbf{A}^{j-1}\| \leq \|\mathbf{A}\|\|\mathbf{A}^{j-1}\| \leq \cdots \leq \|\mathbf{A}\|^j.$$

It follows that the Lax-Richtmyer definition of stability is satisfied by

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

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 $\rho(\mathbf{A}) \leq 1$ since $\rho(\mathbf{A}) \leq \|\mathbf{A}\|$.

If, however, $\rho(\mathbf{A}) \leq 1$,it does not follow that $\|\mathbf{A}\| \leq 1$. 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

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

Example (3.1):

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 $1 - 2r \geq 0$, then $0 < r \leq \frac{1}{2}$

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 $0 < r \leq \frac{1}{2}$.

Alternatively, since matrix **A** is real and symmetric,

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

where μ_s is the sth eigenvalue of **A.** Now **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

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

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

Therefore the equations will be stable when

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

i.e.,

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

The left-hand inequality gives that

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

As $h \to 0, N \to \infty$ and $\sin^2(N-1)\pi/2N \to 1$.

Hence $r \leq \frac{1}{2}$.

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

Example (3.2):

The Crank-Nicolson equations (1.9) 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

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

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

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

from which it follows that matrix $\mathbf{A}$ of equation (3.1) is

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

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 $(2 + 4r\sin^2 s\pi/2N)^{-1}(2 - 4r\sin^2 s\pi/2N)$. [4]

Therefore

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

proving that the Crank-Nicolson equations are unconditionally stable. They are also consistent,[4], so they are also convergent.

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,

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

i.e.

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

As shown by equation (3.3), 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,

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

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

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

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

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

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

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

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

Where $\lambda_s = 1 - 4r\sin^2 s\pi/2N$.

Therefore any perturbation, rounding errors and $u_{i,j}$ will be bounded as j increases if

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

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

A note on eigenvalues and eigenvectors:

Let $\mathbf{x}$ be an eigenvector of the matrix $\mathbf{A}$ corresponding to the eigenvalue λ. Then $\mathbf{Ax} = \lambda\mathbf{x}$. Hence $\mathbf{A}(\mathbf{Ax}) = \mathbf{A}^2\mathbf{x} = \lambda\mathbf{Ax} = \lambda^2\mathbf{x}$, 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 $f(\mathbf{A}) = a_p\mathbf{A}^p + a_{p-1}\mathbf{A}^{p-1} + \cdots + a_0\mathrm{I}$ is a polynomial in $\mathbf{A}$ with scalar coefficients $a_p, \ldots, a_0$, then $f(\mathbf{A})\mathbf{x} = \left(a_p\lambda^p + \cdots + a_0\right)\mathbf{x} = f(\lambda)\mathbf{x}$, 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),

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

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 in [4].

Useful theorems on bounds for eigenvalues:

Gerschgorin's first theorem:

The largest of the moduli of the eigenvalues of the square matrix $\mathbf{A}$ cannot exceed the largest sum of the moduli of the elements along any row or any column. In other words

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

Proof:

Let λ_i be an eigenvalues of the $N \times N$ matrix $\mathbf{A}$, and $\mathbf{x}_i$ the corresponding eigenvectors 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

$$
\begin{array}{c}
a_{1,1}v_1 + a_{1,2}v_2 + \cdots + a_{1,n}v_n = \lambda_i v_1, \\
a_{2,1}v_1 + a_{2,2}v_2 + \cdots + a_{2,n}v_n = \lambda_i v_2, \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots \\
a_{s,1}v_1 + a_{s,2}v_2 + \cdots + a_{s,n}v_n = \lambda_i v_s, \\
\vdots \qquad \vdots \qquad \vdots \qquad \vdots
\end{array}
$$

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

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

Therefore

$$|\lambda_i| \leq |a_{s,1}| + |a_{s,2}| + \cdots + |a_{s,n}|,$$

because

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

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 **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,

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

Hence

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

which completes the proof.

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

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

which can be written as

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

giving

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

where

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

The equations will be stable in the Lax-Richtmyer sense when $\|4\mathbf{B}^{-1} - \mathbf{I}_{N-1}\| \leq 1$. Since $\mathbf{B}$ is real and symmetric it follows that $4\,\mathbf{B}^{-1} - \mathbf{I}_{N-1}$ is real and symmetric, so $\|4\mathbf{B}^{-1} - \mathbf{I}_{N-1}\| = \rho(4\mathbf{B}^{-1} - \mathbf{I}_{N-1})$ [4]. 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

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

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

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

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

from which it follows that

$$-2r \leq \lambda - 2 - 2r \leq 2r,$$

or

$$2 \leq \lambda \leq 2 + 4r$$

Proving that the equations are unconditionally stable, since $\lambda \geq 2$ 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 $|\lambda_i| \leq 1$ is equivalent to $\|\mathbf{A}\|_\infty \leq 1$ or $\|\mathbf{A}\|_1 \leq 1$. The theorem states that

$$|\lambda - a_{s,s}| \leq P_s,$$

Hence

$$-P_s \leq \lambda - a_{s,s} \leq P_s,$$

so that

$$-P_s + a_{ss} \leq \lambda \leq P_s + a_{s,s}.$$

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

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

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

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

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

When both inequalities hold, it follows automatically that $\|\mathbf{A}\|_2 \leq 1$ because $\|\mathbf{A}\|_2^2 \leq \|\mathbf{A}\|_1 \|\mathbf{A}\|_\infty$. 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's circle contains one eigenvalue.

A proof is given in [5].

Stability criteria for derivative boundary conditions:

Consider the equation

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

and the conditions,

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

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

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

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

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

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

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

$$\begin{bmatrix} u_{0,j+1} \\ u_{1,j+1} \\ u_{N-1,j+1} \\ u_{N,j+1} \end{bmatrix}$$

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

$$\times \begin{bmatrix} u_{0,j} \\ u_{1,j} \\ \\ u_{N-1,j} \\ u_{N,j} \end{bmatrix} + \begin{bmatrix} 2rh_1v_1\delta x \\ 0 \\ \\ 0 \\ 2rh_2v_2\delta x \end{bmatrix}$$

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

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

Since the off-diagonal elements of this real matrix are one-signed, all its eigenvalues are real. 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

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

Using Figure (3.1),

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

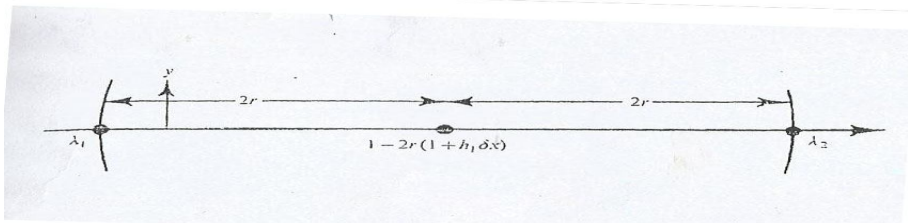


Figure (3.1)

and for stability,

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

Hence

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

and

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

The least of these is $r \leq 1/(2 + h_1\delta x)$.

Similarly, we require $r \leq 1/(2 + h_2\delta x)$.

For rows $2(1)N, a_{s,s} = 1 - 2r, P_s = 2r$, giving $r \leq \frac{1}{2}$.

For overall stability, $r \leq \min\{1/(2 + h_1\delta x); 1/(2 + h_2\delta x)\}$.

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

As mentioned earlier, the circle theorem and $|\lambda| \leq 1$ is equivalent to $\|\mathbf{A}\|_\infty \leq 1$. This is easily verified for this example.

Row 1: if $1 - 2r(1 + h_1\delta x) \geq 0$, the sum of the moduli of the terms along row $1 = 1 - 2rh_1\delta x$.

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

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

If $1 - 2r(1 + h_1\delta x) \leq 0$, 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, $0 < 2r(2 + h_1\delta x) - 1 \leq 1$ implies that $r \leq 1/(2 + h_1\delta x)$. Similarly for 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 $4\mathbf{B}^{-1} - \mathbf{I} \leq 1$ if the eigenvalues of $\mathbf{B} \geq 2$. By the circle theorem,

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

giving

$$2 + 2rh_1\delta x \leq \lambda$$

Similarly, the remaining rows give that $2 \leq \lambda$ and $2 + 2rh_2\delta x \leq \lambda$. 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

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

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

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

Assuming $M > 1$,

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

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

In particular, this must satisfy

$$\|\mathbf{A}\| \leq 1 + \frac{1}{J} \ln M + \cdots$$

$$\geq 1 + \frac{1}{J} \ln M$$

$$= 1 + \frac{k}{T} \ln M.$$

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.

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

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 $\delta x = h \to 0$ and $\delta t = k \to 0$, i.e. as $j \to \infty$. The Fourier series or von Neumann method, first discussed in detail in [2], 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 $u(ph, kq) = u_{p,q}$. 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

$$u_{p,0} = \sum_{n=0}^{N} A_n e^{i\beta_n ph}, \quad p = 0,1, \dots, N, \qquad (3.4)$$

are sufficient to determine the $(N + 1)$ unknowns $A_0, A_1, \dots, A_N$ 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,we put

$$u_{p,q} = \mathrm{e}^{i\beta x}\mathrm{e}^{\alpha t} = \mathrm{e}^{i\beta ph}\mathrm{e}^{\alpha qk} = \mathrm{e}^{i\beta ph}\xi^q \qquad (3.5)$$

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 $q \leq J$ 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

$$|\xi| \leq 1,$$

i.e.

$$-1 \leq \xi \leq 1. \tag{3.6}$$

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

$$|\xi| \leq 1 + Kk = 1 + O(k), \tag{3.7}$$

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 condition (3.6) and (3.7) are always necessary but may not be sufficient. In practice the method often gives useful results even when its application is not fully justified.

Example (3.3):

Investigate the stability of the fully implicit difference equation

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

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

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

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

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

$$\xi - 1 = r\xi\left(\mathrm{e}^{-i\beta h} - 2 + \mathrm{e}^{i\beta h}\right)$$

$$= r\xi(2\cos\beta h - 2)$$

$$= -4r\xi\sin^2(\beta h/2).$$

Hence,

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

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

Example (3.4):

Investigate the stability of the linear difference equation

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

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

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

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

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

This leads to

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

Hence,

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

Satisfies the von Neumann condition (3.7) if

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

This leads to $r \leq 1/2a$, 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 (3.5):

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

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

Investigate its stability.

It is easily shown by the procedure of Example (1.9) that the equation for ξ is

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

where

$$A = 1 - 2r^2\sin^2(\beta h/2) \quad r = k/h. \tag{3.8}$$

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

$$|\xi| \leq 1.$$

As r, k and β are real, $A \leq 1$ by (3.8).

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

When

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

hence

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

showing that a necessary condition for stability is $-1 \leq A \leq 1$. By (3.8),

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

The only useful inequality is

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

giving

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

3.4 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.

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 arbitrary 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

$$\frac{\partial U}{\partial t} = \frac{\partial^2 U}{\partial x^2}, \quad 0 < x < 1, \quad t > 0, \tag{3.9}$$

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

The classical explicit approximation to (3.9) is

$$\frac{u_{i,j+1} - u_{i,j}}{k} = \frac{u_{i-1,j} - 2u_{i,j} + u_{i+1,j}}{h^2}, \quad i = 1(1)(N-1), \tag{3.10}$$

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

$$U_{i,j+1} = kT_{i,j} + rU_{i-1,j} + (1-2r)U_{i,j} + rU_{i+1,j}, \tag{3.11}$$

where $r = k/h^2$. For $i = 1(1)(N-1)$, equations (3.11) can be written in matrix form as

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

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

i.e. as

$$\mathbf{U}_{j+1} = k\mathbf{T}_j + \mathbf{A}\mathbf{U}_j + \mathbf{c}_j, \tag{3.12}$$

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

$$\begin{aligned} U_{j+1} &= kT_j + \mathbf{A}\big(k\mathbf{T}_{j-1} + \mathbf{A}\mathbf{U}_{j-1} + \mathbf{c}_{j-1}\big) + \mathbf{c}_j \\ &= k\big(\mathbf{T}_j + \mathbf{A}\mathbf{T}_{j-1}\big) + \mathbf{A}^2\mathbf{U}_{j-1} + \big(\mathbf{c}_j + \mathbf{A}\mathbf{C}_{j-1}\big) \\ &= \cdots \\ &= k\big(\mathbf{T}_j + \mathbf{A}\mathbf{T}_{j-1} + \cdots + \mathbf{A}^j\mathbf{T}_0\big) + \mathbf{A}^{j+1}\mathbf{U}_0 \\ &\quad + \big(\mathbf{c}_j + \mathbf{A}\mathbf{c}_{j-1} + \cdots + \mathbf{A}^j\mathbf{c}_0\big). \end{aligned} \tag{3.13}$$

As the boundary and initial values for u are the same as for U, it follows from equation (3.10), which can be written as

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

that

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

This leads as before to

$$\mathbf{u}_{j+1} = \mathbf{A}^{j+1}\mathbf{u}_0 + \big(c_j + \mathbf{A}\mathbf{c}_{j-1} + \cdots + \mathbf{A}^j\mathbf{c}_0\big). \tag{3.14}$$

Subtraction of (3.14) from (3.13) shows that

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

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

$$\mathbf{U}_{j+1} - \mathbf{u}_{j+1} = k\big(\mathbf{T}_j + \mathbf{A}\mathbf{T}_{j-1} + \cdots + \mathbf{A}^j\mathbf{T}_0\big). \tag{3.15}$$

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 $0 \leq x \leq 1, 0 \leq t = jk \leq T = Jk, T$ finite.

Hence equation (3.15) gives

$$\left\|\mathbf{U}_{j+1} - \mathbf{u}_{j+1}\right\| \leq k\left\{\left\|T_j\right\| + \|\mathbf{A}\|\left\|\mathbf{T}_{j-1}\right\| + \cdots + \left\|\mathbf{A}^j\right\|\|\mathbf{T}_0\|\right\}, \ \ 0 \leq j \leq J - 1.$$

By the Lax-Richtmyer definition of stability,

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

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

$$\left\|\mathbf{U}_{j+1} - \mathbf{u}_{j+1}\right\| \leq k\left\|\mathbf{T}_j\right\| + kM\left(\left\|\mathbf{T}_{j-1}\right\| + \cdots + \|\mathbf{T}_0\|\right).$$

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 $\leq C$. Using this norm it follows that

$$\left\|\mathbf{U}_{j+1} - \mathbf{u}_{j+1}\right\| \leq kM + kMj\mathrm{C} = kM + MtC.$$

But $jk = t \leq T = Jk$ is finite and $k \to 0$ as $J \to \infty$, 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 $C = \max_{i,j}\left|T_{i,j}\right|$ 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.

References

[1] K. W. Morton, (1980), Stability of finite-difference approximations to a diffusion-convection equation. Int. J. Num. Meth. Eng. 12, 899-916.

[2] C. G. O'Brien, M. A. Hyman, and S. Kaplan, (1951), Astudy of the numerical solution of partial differential equations. J. Math. Phys. 29, 223-251.

[3] R. D. Richtmyer, and K. W. Morton, (1967), Difference methods for initial-value problems, Interscience publishers, New York.

[4] G. D. SMITH, (1985), Numerical Solution Of Partial Differential Equations: Finite Difference Methods, Third Edition Oxford University, New York.

[5] J. H. Wilkinson, (1965), The algebraic eigenvalue problem, Claren-don press, Oxford.