THE STABILITY OF FINITE DIFFERENCE SCHEMES

FOR THE SOLUTION OF

DIFFERENTIAL EQUATIONS

P. W. Williams

A B S T R A C T

The different approaches to stability which have been made by various authors are discussed and an abstract framework is formulated within which considerable unification of the existing theories is possible. This framework is then used as a basis for investigating in detail an initial value problem for an ordinary differential equation and an initial value problem for a partial differential equation. In the case of the initial value problem for the partial differential equation the definition of the problem may include boundary conditions or the problem may be of pure initial value type and it is shown that different stability equations arise in these two cases. The problem without boundary conditions has traditionally been investigated using Fourier methods of stability analysis but it is shown here that the same results can be obtained by direct analysis of the matrices involved. This is achieved by formulating the problem so that the matrices have circulant form and using the special properties of such matrices. An appendix is included in which the properties of these circulant matrices are developed. Also a wide class of problems is found for which it is possible to transform a problem involving boundary conditions to a problem in which the boundary conditions are replaced by periodicity conditions. The transformed problem can then be investigated using Fourier methods or the new direct approach based on circulant

matrices. These methods lead to necessary and sufficient conditions for stability in terms of conditions on the powers of a matrix known as the amplification matrix and a survey on the results associated with the amplification matrix is given.

CONTENTS

1. Introduction

1.1 Introductory Remarks 5

1.2 Inherent Instability 7

1.3 Partial Instability 8

1.4 Stability of Multistep Methods 10

2. An Abstract Setting for Convergence Proofs

2.1 Development of the Abstract Formulation 14

2.2 Application to Multistep Methods for Ordinary Differential Equations 25

2.3 Application to a Parabolic Partial Differential Equation 38

3. Stability Analysis in Partial Differential Equations

3.1 Introduction 66

3.2 The Lax-Richtmyer Theory 70

3.3 The Wendroff Theory 93

3.4 Replacement of Boundary Conditions by Periodicity Conditions 100

3.5 Fourier Theory 110

3.6 A Direct Analysis of Stability 115

3.7 The Amplification Matrix 127

4. Summary and Comments 134

Appendix 1 Circulant Matrices 139

References 149

1. Introduction

1.1 Introductory Remarks

Finite difference schemes for the solution of an initial or boundary value problem for an ordinary or partial differential equation yield approximations to the solution at a set of grid points in the domain in which the true solution is defined. The convergence of the approximate solution to the true solution as the grid spacings approach zero depends on the differential problem, the difference scheme, and (in the case where the domain is a region of more than one dimension) on the manner in which the grid spacings of the various dimensions approach zero. In attempting to design difference schemes which are good approximations to the differential problem it is found that some apparently sensible difference schemes are not convergent. Thus it is essential that we have an adequate theory of difference schemes and their convergence properties. Contributions to the theory have been made by Dahlquist (1956) and (1959) and Henrici (1962) and (1963) for ordinary differential equations and by Lax and Richtmyer (1956), Richtmyer (1957), Godunov and Ryabenki (1964) and Richtmyer and Morton (1967) for partial differential equations. These authors have defined two properties of the difference schemes, consistency and stability, and shown that these two conditions are sufficient and in many cases necessary for convergence. Each of the authors has restricted himself to particular classes of problem and although there is considerable compatibility of the ideas involved some of the concepts and

terminology differ from author to author. The abstract framework which is presented here gives a basis for the unification of existing theories and for the further development of the ideas involved. This abstract framework is presented in Section 2 and applied in detail to typical ordinary and partial differential equations to show the practical implementation of the abstract scheme. At a late stage in the development of the theory it was found that the ideas were rather similar to those used by Stetter (1965a), (1965b), (1966), so that the theory cannot now be regarded as an entirely new development. However in applying this theory in detail to particular examples it became clear that certain areas of stability analysis were in need of clarification and further investigation. The Fourier method of stability analysis is widely used but is only theoretically justified for problems without boundary conditions. In Section 3.4 it is shown that a wide class of problems, which include boundary conditions, can be transformed into an equivalent problem without boundary conditions in which Fourier methods of stability analysis can then be used. In practice the results derived from Fourier methods may often be wrongly used to give stability criteria for problems with boundary conditions but using the theory of Section 3.4 it is possible to design finite difference schemes for which the stability properties can be rigorously stated. A further difficulty associated with the use of Fourier methods in the Lax-Richtmyer theory is the necessity of creating an extended function defined for all values of x from the mesh values given

by a finite difference scheme. It is shown in Section 3.6 that for the problems in which Fourier methods can be used an alternative approach can be adopted which is based on the properties of circulant matrices and involves only the mesh values produced in the finite difference scheme. Both of these approaches lead to conditions on the powers of the amplification matrix and the theory associated with the amplification matrix is presented in Section 3.7.

The work does not include a discussion of the influence of round-off error since it is clearly necessary that a difference scheme be stable in the absence of round-off error. However the reader will note that the quantity Λ in the stability formulation could be modified to include a term for round-off error as well as the consistency term. Convergence could still be established in the theoretical situation where the round-off error becomes zero in the limit as $h \to 0$.

Before giving the precise definitions of consistency and stability it would be helpful to discuss some terms which are used in connection with stability, for example in Fox (1962), Modern Computing Methods (1961), Dahlquist (1959) and Henrici (1962).

1.2 Inherent Instability

Consider the equation $\frac{dy}{dt} = y - t$, $y(o) = s + 1$ which has the solution $y = se^{t} + t + 1$. The exponential part of the solution of the differential equation will be represented by a

part of the finite difference solution which increases in a similar manner. Thus even for schemes which are stable according to the definition given later, it is found that initial, round-off or truncation error gives rise to an exponentially increasing error which will normally be insignificant compared to the true solution. If the initial conditions are chosen so that $s = 0$, i.e. Let $y = 1$ when $t = 0$, then there is no exponential part to the true solution but the exponential growth of the error in the finite difference solution will almost always be present and may eventually dominate the approximation to the true solution. However, this is due to the sensitivity of the result to small variations in the initial data and this type of ill-conditioning of the differential problem is inevitably reflected in the solution of the difference scheme.

1.3 Partial Instability

This type of instability, which is singled out by many authors, is a situation where too large a size of h results in a poor approximation to the solution of the differential equation. It is not true instability since it can be removed by reducing the size of h but in some circumstances the effect of partial instability is particularly noticeable. For example this occurs when a known decreasing true solution is transformed into an increasing finite difference solution.

Consider the initial value problem

$$\left.\begin{aligned} y(o) &= s \\ \frac{dy}{dt} &= -\lambda y \end{aligned}\right\} \quad \ldots\ldots\ldots\ldots \qquad (1.1)$$

with $\lambda > 0$, $0 \leq t \leq T$.

If we define a set of mesh points

$$D^h \equiv \{t^n \equiv nh : h = \frac{T}{N},\quad n = 0,\ 1,\ \ldots\ N\}$$

we can use a simple Runge-Kutta formula to provide a solution on D^h. i.e. for the problem $y' = f(t,y)$, $y(o) = s$ we use $\eta^o = s$,

$$\eta^n = \eta^{n-1} + \frac{h}{2} f(t^{n-1},\ \eta^{n-1}) + \frac{h}{2} f(t^n\ ,\ \eta^{n-1} + hf(t^{n-1},\ \eta^{n-1}))$$

$n = 1,\ 2,\ \ldots\ N$

where η^n is the approximation to $y(t^n)$.

The equation gives $\eta^n = \eta^{n-1}(1 - h\lambda + h^2 \frac{\lambda^2}{2})$

and we see that for $h\lambda > 2$, $1-h\lambda + h^2 \lambda^2/2 > 1$.

The sequence η^n is then increasing and cannot represent $y = se^{-\lambda t}$ in any way satisfactorily. We see from the above example that for large values of λ a small value of h is required. Thus in equations of higher order than one, or in simultaneous equations, which give rise to more than one value of λ the largest value of λ will dictate the step size. This means that a rapidly decreasing exponential solution can require a very small step size even after its contribution to the solution has become negligible.

An example of this is given in Modern Computing Methods (1961). Using a simple fourth order Runge-Kutta method on the initial value problem

$$y' = -10y + 6z$$

$$z' = 13.5y - 10z$$

with the initial conditions $y(0) = 4$, $z(0) = 0$ we find solutions

$y = 2e^{-\lambda_1 x} + 2e^{-\lambda_2 x}$, $z = 3e^{-\lambda_1 x} - 3e^{-\lambda_2 x}$ where $\lambda_1 = 1$, $\lambda_2 = 19$.

In order to avoid increasing terms the step length must be chosen so that

$$h\,\lambda^i < 2.8 \qquad i = 1, 2$$

which give respectively the condition $h < 2.8$ and $h < 2.8/19 \doteq 0.14$. Thus, the step length must be taken less than 0.14 even after the part of solution giving rise to this condition has become negligible.

1.4 Stability of Multistep Methods

1.4.1 Strong Instability

This type of instability is associated with multistep methods in which spurious terms of the finite difference solution increase without limit. A decrease in the size of the mesh cannot remove strong instability. We give an example of strong instability for the problem (1.1) with $s = 1$, $\lambda = 1$ which we solve on the mesh D^h of Section 1.3 using the explicit finite difference equation, involving values at three points, which has maximum accuracy.

i.e. $\eta^n = -4\eta^{n-1} + 5\eta^{n-2} + h(-4\eta^{n-1} - 2\eta^{n-2})$

giving $\eta^n + 4(1 + h)\eta^{n-1} - (5 - 2h)\eta^{n-2} = 0$

with solutions $\eta^n = a_1(p_1)^n + a_2(p_2)^n$ where the quantities a_1 and a_2 are dependent on the initial conditions.

$\lim_{h \to 0} p_1 = 1$ and $\lim_{h \to 0} p_2 = -5$ so that the first term is the approximation to the solution of differential equation but the second term is a spurious solution. If the initial conditions are represented by the equations $\eta^0 = s_0$, $\eta^1 = s_1$, which are consistent if $s_0 = s$, $\lim_{h \to 0} s_1 = s$, then we can solve for a_1 and a_2.

$$a_1 + a_2 = s_o = 1 \qquad\qquad a_1p_1 + a_2p_2 = s_1$$

Hence $a_2 = (s_1 - p_1)/(p_2 - p_1) = (s_1 - p_1)/(-6 + r(h))$

where $r(h) = O(h)$. In the particular case where $s_1(h) = p_1(h)$ then $a_2 = 0$ and the scheme would be convergent. However, it is desirable that a scheme should be convergent for any consistent initial conditions which is not the case in this example. If $s_1(h) = p_1(h) + O(h)$ then $a_2 = O(h)$ and the second term of the approximation at the end of the range equals,

$$O(h).(-5 + O(h))^{\frac{T}{h}}$$

which is unbounded as h approaches zero for fixed finite T. The phenomenon of strong instability has been fully considered by Dahlquist (1956) and (1959) and Henrici (1962) for ordinary differential equations.

1.4.2 Weak Stability

If a finite difference scheme is not strongly unstable it may still possess the property that a term in the finite difference solution, which does not correspond to a solution of the differential equation, increases without limit as the length of the interval is increased. An example of this is the formula based on approximate integration by Simpson's Rule for the problem (1.1) with $s = 1$, $\lambda > 0$ on the mesh D^h

i.e. $$\eta^n = \eta^{n-2} - \frac{h}{3}(\lambda\eta^n + 4\lambda\eta^{n-1} + \lambda\eta^{n-2})$$

which has the solution $\eta^n = c_1(p_1)^n + c_2(p_2)^n$ where p_1 approximates $e^{-\lambda h}$ and p_2 approximates $(-1)e^{\frac{\lambda h}{3}}$. Thus we

have a second term in the finite difference solution

$$c_2(-1)^n e^{\frac{\lambda hn}{3}} \quad \text{which is proportional to} \quad e^{\frac{\lambda t^n}{3}}$$

which increases without limit as t^n increases. c_2 is dependent on the closeness of the finite difference scheme to the differential equation and normally approaches zero as the mesh size is decreased. Thus for a finite T we can theoretically make the spurious term as small as we please by reducing the mesh width. Hence the phenomenon is classified as a property of a stable finite difference scheme and not as a form of instability.

1.4.3 General Multistep Methods

Linear multistep methods for the problem $\frac{dy}{dt} = f(t,y)$, $y(0) = s$ on the infinite strip $t \varepsilon I = [0, T], -\infty < y < +\infty$ have the general form

$$\sum_{j=0}^{J} \alpha_j \eta^{n+j} - h \sum_{j=0}^{J} \beta_j f(t^{n+j}, \eta^{n+j}) = 0$$

For a system of differential equations the same representation can be used but y and η will then be vectors and we have a series of equations

$$\frac{dy_i}{dt} = f_i(t, y_1, y_2, \ldots y_d) \qquad i = 1, \ldots, d.$$

Part of the work of Dahlquist (1956) and Henrici (1962) and (1963) was the investigation of the stability of the above finite difference scheme. Stability was defined in terms of the polynomials

$$\rho(z) = \sum_{j=0}^{J} \alpha_j z^j, \qquad \sigma(z) = \sum_{j=0}^{J} \beta_j z^j.$$

According to their definition the difference scheme is said to be stable if and only if the polynomial $\rho(z)$ satisfies the 'root condition', i.e. the roots of the polynomial z_i are such that $|z_i| \leq 1$ and any roots for which $|z_i| = 1$ have multiplicity one.

If there is more than one root $|z_i| = 1$ then the possibility of weak stability arises.

Dahlquist shows that the stability requirement restricts the order of accuracy which can be achieved and gives the following results.

i) The order of a stable operator where J is odd cannot exceed J + 1.

ii) The order of a stable operator where J is even cannot exceed J + 2. The necessary and sufficient condition for the maximum order of accuracy J + 2 is that J be even, all the roots of $\rho(z)$ are of modulus 1, and the following relations exist between the coefficient of $\rho(z)$ and $\sigma(z)$

$$\alpha_j = -\alpha_{J-j} \qquad \beta_j = \beta_{J-j}$$

Dahlquist also shows that a stable operator with maximum order of accuracy will always have the property of weak stability.

2. An Abstract Setting for Convergence Proofs

2.1 Development of the Abstract Formulation

2.1.1 A Simple Example

A formal definition of consistency and stability is given later in this section but in order to introduce the ideas a particular example of an ordinary differential equation is discussed here. We consider the initial value problem

$$y(o) = s \qquad 1(a)$$

$$\frac{dy}{dt} = f(t, y) \qquad 1(b)$$

$t \in I \equiv [0, T]$, $f(t, y)$ is a continuous function of the variable t in the infinite strip $t \in I$, $|y| < \infty$ and satisfies a Lipschitz condition with respect to y

$$\text{i.e.} \quad |f(t, y + z) - f(t, y)| \leq L |z|.$$

To obtain the finite difference solution, values are calculated on the mesh

$$D^h \equiv \{t^n \equiv nh, \quad h = \frac{T}{N}, \quad n = 0, 1, \ldots N\}$$

by means of the following equations.

$$\eta^0 - s = 0 \qquad 2(a)$$

$$\eta^1 - \eta^0 - hf(t^0, \eta^0) = 0 \qquad 2(b)$$

$$\eta^n - \eta^{n-2} - \frac{h}{3}\left[f(t^n, \eta^n) + 4f(t^{n-1}, \eta^{n-1}) + f(t^{n-2}, \eta^{n-2})\right] = 0 \qquad 2(c)$$

$$n = 2, 3, \ldots N$$

where η^n is the approximation to $y(t^n)$.

In order that the difference scheme is 'consistent' in terms of our later definition the difference scheme will be written in

the following equivalent form

$$\eta^0 - s = 0 \qquad 3(a)$$

$$\frac{\eta^1 - \eta^0}{h} - f(t^0, \eta^0) = 0 \qquad 3(b)$$

$$\frac{\eta^n - \eta^{n-2}}{2h} - \frac{1}{6}[f(t^n, \eta^n) + 4f(t^{n-1}, \eta^{n-1}) + f(t^{n-2}, \eta^{n-2})] = 0 \qquad 3(c)$$

For a sensible difference scheme the finite difference operator must approximate the differential operator in the limit as $h \to 0$. If η^n is replaced by $y(t^n)$ on the left hand side of equations (3) we have

$$y(t^0) - s = \tau^0 \qquad 4(a)$$

$$\frac{y(t^1)-y(t^0)}{h} - f(t^0, y(t^0)) = \tau^1 \qquad 4(b)$$

$$\frac{y(t^n)-y(t^{n-2})}{2h} - \frac{1}{6}[f(t^n, y(t^n)) + 4f(t^{n-1}, y(t^{n-1})) + f(t^{n-2}, y(t^{n-2}))] = \tau^n \qquad 4(c)$$

where $\tau^n \to 0$ as $h \to 0$. For $n = 0$, $\tau^n \equiv 0$ and for $n = 2, 3, \ldots N$ the result can be established using only the continuity of $y'(t)$. (e.g. See Henrici (1963)). Note that the condition $\lim_{h \to 0} \tau^n = 0$ is not meaningful unless the difference scheme is in a special form. It is clear that a factor of h^p, $p >> 0$ could be introduced into a difference scheme to make almost any schemes satisfy $\lim_{h \to 0} \tau^n = 0$.

If we denote the left hand side of equations (2) by

$$F_0^h(\eta^0, \eta^1 \ldots \eta^N), \; F_1^h(\eta^0, \eta^1, \ldots \eta^N), \; F_n^h(\eta^0, \eta^1 \ldots \eta^N)$$

respectively then the difference scheme can be written concisely as

$$F_n^h(\eta) \equiv F_n^h(\eta^o, \eta^1, \ldots \eta^N) = 0, \quad n = 0, 1, \ldots N.$$

Equations (4) can then be written

$$Q_n F_n^h(\Delta_h y) = \tau^n$$

where

$$Q_o F_o^h \equiv F_o^h \qquad 5(a)$$

$$Q_1 F_1^h \equiv \frac{1}{h} F_1^h \qquad 5(b)$$

$$Q_n F_n^h \equiv \frac{1}{2h} F_n^h \qquad n = 2, 3, \ldots N \qquad 5(c)$$

and where

$$(\Delta_h y)_i \equiv y(t^i)$$

Δ_h can be considered as an operation of discretization.

2.1.2 Abstract Formulation

By the introduction of some simple ideas in functional analysis we can formalize the concepts of consistency and stability and show how this particular problem fits into the general framework. We introduce the normed spaces $E \equiv C'[0,T]$ and $E^o \equiv RxC[0,T]$. Thus an element $w \in E^o$ consists of an ordered pair the first component of which is a real number and the second of which is a function defined on I

i.e. $\quad w = (w^R, w^C)$

If we define the mapping $F:E \to E^o$ by

$$(F(u))^R \equiv u(0) - s \qquad u \in E$$

$$(F(u))^C \equiv \frac{du}{dt} - f(t, u) \qquad (6)$$

then the initial value problem can be written as $F(y) = 0$ where 0 is the zero element of E^o.

In the finite difference scheme (3) the vector η is an element of a normed space E_h (which in this case is identical to R_{N+1}) and the right hand side belongs to a normed space E_h^o (in this case E_h^o is also identical to R_{N+1}). Thus the left hand side of the finite difference scheme (3) can be interpreted as a mapping F^h: $E_h \rightarrow E_h^o$ defined by

$$(F^h \eta)^o \equiv \eta^o - s$$

$$(F^h \eta)^1 \equiv \frac{\eta^1 - \eta^o}{h} - f(t^o, \eta^o)$$

$$(F^h \eta)^n \equiv \frac{\eta^n - \eta^{n-2}}{2h} - \frac{1}{6}\left(f(t^n, \eta^n) + 4f(t^{n-1}, \eta^{n-1}) + f(t^{n-2}, \eta^{n-2})\right)$$

$$n = 2, 3, \ldots N \qquad (7)$$

where the components of an element $\chi \in E_h^o$ are $\chi^o, \chi^1, \ldots \chi^N$. The finite difference solution Y is given by the solution of the equations $F^h Y = 0$.

In order to compare the true solution with the finite difference solution we associated with any member of the space E an element of E_h by a mapping $\Delta_h : E \rightarrow E_h$ where in this case the mapping could be defined by

$$(\Delta_h u)_i \equiv u(t^i) \qquad i = 0, 1, \ldots N \qquad (8)$$

We also discretize the elements of E^o which we represent by a mapping $\Delta_h^o : E^o \rightarrow E_h^o$.

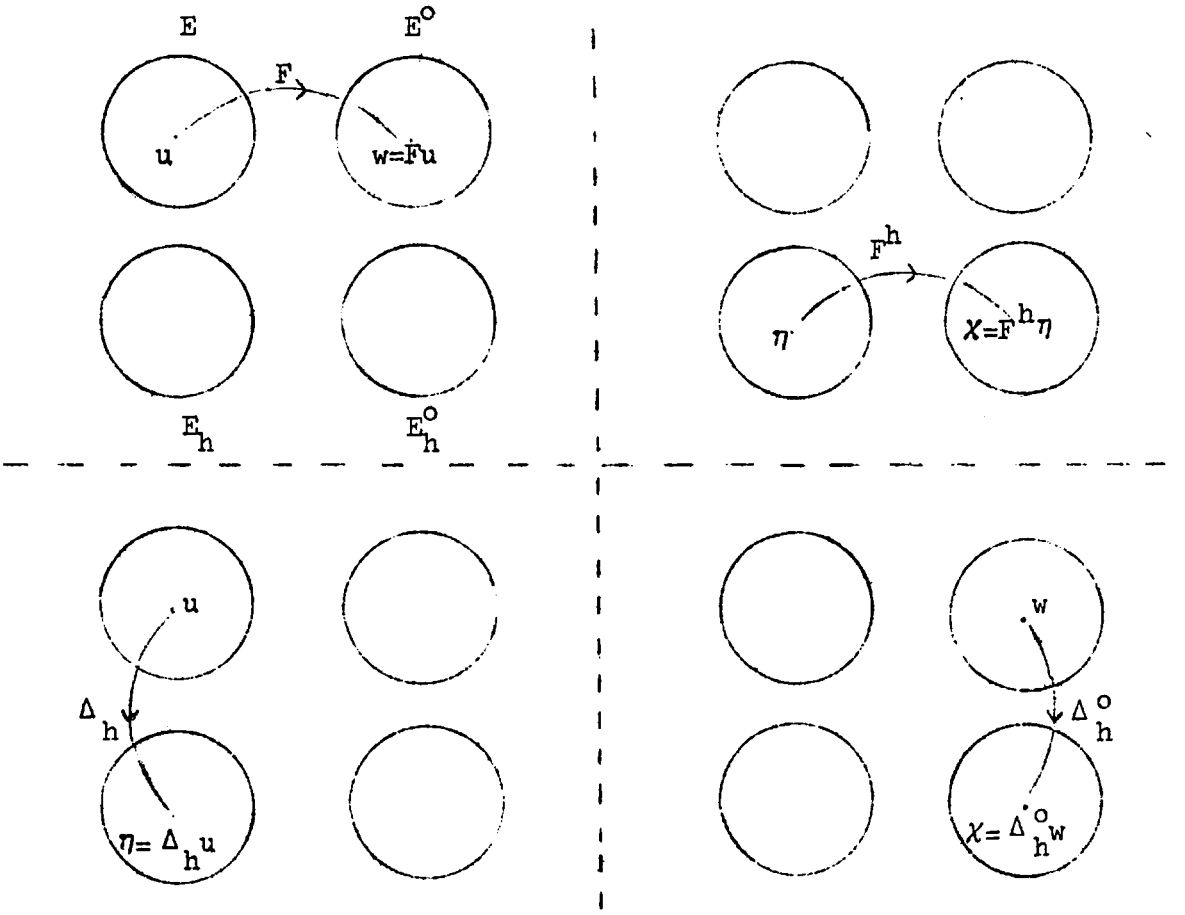
For example we could define Δ_h^o by

$$(\Delta_h^o w)^R = w^R$$

$$(\Delta_h^o w)_i^C = w^C(t^i) \qquad i = 1, 2, \ldots N \qquad (9)$$

These are typical examples of discretization and other situations may suggest different definitions of this operation.

We can represent this abstract structure diagrammatically in terms of the spaces and mappings involved.



In each of the four spaces a norm will be defined but in the future analysis it will be clear from the context which norm is involved and so no distinguishing subscripts will generally be used.

In order that the discrete spaces correspond adequately with the continuous spaces we will choose norms in corresponding spaces that are consistent; i.e. the norms in the spaces E and E_h and in the spaces E^o and E_h^o must be equal in the limit as $h \to 0$.

$$\lim_{h \to 0} \| \Delta_h u \| = \| u \| \quad u \in E$$

$$\lim_{h \to 0} \| \Delta_h^o w \| = \| w \| \quad w \in E^o$$

For example $\| \eta \| = h \max_n | \eta^n |$, $\eta \in E_h$ would not be a suitable norm for comparison with $\| u \| = \max | u(t) |$, $u \in E$ since $\lim_{h \to 0} \| \Delta_h u \| = 0$ would not imply $\| u \| = 0$.

The two norms of most interest in a linear space C[a, b] are

a) $\| u \|_p = \left(\int_a^b |u(t)|^p \, dt \right)^{\frac{1}{p}}$ p a positive integer

b) $\| u \|_\infty = \max_I |u(t)|$

for $u \in E$, and for the corresponding finite dimensional space

a) $\| \eta \|_p = \left[\left(\frac{b-a}{N} \right) (| \eta^0 |^p + | \eta^1 |^p + \ldots | \eta^N |^p \right]^{\frac{1}{p}}$

b) $\| \eta \|_\infty = \max_n (| \eta^n |)$ $\quad n = 0, 1 \ldots N$

for $\eta \in E_h$. We will also ensure that the norms in the spaces E^o and E_h^o are consistent.

2.1.3 Consistent Difference Schemes

The spaces E_h and E_h^o and the mapping F^h are designed to 'mirror' the properties of the spaces E and E^o and the mapping F. Thus if a difference scheme is to be sensible we should expect that

the associated mapping F^h would be such that the mappings $F^h\Delta_h$ and $\Delta_h^o F$ from E to E_h^o would approximate one another. i.e. $\| F^h \Delta_h u - \Delta_h^o F(u) \|$ should be small and certainly should approach zero as $h \to 0$. In order to prove convergence the minimum requirement is that a condition of this kind should hold at the solution $y \in E$. But since the solution is unknown it is reasonable to expect that F^h should approximate F over a range of elements $u \in E$ and in many cases this will be the entire space E. However we have seen previously that the finite difference schemes (2) and (3) are different formulations of the same computational procedure. Any finite difference scheme can generate an infinite number of schemes which have the same solution and not all of these will correspond to F in the limit as $h \to 0$. Before discussing which finite difference scheme is most appropriate when defining consistency the concept of equivalent difference schemes will be expressed more formally. Two operators F^h and G^h will be said to represent equivalent difference schemes if $F^hY = 0 \Leftrightarrow G^hY = 0$. In particular if Q^h is a mapping of E_h^o into E_h^o such that $\chi = 0 \Leftrightarrow Q^h\chi = 0$ then Q^hF^h is equivalent to F^h. In the example given previously the operator Q^h which transforms equations (2) to equations (3) is defined by equations (5).

The requirement that a difference scheme has the property $\lim_{h \to 0} \| F^h \Delta_h y \| \to 0$ is now seen to be a very weak condition unless the difference scheme is carefully chosen from the range of equivalent schemes. This choice can be made by specifying a mapping F^h which approximates F not only at the solution y but at least in some neighbourhood of the solution $\| u - y \| \leq \partial$, $\partial > 0$.

We therefore define u-consistency as follows.

Definition 1

The difference scheme represented by F^h is said to be u-consistent if there exists an equivalent difference scheme $Q^h F^h$ and a domain $D \subset E$ $y \in D$ such that $\lim_{h \to o} \| Q^h F^h \Delta_h u - \Delta_h^o F u \| \to 0$ for any $u \in D$.

This is a more demanding definition than is required to prove convergence since D need only be the single point y. In the case where $D \equiv y$ since $Fy = 0$ the second term in the consistency definition becomes zero giving $\lim_{h \to o} \| Q^h F^h \Delta_h y \| \to 0$. Although the form $Q^h F^h$ which gives a direct comparison with the initial value problem is often easily found there is no reason why y-consistency should not be defined in terms of any difference scheme represented by F^h. For example the discussion by Henrici (1963) on multistep methods considers a form similar to equations (2) with a consistency condition $\lim_{h \to o} \| F^h \Delta_h y \| = o(h)$.

In the subsequent presentation difference schemes will always be formulated so that the operator F^h will be in a form which can be used directly in the definition of consistency without an operator Q^h.

2.1.4 Stability and Convergence

The term convergent which has been used previously will now be formally defined.

Definition 2

The finite difference scheme associated with F^h is said to be convergent to the solution of the problem $Fy = 0$

if the solutions of $F^h Y = 0$ are such that $\lim_{h \to o} \| \Delta_h y - Y \| = 0$.

Note that we have a sequence of spaces E_h and vectors Y whose dimensionality increases with n.

The most useful methods will be those which are convergent for arbitrary consistent starting methods and applicable to a class of initial value problems although the minimum requirement to prove convergence in a particular case may be less restrictive.

If we define $E = \Delta_h y - Y$ we have $F^h(\Delta_h y) \equiv F^h(Y + E) = Q$ and since $\lim_{h \to o} \| Q \| = 0$ by consistency we require for convergence that $\lim_{h \to o} \| F^h(Y + z) \| = 0$ implies $\lim_{h \to o} \| z \| = 0$

A scheme which satisfies this condition has the property of stability. This definition of stability is a special case of the following definition which is more suitable for the further theoretical development.

Consider the equation

$$G^h(z) \equiv F^h(\xi + z) - F^h(\xi) = \Lambda \qquad (10)$$

$$z \in E_h, \quad \xi \in S \subset E_h, \quad \Lambda \in T \subset E_h^o$$

Definition 3

The finite difference scheme is said to be S-stable in a region T if there exists a function $H_\xi(x)$ satisfying the conditions.

i) $H_\xi(x) > 0$ for $x > 0$, for all $\xi \in S$

ii) $\lim_{x \to o} H_\xi(x) = 0$ for all $\xi \in S$

and for any $z \in E_h$ the solution of the equation $G^h(z) = \Lambda$ satisfies the inequality $\| z \| \leq H_\xi (\| \Lambda \|)$ for all $\xi \in S$ and all $\Lambda \in T$. S,T may be defined as the entire spaces E_h or E_h^o or as a restricted region such as $S \equiv \{\eta : \| \eta \| \leq \eta_o, \eta_o \text{ constant}\}$.

In the special case where $S \equiv Y$ since $F^h Y = 0$ the equation $G^h(z) = \Lambda$ reduces to $F^h(Y + z) = \Lambda$ so that the less demanding condition $\| z \| \leq H_Y(\| \Lambda \|)$ is clearly sufficient for convergence. If $S \equiv Y$ and T is the region $\|\Lambda\| \leq \Lambda_o$, Λ_o constant this is the least demanding situation which will be sufficient for convergence if the scheme is consistent. Most definitions of stability are more restrictive examples of definition 3. For example Godunov and Ryabenki (1964) take the special case where F^h is linear so that $G^h(z) = \Lambda$ reduces to $F^h z = \Lambda$ and use for the function H, (which is now independent of ξ), a linear function $K \|\Lambda\|$ where K is independent of h. We will use the term global stability when $T \equiv E_h^o$ and the term local stability when T is a restricted region. There are finite difference schemes in which the stability condition only holds in a region T which is dependent on h e.g. $\|\Lambda\| \leq \Lambda_o h^m$, $m > 0$, and in this situation it is possible to achieve convergence by means of a stronger condition of consistency. Stetter (1966) considers non-linear equations and defines 'm-restricted stability' where T is defined by $\|\Lambda\| \leq \Lambda_o h^m$ and $H_\xi (\| \Lambda \|) \equiv K \| \Lambda \|$. If the consistency condition is strengthened to $\| Q \| = o(h^m)$ then

$$\lim_{h \to o} \frac{\|Q\|}{h^m} = 0$$

$\therefore$ There exists h_o such that

$$\frac{\|Q\|}{h^m} < \Lambda_o$$

i.e. $\| Q \| < \Lambda_o h^m$ for all $h \leq h_o$.

Hence there exists a value $h_o > 0$ such that for all $h \leq h_o$ the consistency error lies within the region of stability T and this will give convergence.

The two conditions of stability and consistency are interrelated. If a consistency condition $\| Q \| = O(h^m)$, m positive, is true then it is possible to relax the stability condition (ii) on the function H_ξ. All that would be required for convergence is

$$\lim_{x \to o} H_\xi(x) = o\left(\frac{1}{h^m}\right) \quad \text{for all} \quad \xi \in S.$$

However such a condition would be more dangerous in practice owing to the presence of round-off error which would not normally have the same behaviour as the consistency error.

In the above development the abstract formulation has been in terms of normed spaces whereas some of the other presentations are based on Banach spaces, (e.g. Lax-Richtmyer, Stetter). For many problems it is possible to show that consistency and stability are sufficient to ensure convergence without using the completeness property associated with the Banach space. However, in certain problems the elements which are involved can only be defined as the limit points of a space and the space must be carefully defined to form a Banach space so that the limit points are members of the space. For example if the space E is composed

of functions u(x) which are twice continuously differentiable with respect to x and the norm used is

$$\| u \|^2 = \int_a^b | u(x) |^2 \, dx$$

then a discontinuous initial function would not be a member of the space but a sequence of elements of the space E could be found which approach the discontinuous function in the limit. The use of a Banach space would ensure that the limit points were members of the space. In the subsequent analysis normed spaces will be used for the sake of simplicity but the extension to Banach spaces can be made if required in a particular problem.

2.2 Application to Multistep Methods for Ordinary Differential Equations

A linear multistep method applied to problem (1) can be represented by the mapping

$$(F^h \eta)^i \qquad i = 0, 1, \ldots J-1. \qquad 11(a)$$

$$(F^h \eta)^{n+J} = \sum_{j=0}^{J} \alpha_j \eta^{n+j} - \sum_{j=0}^{J} \beta_j f(t^{n+j}, \eta^{n+j}) \qquad 11(b)$$

$n = 0, 1, \ldots N-J.$

The α_j and β_j may depend upon h and a suitable form for these coefficients will arise from the consistency requirement. $(F^h \eta)^i$ represents a method for generating the starting values η^i required for the multistep formula and in a practical example this could be a Runge-Kutta or Taylor Series formula. In order to concentrate on the multistep formula Euler's method will be

used for the starting method since although it is unsuitable practically the analysis involved is simple. The discretization will be on the mesh D^h according to equations (8) and (9) and the appropriate maximum norm will be used in spaces E, E_h, E_h^o. The norm in space E^o will be defined by

$$\| w\|^o = \max(| w^R| , \max_I | w^C |).$$

If $F^h \Delta_h u - \Delta_h^o F u = \chi$

then sufficient conditions for u-consistency are that

$$\lim_{h \to o} \max_i | \chi^i | = 0 \qquad i = 0, 1, \ldots J-1$$

and $$\lim_{h \to o} \max_j | \chi^j | = 0 \qquad j = J, \ldots N$$

since $$\| \chi \|_\infty = \max(\max_i |\chi^i| , \max_j | \chi^j |)$$

the starting method and the multistep method can be considered separately.

$$\chi^o = u(o) - s - [u(o) - s] \equiv 0$$

$$\chi^i = \frac{u(t^i) - u(t^{i-1})}{h} - f(t^{i-1}, u(t^{i-1})) - (\frac{du}{dt})_{t=t^i} + f(t^{i-1}, u(t^{i-1}))$$

$i = 1, \ldots J-1.$

For any function v(t) continuous on I there exists an associated function $w(v, \delta)$, the modulus of continuity, defined by

$$w(v, \delta) = \max_{\substack{|t^1 - t^2| \leq \delta \\ t^1, t^2 \in I}} |v(t^1) - v(t^2)|$$

and clearly $\lim_{\delta = 0} w(v, \delta) = 0.$

Thus since $u \in C'[0,T]$ and $g(t) \equiv f(t, u(t)) \in C[0,T]$ there exist $w(u', \delta)$ and $w(g, \delta)$ and for brevity we denote $w(u', \delta) \equiv w(\delta)$ $w(g, \delta) \equiv \Omega(\delta)$.

Hence $$u(t^{n+j}) = u(t^n) + jhu'(t^n) + jh\theta_1 w(jh) \qquad (12)$$

$$f(t^{n+j}, u(t^{n+j})) = f(t^n, u(t^n)) + \theta_2 \Omega(jh) \qquad (13)$$

$$|\theta_1|, |\theta_2| \leq 1$$

Thus $|\chi^i| \leq (\max |\theta_1 w(h)|) \to 0$ as $h \to 0$

for $i = 1, \ldots J-1$.

For the multistep formula

$$\chi^{n+J} = \sum_{j=0}^{J} \alpha_j u(t^{n+j}) - \sum_{j=0}^{J} \beta_j f(t^{n+j}, u(t^{n+j}))$$

$$- \left(\frac{du}{dt}\right)_{t=t^n} + f(t^n, u(t^n))$$

$$= \sum_{j=0}^{J} \alpha_j [u(t^n) + jhu'(t^n) + jh\theta_1 w(jh)] - u'(t^n)$$

$$- \sum_{j=0}^{J} \beta_j [f(t^n, u(t^n)) + \theta_2 \Omega(jh)] + f(t^n, u(t^n)) \qquad (14)$$

and the following conditions will give u-consistency

$$\sum_{j=0}^{J} \alpha_j = 0 \qquad \sum_{j=0}^{J} jh\, \alpha_j = 1 \qquad \sum_{j=0}^{J} \beta_j = 1$$

If we put $\alpha'_j = h\alpha_j$ then the following conditions are sufficient for u-consistency.

$$\sum_{j=0}^{J} \alpha'_j = 0 \qquad \sum_{j=0}^{J} j\,\alpha'_j = 1 \qquad \sum_{j=0}^{J} \beta_j = 1 \tag{15}$$

For y-consistency when $y' = f(t, y)$ the conditions become

$$\sum_{j=0}^{J} \alpha'_j = 0$$

$$\sum_{j=0}^{J} (j\alpha'_j - 1 - \beta_j + 1) = 0 \qquad \text{i.e.} \quad \sum_{j=0}^{J} j\alpha'_j = \sum_{j=0}^{J} \beta_j \tag{16}$$

In terms of the polynomials $\rho(z) = \sum_{j=0}^{J} \alpha'_j z_j$ and $\sigma(z) = \sum_{j=0}^{J} \beta_j z^j$ the above results are the same as those of Dahlquist (1956) and Henrici (1962) namely

$$\rho(1) = 0 \qquad \rho'(1) = \sigma(1) .$$

In the future analysis the superscript ' on α will be dropped.

In order to establish stability we will first show that the part of the equation $G^h(z) = \Lambda$ associated with the starting method satisfies

$$\max(|z^0| , |z^1| , \ldots . |z^{J-1}|) \leq N_1 || \Lambda || \tag{17}$$

(where N_1 is independent of h).

We will then show that the values $|z^n|$ $n = J, J+1, \ldots N$ obtained from the multistep formula satisfy a condition.

$$\max_{\eta}(|z^n|) \leq N_2 || \Lambda || \tag{18}$$

(where N_2 is independent of h) if the initial values of the J-th order finite difference equation satisfy condition (17). Then

$\max_n(|z^n|) \leq \max(N_1, N_2) \|\Lambda\|$ $n = 0, 1, \ldots N$ and the

finite difference scheme is globally stable at any point $\eta \in E_h$.

$$(G^h(z))^o \equiv z^o = \Lambda^o$$

$$(G^h(z))^i \equiv \frac{z^i - z^{i-1}}{h} - [f(t^{i-1}, \eta^{i-1} + z^{i-1}) - f(t^{i-1}, \eta^{i-1})] = \Lambda^i$$

$$i = 1, 2, \ldots J-1.$$

Hence $|z^o| \leq \|\Lambda\|$

$$|z^i| \leq |z^{i-1}| + hL|z^{i-1}| + h\|\Lambda\|$$

$$\leq (1 + hL)|z^{i-1}| + h\|\Lambda\|$$

Using the formula recursively gives

$$|z^i| \leq (1 + hL)^i |z^o| + h[\sum_{j=0}^{i-1} (1 + hL)^j] \|\Lambda\|$$

$$\leq (1 + hL)^N \|\Lambda\| + h\left(\frac{(1 + hL)^N - 1}{1 + hL - 1}\right)\|\Lambda\|$$

$$\leq \left(e^{hLN} + \frac{e^{hLN} - 1}{L}\right)\|\Lambda\|$$

since $hN = T$ we have the condition

$$|z^i| \leq K_1 \|\Lambda\| \qquad i = 0, 1, \ldots J-1 \qquad (19)$$

and hence condition (17) is satisfied with $N_1 = K_1$.

The multistep algorithm gives the equation

$$(G^h(z))^{n+J} \equiv \sum_{j=0}^{J} \frac{\alpha_j}{h} z^{n+j} - \sum_{j=0}^{J} \beta_j [f(t^{n+j}, \eta^{n+j} + z^{n+j}) - f(t^{n+j}, \eta^{n+j})]$$

$$= \Lambda^{n+j}$$

$$n = 0, 1, \ldots N-J$$

Hence $$\sum_{j=0}^{J} \alpha_j z^{n+j} = C^{n+J} \qquad (20)$$

where $$C^{n+J} = h \sum_{j=0}^{J} \beta_j [f(t^{n+j}, \eta^{n+j} + z^{n+j}) - f(t^{n+j}, \eta^{n+j})] + h \Lambda^{n+J}.$$

The analysis will be concerned with methods which are applicable to any problem of the type described by equations (1) and we will first show that a necessary condition for stability is that the solution of the difference equation

$$\sum_{j=0}^{J} \alpha_j z^{n+j} = 0 \qquad n = 0, 1, \ldots N-J \qquad (21)$$

subject to initial values z^i $i = 0, 1, \ldots J-1$ must satisfy the 'root condition'; i.e. any root X_p of the associated polynomial

$$\sum_{j=0}^{J} \alpha_j X^j = 0 \qquad (22)$$

must satisfy the condition $|X_p| \leq 1$ if it is a simple root or $|X_p| < 1$ if X_p is a multiple root. Since the scheme must be stable for arbitrary Λ we choose $\Lambda^{n+J} = 0$ $n = 0, 1, \ldots N-J$. If we wish the multistep method to be stable for all functions $f(t, y)$ satisfying a Lipschitz condition then it must be stable for $f(t, y) \equiv 0$, $s = 0$. Since $C^{n+J} \equiv 0$, $n = 0, 1, \ldots N-J$ the equation (20) now reduces to the homogeneous form (21). The solution of (21) can be expressed in terms of the solutions of the polynomial (22)

$$\sum_{j=0}^{J} \alpha_j X^j = 0.$$

Each root X_p of multiplicity m_p gives rise to m_p linearly independent solution of the difference equation and the complete solution is

$$z^{n+J} = \sum_{p=1}^{r} \sum_{q=0}^{m_p-1} b^{(pq)} v_{n+J}^{(pq)} \qquad n = 0, 1, \ldots N-J$$

$$v_n^{(pq)} = n^q X_p^n \qquad q = 0, 1, \ldots m_p-1 \qquad (23)$$

where the superscripts q and n on the right hand side are indices. $\sum_{p=1}^{r} m_p = J$ where the polynomial (22) has r distinct roots. We consider the starting method $\eta^i = 0 \quad i = 0, 1, \ldots J-1$ which gives the exact solution and satisfies condition (17) since

$$(G^h(z))^i \equiv z^i = \Lambda^i \qquad (24)$$

$$i = 0, 1, \ldots J-1$$

and a solution of (24) and (21) is given by

$$z^n = h\, n^q X_p^n \qquad n = 0, 1, \ldots N. \qquad (25)$$

In this case $\Lambda^i = hi^q X_p^i \quad i = 0, 1, \ldots J-1$ and $\lim_{h \to 0} \Lambda^i = 0$ so that the scheme will be unstable unless $\lim_{h \to 0} z^n = 0$

$$n = J, J+1 \ldots N$$

Now $$z^n = h\left(\frac{t}{h}\right)^q (X_p)^{\frac{t}{h}}, \qquad t \leq T$$

and this quantity has limit zero as $h \to 0$ unless $|X_p| > 1$ for a simple root or $|X_p| \geq 1$ for a multiple root (e.g. See Henrici (1962)p219).

It is therefore necessary for stability that simple roots of (22) satisfy $| X_p | \leq 1$ and multiple roots satisfy $| X_p | < 1$.

It can be shown that the root condition is also sufficient for stability. Equation (20) has J linearly independent solutions defined by (23) but in order to simplify the analysis we take as the fundamental set of solutions a linear combination of the $v^{(pq)}$

$$u^n(i) = \sum_{p=1}^{r} \sum_{q=0}^{m_p-1} l_{(i)}^{(pq)} v_n^{(pq)} \qquad i = 0, 1, \ldots J-1 \tag{26}$$

where the coefficients are independent of h and are determined from the equations

$$u_{(i)}^j = \sum_{p=1}^{r} \sum_{q=0}^{m_p-1} v_j^{(pq)} l_{(i)}^{(pq)} = \delta_{ji} \qquad j = 0, 1, \ldots J-1$$

where $\delta_{ji} = 1 \quad i = j, \quad \delta_{ji} = 0 \quad i \neq j.$

Using a discrete analogue of Duhamel's principle (e.g. See Isaacson and Keller (1966)) we can show that the solution of (20) has the form

$$z^n = \sum_{j=0}^{J-1} z^j u_{(j)}^n + \frac{1}{\alpha_J} \sum_{i=0}^{n-J} C^{J+i} u_{(J-1)}^{n-i-1} \tag{27}$$

$$n = 0, 1, \ldots N$$

and $$| C^{i+J} | \leq h[L \sum_{j=0}^{J} | \beta_j | \quad | z^{i+j} | + \| \Lambda \|] \tag{28}$$

Since the $u^n_{(j)}$ are linear combinations of the $v_n^{(pq)}$ with coefficients independent of h then if the root condition holds

$$\max |u^n_{(j)}| \leq K_2, \quad 0 \leq j \leq J-1 \quad \text{and we note that } K_2 \geq 1.$$

$$\text{Hence } |z^n| \leq K_2 \sum_{j=0}^{J-1} |z^j| + \frac{K_2 h}{|\alpha_J|} \sum_{i=0}^{n-J} \left(L \sum_{j=0}^{J} |\beta_j| \; |z^{i+j}| + \|\Lambda\| \right)$$

$$\leq K_2 \sum_{j=0}^{J-1} |z^j| + \frac{K_2 hL}{|\alpha_J|} \sum_{i=0}^{n-J} \sum_{j=0}^{J} |\beta_j| \; |z^{i+j}| + \frac{hK_2 N}{|\alpha_J|} \|\Lambda\|.$$

If the starting method is stable then

$$|z^i| \leq K_1 \|\Lambda\| \qquad i = 0, 1, \ldots J-1$$

and if $B = \max |\beta_j| \qquad j = 0, 1, \ldots J-1$

$$\text{then } |z^n| \leq \left(K_1 K_2 J + \frac{hK_2 N}{|\alpha_J|} \right) \|\Lambda\| + \frac{hLK_2 B(J+1)}{|\alpha_J|} \sum_{i=0}^{n} |z^i|.$$

$$\text{Let } B^* = \frac{hLK_2 B(J+1)}{|\alpha_J|}$$

$$hL^* = B^*/(1 - B^*)$$

$$K^* = \left(K_1 K_2 J + \frac{hK_2 N}{|\alpha_J|} \right) / (1 - B^*).$$

$$\text{Hence } |z^n| \leq hL^* \sum_{i=0}^{n-1} |z^i| + K^* \|\Lambda\| \qquad (29)$$

$$n = J, J+1, \ldots N.$$

Note that for sufficiently small h

$$1 \geq 1 - B^* > 0$$

and since $K_2 \geq 1$, $J \geq 1$ then using (29) $K^* \geq K_1$.

Hence $|z^i| \leq K_1 \|\Lambda\| \leq K^*(1 + hL^*)^i \|\Lambda\|$

$i = 0, 1, \ldots J-1$

Now $|z^J| \leq hL^*K^* \sum_{i=0}^{J-1} (1+hL^*)^i \|\Lambda\| + K^* \|\Lambda\|$

$$= K^* \|\Lambda\| \left[hL^* \left(\frac{(1+hL^*)^J - 1}{hL^*}\right) + 1\right]$$

$$= K^*(1 + hL^*)^J \|\Lambda\|$$

and by induction we have

$$|z^n| \leq K^*(1 + hL^*)^n \|\Lambda\| \leq K^* e^{L^*hn} \|\Lambda\| \qquad (30)$$

$n = 0, 1, \ldots N.$

Since this is true for all n we have the stability condition

$$\|z\| \leq K \|\Lambda\| \qquad (31)$$

where $K = K^* e^{L^*T}$, $\qquad T = Nh.$

If a predictor-corrector method is used with a single application of the corrector similar analysis can be applied to a modified equation 11(b) in which the β_J term is replaced by $\beta_J f(t^{n+J}, \tilde{\eta}^{n+J})$ is determined from a predictor equation of the type

$$\tilde{\eta}^{n+J} = \sum_{r=J-R}^{J-1} a_r \eta^{n+r} - h \sum_{r=J-R}^{J-1} b_r f(t^{n+r}, \eta^{n+r}).$$

In the consistency equation (14) the β_J term is replaced by

$$f(t^{n+J}, \tilde{u}(t^{n+J})) = f[t^{n+J}, \sum_{r=J-R}^{J-1} a_r u(t^{n+r}) - h \sum_{r=J-R}^{J-1} b_r f(t^{n+r}))] \tag{32}$$

$$= f(t^{n+J}, \sum_{r=J-R}^{J-1} a_r u(t^{n+J}) + O(h))$$

$$= f(t^{n+J}, \sum_{r=J-R}^{J-1} a_r u(t^{n+J})) + O(h) \quad \text{by continuity of } f,$$

$$= f(t^{n}, \sum_{r=J-R}^{J-1} a_r u(t^{n})) + \theta\, \Omega'(Jh) + O(h),$$

where Ω' is the modulus of continuity of $f(t^n, \sum_{r=J-R}^{J-1} a_r u(t^n))$ and $|\theta| < 1$. Hence the equations (15) will be unchanged if

$$\sum_{r=J-R}^{J-1} a_r = 1$$

and this weak condition on the predictor equation is sufficient for the consistency of the combined predictor-corrector method. However the rate of convergence will be adversely affected unless the usual consistency conditions are valid for the predictor formula also.

The proof of the necessity of the root condition for the corrector equation is unaltered. In the proof of sufficiency the solution is given by (27) with the modified value of C^{i+J} defined by the new term (32)

$$|C^{i+J}| \leq h(L \sum_{j=0}^{J-1} |\beta_j| \cdot |z^{i+j}| + \|\Lambda\|)$$

$$+ hL|\beta_J|[\sum_{r=J-R}^{J-1} |a_r| \cdot |z^{i+r}| + h \sum_{r=J-R}^{J-1} L|b_r|\,|z^{i+r}|]$$

Assume $R \geq J$ for simplicity. The argument would follow the same lines with minor notational changes if $R < J$.

Let $B = \max(|\beta_j| + |\beta_J| . |a_j|)$ $\quad j = J-R, \ldots J-1$, where $\beta_j = 0$ for $j < 0$.

$$\text{Then } |C^{i+J}| \leq hLB \sum_{r=J-R}^{J-1} |z^{i+r}| + h\|A\|$$

$$+ h^2 L^2 |\beta_J| \sum_{r=J-R}^{J-1} |b_r| \, |z^{i+r}|$$

Hence if the starting method is stable

$$|z^n| \leq K_1 K_2^J \|A\| + \frac{hK_2 N \|A\|}{|\alpha_J|} + \frac{K_2 hL}{|\alpha_J|} \sum_{i=0}^{n-J} \sum_{r=J-R}^{J-1} (B + hL|\beta_J| . |b_r|) |z^{i+r}|$$

for $h \leq ho$ $\quad B + hL|\beta_J| . |b_r| \leq B' \qquad r = J-R, \ldots J-1.$

$$\text{Hence } |z^n| \leq \left(K_1 K_2^J + \frac{h.K_2.N}{|\alpha_J|}\right) \|A\| + \frac{K_2 h.LB'.R}{|\alpha_J|} \sum_{i=0}^{n-1} |z^i|$$

$$\leq K^* \|A\| + hL^* \sum_{i=0}^{n-1} |z^i| \qquad (33)$$

which has the same form as equation (29) but with different values for the constants K* and L*. Hence condition (31) is still true.

Thus for the predictor-corrector scheme to be stable the only condition on the coefficients of the predictor method is that they are bounded. The previous discussion has been concerned with establishing convergence and therefore it is not surprising that the results are weaker than the normal requirements for a predictor

method. In the practical case where a high rate of convergence is required more restrictive conditions are necessary for the coefficients of the predictor formula. (e.g. See Henrici (1962)).

2.3 Application of the Theory to a Parabolic Partial Differential Equation

2.3.1 Introduction

In the case of partial differential equations the formulation and analysis become more complicated since the definition of a problem may involve conditions on several boundaries. In particular some methods of stability analysis can only be applied to restricted classes of problem with special types of boundary conditions and simple shapes of domain.

In order to apply these methods of stability analysis it is necessary to replace the boundary conditions by periodicity conditions and this topic is discussed in Section 3.4. In this section stability conditions will be derived both for the case where periodicity conditions can replace the boundary conditions and also for the case where boundary conditions are present.

The analysis presented here goes beyond that required as a simple illustration of the general theory since it is convenient to discuss the problem fully at this point. In this simple problem stability can be established for all $\eta \in E_h$ and all $\Lambda \in E_h^o$ so that the stability regions are the entire spaces E_h and E_h^o.

2.3.2 Formulation of the Problem

The general problem is described by the equations

$$\frac{\partial y}{\partial t} = \sum_{l=1}^{s} \frac{\partial^2 y}{\partial x^l \partial x^l} + g(x,t) \qquad \begin{matrix} 0 \leq t \leq T \\ 0 < x^l < 1 \end{matrix} \qquad 34(a)$$

$$y(x,0) = f(x) \qquad 0 \leq x^l \leq 1 \qquad l = 1, 2, \ldots s \qquad 34(b)$$

where x is a vector of dimension s with components x^l, l = 1, 2, ... s. The definition of the problem is completed by the formula describing the boundary conditions which typically give the value of a linear combination of the function y and its first derivative on the boundary. In the more complex problems the solution y may be a vector of dimension d but in this section attention will be restricted to the case s = d = 1 with boundary conditions given by

$$y(0,t) = p_1(t) \qquad 34(c)$$

$$0 \leq t \leq T$$

$$y(1,t) = p_2(t) \qquad 34(d)$$

The functions considered will be members of a space E consisting of functions which are twice continuously differentiable with respect to t and four times continuously differentiable with respect to x and have mixed derivatives up to these orders in x and t which are continuous in the region

$$D \equiv \{x,t: 0 \leq x \leq 1, 0 \leq t \leq T\} \qquad (35)$$

This will simplify the discussion of consistency although consistency can be established for a wider class of functions if required.

The space E^o of the general theory is formed as the product space $C_{2,1}[D]_x C[D_0]_x C[D_1]_x C[D_1]$ where $C_{2,1}$ denotes a space of functions twice continuously differentiable with respect to x and continuously differentiable with respect to t

$$D_0 \equiv \{x : 0 < x < 1\}$$

$$D_1 \equiv \{t : 0 \leq t \leq T\} \qquad (36)$$

The equations (34) can be represented by the mapping

$$(Fu)_1 \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - g(x,t) \qquad (x,t) \in D \qquad 37(a)$$

$$(Fu)_2 \equiv u(x,0) - f(x) \qquad x \in D_0 \qquad 37(b)$$

$$(Fu)_3 \equiv u(0,t) - p_1(t) \qquad t \in D_1 \qquad 37(c)$$

$$(Fu)_4 \equiv u(1,t) - p_2(t) \qquad t \in D_1 \qquad 37(d)$$

$$u \in E$$

Corresponding to the spaces E and E^o there are the finite spaces E_h and E_h^o which have vector elements of dimension (M+1)x(N+1). A simple discretization is defined by setting up the mesh

$$D^h \equiv \left\{ \begin{array}{l} x_j, t^n : x_j = jh, j = 0, 1, \ldots M \\ \qquad\qquad t^n = nk, n = 0, 1, \ldots N \end{array} \right\} \qquad (38)$$

where M and N are integers $h = 1/M$, $k = T/N$, and setting $(\Delta_h u)_j^n = u(x_j, t^n)$ $j = 0, 1, \ldots M$; $n = 0, 1, \ldots N$. As shown by equation (37) an element $w \in E^o$ has four components which we discretize as follows

$$(\Delta_h^o w)_j^n = w_1(x_j, t^n) \qquad j = 1, 2, \ldots M-1; \quad n = 1, 2, \ldots N \qquad 39(a)$$

$$(\Delta_h^o w)_j^o = w_2(x_j) \qquad j = 1, 2, \ldots M-1 \qquad 39(b)$$

$$(\Delta_h^o w)_o^n = w_3(t^n) \qquad n = 0, 1, \ldots N \qquad 39(c)$$

$$(\Delta_h^o w)_M^n = w_4(t^n) \qquad n = 0, 1, \ldots N \qquad 39(d)$$

The finite difference solution, denoted by Y, is found from the equation $F^h Y = 0$ where a general finite difference scheme is used which can be represented by the mapping F^h defined by

$$(F^h \eta)_j^n = \frac{\eta_j^n - \eta_j^{n-1}}{k} - \frac{\theta\delta^2 \eta_j^n + (1-\theta)\delta^2 \eta_j^{n-1}}{h^2} - g(x_j, t^n) \qquad 40(a)$$

$$j = 1, 2, \ldots M-1 \quad , \quad n = 1, 2, \ldots N$$

$$(F^h \eta)_j^o = \eta_j^o - f(x_j) \qquad j = 1, 2, \ldots M-1 \qquad 40(b)$$

$$(F^h \eta)_o^n = \eta_o^n - p_1(t^n) \qquad n = 0, 1, \ldots N \qquad 40(c)$$

$$(F^h \eta)_M^n = \eta_M^n - p_2(t^n) \qquad n = 0, 1, \ldots N \qquad 40(d)$$

$\delta^2 \eta_j^n = \eta_{j-1}^n - 2\eta_j^n + \eta_{j+1}^n$, θ is a constant $0 \leq \theta \leq 1$ and we let $\lambda = k/h^2$ be constant.

The norm used in initial value problems usually takes the form

$$\| \eta \| = \max_n \| \eta^n \| \qquad \eta \in E_h \tag{41}$$

where $\| \eta^n \|$ will be the maximum norm or the p-norm as defined in Section 2.1. A similar definition is used in E_h^o.

2.3.3 Consistency

It will be shown below that for all mesh points the consistency term $(F^h \Delta_h u - \Delta_h^o F u)_j^n$ has a limiting value of zero. It follows immediately that $\| F^h \Delta_h u - \Delta_h^o F u \|$ has a limiting value of zero as $h \to 0$ in the case of the maximum norm. In the case of the p-norm we have

$$\| w^n \| = \left(h \sum_{j=0}^{M} \left(|(F^h \Delta_h u - \Delta_h^o F u)_j^n| \right)^p \right)^{1/p}$$

$$\leq \left(h.(M+1). \max_j |(F^h \Delta_h u - \Delta_h^o F u)_j^n|^p \right)^{1/p}$$

Since $h(M+1) = 1 + h$ then if $(F^h \Delta_h u - \Delta_h^o F u)_j^n$ has limit zero as $h \to 0$ it follows that $\| w^n \| \to 0$ for all n and $\lim_{h \to o} \| w \| = 0$. Thus a sufficient conditions for u-consistency in either the maximum norm or the p-norm is that at each mesh point $\lim_{h \to o} |(F^h \Delta_h u - \Delta_h^o F u)_j^n| = 0$.

At the boundary points $n = 0$ or $j = 0, M$ we have $(F^h \Delta_h u - \Delta_h^o Fu)_j^n \equiv 0$. For $j = 1, 2, \ldots M-1$, $n = 1, 2, \ldots N$ we have

$$(F^h \Delta_h u - \Delta_h^o Fu)_j^n \equiv \frac{u(x_j,t^n) - u(x_j,t^{n-1})}{k} - \left(\frac{\partial u}{\partial t}\right)_{x = x_j}^{t=t^n}$$

$$- \theta[u(x_{j-1},t^n) - 2u(x_j,t^n) + u(x_{j+1},t^n)]/h^2$$

$$- (1-\theta)\,[u(x_{j-1},t^{n-1}) - 2u(x_j,t^{n-1}) + u(x_{j+1},t^{n-1})]/h^2$$

$$+\left(\frac{\partial^2 u}{\partial x^2}\right)_{x=x_j}^{t=t^n}$$

$$= k[-u_{tt}(x_j,\tau_1) + (1-\theta)u_{txx}(x_j,\tau_2)]$$

$$- \frac{\theta h^2}{24}[\phi(\xi_1,t^n) + \phi(\xi_2,t^n)] - \frac{(1-\theta)h^2}{24}[\phi(\xi_3,t^{n-1}) + \phi(\xi_4,t^{n-1})] \qquad (42)$$

where $t^{n-1} \leq \tau_i \leq t^n$ $\qquad$ $x_{j-1} \leq \xi_i \leq x_{j+1}$.

and $\phi \equiv \dfrac{\partial^4 u}{\partial x^4}$

Since $k = \lambda h^2$ where λ is a constant then the right hand side of equation (42) has limit zero as $h \to 0$ and hence

$$\lim_{h \to o} \max_{n,j} (F^h \Delta_h u - \Delta_h^o Fu)_j^n = 0$$

Therefore the scheme is u-consistent both in the maximum norm and in the p-norm.

2.3.4 Stability

2.3.4.1 Introduction

The stability equation $G^h Z = \Lambda$ derived from equations (40) are given by

$$Z_j^o = \Lambda_j^o \qquad j = 1, 2, \ldots M-1 \qquad 43(a)$$

$$Z_o^n = \Lambda_o^n \qquad n = 0, 1, \ldots N \qquad 43(b)$$

$$Z_M^n = \Lambda_M^n \qquad n = 0, 1, \ldots N \qquad 43(c)$$

and at interior points

$$-\lambda\theta Z_{j-1}^n + (1+2\lambda\theta) Z_j^n - \lambda\theta Z_{j+1}^n = \lambda(1-\theta) Z_{j-1}^{n-1} + (1+2\lambda(1-\theta)) Z_j^{n-1} + \lambda(1-\theta) Z_{j+1}^{n-1} + k\Lambda_j^n \qquad 43(d)$$

$$= \psi_j^n$$

$j = 1, 2, \ldots M-1 \qquad n = 1, 2, \ldots N$

The most frequently used method of stability analysis for partial differential equation problems is the Fourier method but this cannot be applied directly to equations (43). In order to apply the Fourier method it is necessary that the problem is of pure initial value type or that the boundary conditions can be replaced by periodicity conditions as described in Section 3.4. The alternative approach to stability analysis presented in Section 3.6 is also restricted to the periodic case. We will first consider the problem of equation (40) transformed to

periodic type and show how the theory based on circulant matrices can be used to find stability conditions both for the p-norm, p = 2 and for the maximum norm. The full stability equations (43) will then be considered but in this case the sufficient conditions for stability which are obtained are more restrictive than for the periodic case. It should be noted that in practice a difference scheme including boundary conditions is used for computation but Fourier analysis is used to investigate stability. Thus the theoretical results which are devised do not apply to the computation scheme which is used. In practice the results derived by Fourier methods appear to give satisfactory stability criteria but it should be realized that strictly speaking these methods should only be applied to computation schemes which are based on periodic properties.

2.3.4.2 Stability in the Periodic Case

Following the arguments of Section 3.4 the problem can be transformed to one of periodic type in which case the stability equations are all of similar type to equation 43(d) and can be represented by

$$AZ^n = BZ^{n-1} + k\,\Lambda^n \qquad n = 1, 2, \ldots N \qquad (44)$$
$$Z^o = \Lambda^o$$

where Z^n and Λ^n are vectors of dimension 2M and A and B are matrices of circulant form (see Appendix 1) with elements defined by the first rows $1 + 2\lambda\theta$, $-\lambda\theta$, 0 ... 0, $-\lambda\theta$

and $1-2\lambda(1-\theta)$, $\lambda(1-\theta)$, 0, ... 0, $\lambda(1-\theta)$ respectively. The extension of Z^n to dimension 2M is required in order to produce a problem of periodic type.

A simple necessary and sufficient conditions can easily be derived and this is then investigated for the case of the p-norm, p=2 and for the maximum norm. From equation (44)

$$Z^n = A^{-1}BZ^{n-1} + k\,A^{-1}\,\Lambda^n$$

and letting $C = A^{-1}B$ and $\tilde{\Lambda}^n = A^{-1}\,\Lambda^n$

$$Z^n = C^n\,Z^o + k \sum_{j=1}^{n} C^{n-j}\,\tilde{\Lambda}^j \qquad (45)$$

Theorem 1

If, for all M, $\| A^{-1}w \| \leq L \| w \|$ for an arbitrary vector w of dimension 2M where L is constant independent of M, then the necessary and sufficient conditions for stability is that there exists a constant K such that

$$\| C^n(k) \| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \qquad (46)$$

Proof

The sufficiency is established first. From equation (45) using the uniform boundedness of $\|A^{-1}\|$ we have

$$\| Z^n \| \leq \| C^n(k) \| . \| Z^o \| + k \sum_{j=1}^{n} \| C^{n-j} \| .L. \| \Lambda^j \|$$

and using the norm $\| \Lambda \| = \max_n \| \Lambda^n \|$ and the condition (46)

we have

$$\| Z^n \| \leq K \| \Lambda \| + k.n.K.L. \| \Lambda \|$$

$$\leq K[1+T.L] \| \Lambda \|$$

Since it is true for all n we have

$$\| Z \| \leq K[1+T.L] \| \Lambda \| \qquad (47)$$

and condition (46) is sufficient for stability.

The necessity of condition (46) is established by considering the situation where $\Lambda^i = 0$, $i \neq 0$ in which case

$$Z^n = C^n(k) \Lambda^o$$

We assume condition (46) is not satisfied and show that this leads to a contradiction. If condition (46) is not satisfied then there exists a sequence S of $C^n(k)$ such that $\| C^n(k) \| \geq M$ where M can be as large as we please. Now for any $\| C^n(k) \| = Q$ there exists a Λ^o such that $\| C^n(k) \Lambda^o \| = Q \| \Lambda^o \|$ and a sequence of Λ^o is chosen corresponding to the sequence S of $C^n(k)$ so that $\| Z^n \| = \| C^n(k) \| . \| \Lambda^o \| \geq M \| \Lambda \|$ which contradicts condition (46). Hence condition (46) is necessary for stability.

Stability in the 2-norm

The analysis of stability involves the comparison of $\| Z \|$ with $\| \Lambda \|$ and for convenience the common factor $h^{\frac{1}{2}}$ in the norm

$$\| \eta^n \| = h^{\frac{1}{2}} \left(\sum_{j=1}^{2M} | \eta_j^n |^2 \right)^{\frac{1}{2}} \qquad (48)$$

can be omitted. The norm used will then be the standard Euclidean vector norm. The matrix $C = A^{-1}B$ is a product of two circulant matrices and hence C and its integer powers C^n are also circulant in form. Since a circulant matrix is normal the matrix norm subordinate to the Euclidean vector norm will equal the spectral radius for any circulant matrix. Hence

$$\| C^n(k) \|_2 = (\max_j | \mu_j |)^n \qquad \text{for } n = 1, 2, \ldots N \qquad (49)$$

For this matrix and for the 2-norm we show that the necessary and sufficient condition $\| C^n(k) \| \leq K$ reduces to the condition

$$| \mu_j | \leq 1 \quad \text{for } j = 1, 2, \ldots 2M, \text{ for all } M \qquad (50)$$

The necessity is established first. We note that if an eigenvalue $\mu_{j_o} = \delta$ for a matrix C of dimension 2Mo then there is an eigenvalue equal to δ for all matrices C of dimension 2rMo where r is a positive integer. This follows by considering the eigenvalues of C which are given by the results of Appendix 1.

$$\mu_j = \frac{1-2\lambda(1-\theta)+\lambda(1-\theta)\left[e^{2\pi ji/2M} + e^{-2\pi ji/2M}\right]}{1+2\lambda\theta-\lambda\theta\left[e^{2\pi ji/2M} + e^{-2\pi ji/2M}\right]}$$

$$= \frac{1-4\lambda(1-\theta)\sin^2(\pi j/2M)}{1+4\lambda\theta\sin^2(\pi j/2M)} \qquad (51)$$

$j = 1, 2, \ldots 2M$

Note that

$$\mu_j \equiv \mu_j(M) = F(j/M) \qquad (52)$$

for $j = 1, 2, \ldots 2M$ and hence for $r = 1, 2, \ldots$ we have

$$\mu_j(M) = \mu_{rj}(rM)$$

Thus if for some matrix C of dimension 2Mo and some $j = j_o$

$$\mu_{j_o}(Mo) = \delta$$

then for every matrix C of dimension 2r.Mo $r = 1, 2, \ldots$ there is an eigenvalue $\mu_{rj_o}(rMo) = \delta$ where δ is independent of h and k.

Let us assume that the scheme is stable but that condition (50) does not hold and show that this leads to a contradiction. If condition (50) does not hold then there exists j_o and Mo such that

$$|\mu_{j_o}(Mo)| = \delta > 1$$

Let the r.Mo x r.Mo matrix C have an eigenvector v(r) corresponding to the eigenvalue $\mu \equiv \mu_{rj_o}(r.Mo) = \mu_{j_o}(Mo)$.

Then taking a sequence of $\Lambda^o(r) = v(r)$, $\Lambda^i = 0$, $i \neq 0$

we have

$$Z^n(r) = C^n(r)v(r) = \mu^n v(r)$$

$$n = 0, 1, \ldots N \quad \text{where } N = T/k = T(rMo)^2/\lambda$$

Thus in particular $Z^N(r) = \mu^N v(r)$

$$\| Z^N(r) \| = |\mu|^N \| v(r) \| = \delta^N \| v(r) \|$$

$$\| Z(r) \| = \max_n \| Z^n(r) \| = \delta^N \| v(r) \|$$

$$= \delta^N \| \Lambda(r) \| \tag{53}$$

For a sufficiently large value of r we have $\delta^N > K$ for any K and hence

$$\| Z(r) \| > K \| \Lambda(r) \| \qquad r > R$$

which contradicts the stability condition

$$\| Z \| \leq K \| \Lambda \|$$

Hence condition (50) is necessary for stability.

The sufficiency of condition (50) follows from equation (49). If $|\mu_j| \leq$ then

$$\| C^n(k) \|_2 = (\max_j |\mu_j|)^n$$

$$\leq 1 \tag{54}$$

which gives condition (46) of the theorem. It must also be shown that $\| A^{-1} \Lambda^n \| \leq L \| \Lambda^n \|$ to establish the theorem.

Let the eigenvalues of A be ν_j, $j = 1, 2, \ldots 2M$. Since A and A^{-1} are circulant matrices and hence normal we have $\| A^{-1} \| = \max_j | \nu_j^{-1} |$. Since $\nu_j = 1 + 4\lambda\theta \sin^2(\pi j/2M)$ then it follows that $\| A^{-1} \| \leq 1$ and

$$\| A^{-1} \Lambda^n \| \leq \| \Lambda^n \| \tag{55}$$

Hence condition (50) is sufficient to establish condition (46) of the Theorem 1, and this gives a stable scheme.

Condition (50) on μ_j is equivalent to a condition on the parameters λ and θ. Since μ_j never exceeds one condition (50) will hold if $\mu \geq -1$, i.e., the condition reduces to

$$1-4\lambda(1-\theta)\sin^2(\pi j/2M) \geq -1 - 4\lambda\theta \sin^2(\pi_j/2M)$$

$$2\lambda(1-2\theta)\sin^2(\pi j/2M) \leq 1 \qquad \text{for } j = 1, 2, \ldots 2M \tag{56}$$

Since this must hold for all M we have the necessary and sufficient condition for stability

$$\begin{array}{ll} 2\lambda(1-2\theta) \leq 1 & 0 < \theta < \frac{1}{2} \\ \text{No restriction} & \frac{1}{2} \leq \theta \leq 1 \end{array} \tag{57}$$

which is the condition obtained by Richtmyer using Fourier analysis.

Stability in the Maximum Norm

It will first be shown that condition (50) namely $|\mu_j| \leq 1$ for $j = 1, 2, \ldots 2M$ for all M is a necessary condition for stability in the maximum norm

$$\| u^n \| = \max_j | u_j^n | \qquad (58)$$

$$\| u \| = \max_n \| u^n \|$$

The argument of the previous section is followed; it is assumed that condition (50) does not hold and it is shown that this contradicts the stability condition. As in the previous section we have

$$Z^N(r) = \mu^N v(r) \qquad ,N = T(rMo)^2/\lambda$$

$$\| Z(r) \| = \delta^N \| \Lambda(r) \|$$

and for a sufficiently large value of r we have $\delta^N > K$ for any K and hence

$$\| Z(r) \| > K \| \Lambda(r) \| \quad \text{for any K if } r > R$$

which contradicts the stability condition $\| Z \| \leq K \| \Lambda \|$ Hence condition (50) is necessary for stability in the maximum norm.

The sufficient conditions for stability in the maximum norm are stronger than condition (50) and some preliminary definitions and results are required. The families of matrices

A and B which form $C = A^{-1}B$ in equation (45) have a special property which is defined below. For such matrices it can be shown that the eigenvalues of C are given by taking values of a function g(x) which is independent of M. Lemma 1 shows that $\| C^p \|$ will be uniformly bounded with respect to p if the positive integer powers of g(x) have absolutely convergent Fourier series with a bound which is independent of the power p. Lemma 3.4 (Stetter) investigates the problem described by equation (44) and gives conditions on λ and θ so that the conditions on $g^p(x)$ are satisfied and lemma 1 will hold. A further lemma establishes that $\| A^{-1}w \| \leq L \| w \|$ for an arbitrary vector w, where L is a constant independent of M and together with lemma 1 this gives sufficient conditions for stability according to Theorem 1.

Definition of a Standard Family

The following lemmas are concerned with families of related circulant matrices which we will call 'standard' families in accordance with the following definition.

Definition

A family Q of circulant matrices D^M of dimension $M \geq M_1$ with elements d^M_{jk} will be called a 'standard' family if there exist constants M_1 and j_1 such that for every $D^M \in Q \qquad M \geq M_1$ there is a row $j_1 \leq M_1$ such that $d^M_{j_1 k} = d^{M_1}_{j_1 k} \quad k = 1, 2, \ldots M_1$

$$d^M_{j_1 k} = 0 \qquad M_1 < k \leq M \tag{59}$$

Expressing this less formally the definition means that every matrix of a standard family has in each row a group of constant elements which appear in a banded fashion about the diagonal and the remainder of the matrix has zero elements.

The Eigenvalues of C

The periodic problems involve a matrix $C^{2M} = (A^{2M})^{-1} B^{2M}$ where A^{2M} and B^{2M} each belong to a standard family $A^{2M} \varepsilon Q_1$, $B^{2M} \varepsilon Q_2$. In this case we can show that the eigenvalues of C^{2M} are given by the values of a function g(x), which is independent of M, at a set of equally spaced values of the argument x. Since a number of superscripts and subscripts are required on the matrices the superscript 2M will be omitted in the subsequent analysis where no confusion can arise.

Let the eigenvalues of C be denoted by μ_j and the elements of the defining rows of the standard families Q_1 and Q_2, $A \varepsilon Q_1$, $B \varepsilon Q_2$ be denoted respectively by $q^1_{-r_1}, \ldots q^1_{-1}, q^1_0, q^1_1, \ldots q^1_{s_1}$

$$q^0_{-r_2}, \ldots q^0_{-1}, q^0_0, q^0_1 \ldots q^0_{s_2}$$

where subscript 0 denotes the diagonal elements. The eigenvalues of C are the solutions of $\mu AX = BX$ and using arguments similar to those of Appendix 1, this gives

$$\mu_j = \frac{\sum_{k=-r_1}^{s_1} q^0_k \, e^{2\pi kj\, i/2M}}{\sum_{k=-r_2}^{s_2} q^1_k \, e^{2\pi kj\, i/2M}} \qquad (60)$$

$j = 1, 2, \ldots 2M$

These eigenvalues are given by the values $\mu_j = g(x_j)$, $x_j = j/2M$, $j = 1, 2, \ldots 2M$ where

$$g(x) = \frac{\sum_{k=-r_1}^{s_1} q_k^0 e^{2\pi kxi}}{\sum_{k=-r_2}^{s_2} q_k^1 e^{2\pi kxi}} \qquad 0 \leq x \leq 1 \qquad (61)$$

Lemma 1

If a family of circulant matrices C^{2M} $(M = M_1, M_1 + 1, \ldots)$ are formed from two standard families Q_1 and Q_2 by the formula

$$C^{2M} = (A^{2M})^{-1} B^{2M}, \; A^{2M} \in Q_1, \; B^{2M} \in Q_2.$$

the eigenvalues of C^{2M} are given by

$$\mu_j = g(j/2M) \quad , \quad j = 1, 2, \ldots 2M$$

where $g(x) = \dfrac{\sum_{k=-r_1}^{s_1} q_k^0 e^{2\pi ikx}}{\sum_{k=-r_2}^{s_2} q_k^1 e^{2\pi ikx}} \qquad 0 \leq x \leq 1$

and $g(x)$ is independent of M.

If $g(x)$ and all its positive integer powers have absolutely convergent Fourier Series for $0 \leq x \leq 1$

i.e.

$$g^p(x) = \sum_{l=-\infty}^{+\infty} \bar{a}_l^{(p)} e^{2\pi i l x} \tag{62}$$

with $\sum_{l=-\infty}^{+\infty} |\bar{a}_l^{(p)}| \leq K, \quad K$ independent of p

it follows that for an arbitrary vector w

$$\| C^p w \| \leq K \| w \| \tag{63}$$

where $\| w \| = \max_i | w_i |$

i.e. The maximum norm of C^p is uniformly bounded.

Proof

Consider the eigenvector v_j corresponding to the eigenvalue μ_j. If $a_l^{(p)}$, $l = 1, 2, \ldots 2M$ are the elements of the first row of C^p then using the results of Appendix 1 we have

$$[C^p v_j]_1 = \sum_{l=1}^{2M} a_l^{(p)}(2M) e^{2\pi i (l-1)/2M}$$

$$= \mu_j^p [v_j]_1 = \mu_j^p \qquad j = 1, 2, \ldots 2M$$

Hence the powers of the function g(x) defined by equation (61) are also given at the mesh points by

$$g^p(x_j) = \sum_{l=1}^{2M} a_l^{(p)}(2M) e^{2\pi i x_j (l-1)}$$

We define a function

$$G_{2M}^{(p)}(x) = \sum_{l=1}^{2M} a_l^{(p)}(2M)\, e^{2\pi i x(l-1)} \qquad (64)$$

This function coincides with $g^p(x)$ defined by equation (61) at the mesh points $x = x_j \equiv j/2M$, $j = 1, 2, \ldots 2M$ and provides a Fourier interpolation formula for $g^p(x)$. By the assumptions of the lemma there also exists an absolutely convergent Fourier series for $g^p(x)$

$$g^p(x) = \sum_{l=-\infty}^{+\infty} \bar{a}_l^{(p)}\, e^{2\pi i l x} \qquad 0 \leq x \leq 1 \qquad (65)$$

where $\sum_{l=-\infty}^{+\infty} |\bar{a}_l^{(p)}| \leq K < \infty$, where K is independent of p.

It can easily be shown that the following relation exists between the coefficients of the Fourier interpolation formula and the Fourier series coefficients (following similar arguments to those of Hamming (1962) p.279).

$$a_{l+1}^{(p)}(2M) = \sum_{k=-\infty}^{+\infty} \bar{a}_{l+2Mk}^{(p)} \qquad (66)$$

Hence

$$\sum_{l=0}^{2M-1} |a_{l+1}^{(p)}(2M)| = \sum_{l=0}^{2M-1} \left| \sum_{k=-\infty}^{+\infty} \bar{a}_{l+2Mk}^{(p)} \right|$$

$$= \sum_{l=-\infty}^{+\infty} |\bar{a}_l^{(p)}| \leq K \qquad (67)$$

Since $||C^p||_{\max} = \sum_{l=1}^{2M} |a_l^{(p)}(2M)| \leq K$, K independent of p

then C^p is uniformly bounded in the maximum norm and the lemma is proved.

The conditions on $g^p(x)$ which are required to satisfy lemma 1 will now be considered. This problem has been investigated by Stetter (1963) and the lemma 3.4 of his paper is quoted here.

Lemma 3.4 (Stetter)

If $|g(\phi_k)| = 1$ for $k = 1, 2, \ldots k_1$, $\phi_k \varepsilon (-\pi, +\pi]$ and $|g(\phi)| < 1$ otherwise then for condition (65) to hold it is sufficient that $g(\phi)$ is twice continuously differentiable in $[-\pi, +\pi]$ and

$$\left.\begin{aligned} |g(\phi)| &\leq 1 - M_k(\phi - \phi_k)^{2N_k} \\ |g''(\phi)| &\leq M_k''(\phi - \phi_k)^{2N_k - 2} \end{aligned}\right\} \quad \text{in} \quad |\phi - \phi_k| \leq \delta_k \tag{68}$$

where N_k(integer), M_k and M_k'', and δ_k are positive constants. The function $g(x)$, $0 \leq x \leq 1$ can be transformed to the range required by the lemma by letting $\phi = \pi(2x - 1)$ then

$$g(x) = g\left(\frac{\phi + \pi}{2\pi}\right) = \mathfrak{g}(\phi) .$$

Firstly we require that $|\mathfrak{g}(\phi)| \leq 1$, which is similar to condition (50), and any additional conditions required to satisfy (68) must also be investigated. For the example studied in equation (44) using equation (61)

$$g(x) = \bar{g}(\phi) = 1 - \frac{4\lambda \sin^2(\frac{\phi + \pi}{2})}{1 + 4\lambda\theta \sin^2(\frac{\phi + \pi}{2})} \qquad (69)$$

and $|\bar{g}(\phi)| = 1$ when $\phi = \pi$, or when $2\lambda(1-2\theta).\sin^2(\frac{\phi + \pi}{2}) = 1$

The differentiability properties required by the lemma can be seen to be satisfied since

$$\bar{g}'(\phi) = \frac{-4\lambda \sin(\frac{\phi + \pi}{2}) \cos(\frac{\phi + \pi}{2})}{[1+4\lambda\theta \sin^2(\frac{\phi + \pi}{2})]^2} \qquad (70)$$

and

$$\bar{g}''(\phi) = \frac{2\lambda \sin^2(\frac{\phi + \pi}{2}) - 2\lambda \cos^2(\frac{\phi + \pi}{2}) + 16\lambda^2\theta \sin^2(\frac{\phi + \pi}{2})}{[1+4\lambda\theta \sin^2(\frac{\phi + \pi}{2})]^3} \qquad (71)$$

Note that for a real function the second condition of (68) is automatically satisfied if the first condition holds. We consider the point $\phi = \pi$ where $\bar{g}(\phi) = 1$ and expand $\bar{g}(\pi + 2\delta)$ by Taylor Series

$$\bar{g}(\pi + 2\delta) = 1 - \frac{2\lambda\delta^2}{2} + \frac{K(\delta)\,\delta^3}{3!}$$

where $K(\delta) = \bar{g}'''(\pi + 2\delta p)$, $0 \leq p \leq 1$ is uniformly bounded $|K(\delta)| \leq K^1$.

Hence $|\bar{g}(\pi + 2\delta)| \leq 1 - \delta^2[\lambda - K^1.\delta]$

If δ is chosen so that $K^1.\delta < \lambda/2$ then

$$|\bar{g}(\pi + 2\delta)| \leq 1 - \delta^2.\ \lambda/2 \qquad (72)$$

and condition (68) is satisfied.

At the point ϕ where $2\lambda(1-2\theta)\sin^2\left(\frac{\phi+\pi}{2}\right) = 1$, $\bar{g}(\phi) = -1$.

Let $(\phi + \pi)/2 = \psi$. Then

$$\bar{g}(\psi+\delta) = 1 - \frac{4\lambda\sin^2\psi}{1+4\lambda\theta\sin^2\psi} + \delta\left[\frac{-4\lambda\sin\psi\cos\psi}{(1+4\lambda\theta\sin^2\psi)^2}\right] + \delta^2 K_1(\delta)$$

where $K_1(\delta) = \bar{g}(\psi + p\delta)$, $0 \leq p \leq 1$ is uniformly bounded $K_1(\delta) \leq K_1^1$. Since $\sin^2\psi = \frac{1}{2\lambda(1-2\theta)}$

$$\bar{g}(\psi+\delta) = 1 - \frac{4\lambda}{2\lambda} \pm \frac{4\lambda\delta\sqrt{2\lambda(1-2\theta)-1}}{4\lambda^2} + K_1(\delta)\delta^2 \tag{73}$$

Hence condition (68) cannot be satisfied for this value of ϕ. If the condition on g(x) is strengthened to

$$-1 < g(x) \leq 1 \qquad 0 \leq x \leq 1 \tag{74}$$

then the only point to be considered in applying lemma 3.4 is the point $\phi = \pi$, $x = 1$ and we have shown that the conditions of the lemma are satisfied at this point and therefore lemma 1 also holds.

To establish the conditions of Theorem 1 it must also be shown that there exists a constant L independent of M such that $\|A^{-1}w\| \leq L\|w\|$ where w is an arbitrary vector. This can be done by using some of the arguments of lemma 1 connecting Fourier interpolation coefficients and Fourier series coefficients.

Let $\nu(x)$ be the function defining the eigenvalues of A^{-1}

i.e. $\nu_j = \nu(x_j)$, $x_j = j/2M$, $j = 1, 2, \ldots 2M$ where

$$\nu(x) = \frac{1}{1+4\lambda\theta \sin^2 \pi x} \tag{75}$$

This function has an absolutely convergent Fourier series

$$\nu(x) = \sum_{l=-\infty}^{+\infty} \bar{a}_l^{(-1)} e^{2\pi l x i} \qquad 0 \leq x \leq 1 \tag{76}$$

Following the line of argument of lemma 1 we also have

$\nu_j = g_{2M}^{(-1)}(x_j)$ where $x_j = j/2M$, $j = 1, 2, \ldots 2M$

$$g_{2M}^{(-1)}(x) = \sum_{l=0}^{2M-1} a_{l+1}^{(-1)}(2M)\, e^{2\pi l x i} \qquad 0 \leq x \leq 1 \tag{77}$$

where $a_{l+1}^{(-1)}(2M)$ are the elements of the first row of A^{-1}.
Using the relation (67) between the coefficients of (76) and (77) we have

$$\sum_{l=0}^{2M-1} | a_{l+1}^{(-1)} | \leq \sum_{l=-\infty}^{+\infty} | \bar{a}_l^{(-1)} | \leq L \tag{78}$$

and since $|| A^{-1} || = \sum_{l=0}^{2M-1} | a_{l+1}^{(-1)} |$ we have

$$|| A^{-1} w || \leq L || w || \text{ , for an arbitrary vector w.} \tag{79}$$

The condition (79) together with condition (74)

$$-1 < g(x) \leq 1 \qquad 0 \leq x \leq 1,$$

which gives the condition for lemma 1 to hold, are sufficient conditions for stability in the maximum norm according to

Theorem 1. Condition (74) reduces to a condition on λ and θ in the same way as for the 2-norm giving

$$\begin{array}{ll} 2\lambda(1-2\theta) < 1 & 0 < \theta < \frac{1}{2} \\ \text{No restriction} & \frac{1}{2} \leq \theta \leq 1 \end{array} \qquad (80)$$

Thus for the sufficient conditions $2\lambda(1-2\theta) = 1$ is not included for the maximum norm but is included for the 2-norm in condition (57).

2.3.4.3. Stability in the Presence of Boundary Conditions

In this case the full stability equations given by equations (43) must be considered, i.e.,

$$Z_j^o = \Lambda_j^o \qquad j = 1, 2, \ldots M-1 \qquad 81(a)$$

$$Z_o^n = \Lambda_o^n \qquad n = 0, 1, \ldots N \qquad 81(b)$$

$$Z_M^n = \Lambda_M^n \qquad n = 0, 1, \ldots N \qquad 81(c)$$

and at interior points

$$-\lambda\theta Z_{j-1}^n + (1+2\lambda\theta) Z_j^n - \lambda\theta Z_{j+1}^n =$$

$$\lambda(1-\theta) Z_{j-1}^{n-1} + (1-2\lambda(1-\theta)) Z_j^{n-1} + \lambda(1-\theta) Z_{j+1}^{n-1} + k\Lambda_j^n \qquad 81(d)$$

$$j = 1, 2, \ldots M-1 \qquad n = 1, 2, \ldots N$$

These equations can also be written in matrix form as

$$AZ^n = BZ^{n-1} + \lambda\theta \begin{bmatrix} \Lambda_o^n \\ o \\ \vdots \\ o \\ \Lambda_M^n \end{bmatrix} + \lambda(1-\theta) \begin{bmatrix} \Lambda_o^{n-1} \\ o \\ \vdots \\ o \\ \Lambda_M^{n-1} \end{bmatrix} + k\Lambda^n$$

$n = 1, 2, \ldots N$

and $Z^o = \Lambda^o$

where Z^n and Λ^n are vectors of dimension M - 1 corresponding to the internal points Z_j^n, Λ_j^n, $j = 1, 2, \ldots M-1$ and the matrices A and B are defined by $A = P(\lambda\theta)$, $B = P(-\lambda(1-\theta))$ where P(q) is the symmetric tridiagonal matrix of dimension M-1 with diagonal elements 1+2q and off-diagonal elements -q. If this equation (82) is compared with equation (44) for the periodic case it is seen that there are two significant differences. The matrices A and B are tridiagonal instead of circulant and there are extra terms in $\lambda\theta$ and $\lambda(1-\theta)$ in equation (82). It is these extra terms which necessitate more stringent sufficient conditions than for the periodic case. For the investigation of necessary condition we let $\Lambda^i = 0$, $i \neq 0$ and the extra terms are then no longer present and the proof for necessary conditions follows the same lines as the proof in the periodic case. In this section necessary conditions and sufficient conditions for stability in the maximum norm will be established.

Necessary conditions will be considered first.

Let $\Lambda^i = 0$, $i \neq 0$ and then equations (82) become

$$Z^n = A^{-1}BZ^{n-1}$$

$$= CZ^{n-1} = C^n\Lambda^o \qquad (83)$$

$n = 1, 2, \ldots N$

where $C = A^{-1}B$.

As for Theorem 1 of the last section the following condition is necessary for stability;

$$\| C^n(k) \| \leq K \quad , \quad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \tag{84}$$

Following the argument of the last section on necessary conditions for stability in the maximum norm we see that condition (84) reduces to

$$| \mu_j | \leq 1 \qquad j = 1, 2, \ldots M-1 \tag{85}$$

This follows since the eigenvalues of C are

$$\mu_j = 1-4\lambda(1-\theta)\sin^2(\pi j/2M)/1+4\lambda\theta\sin^2(\pi j/2M) \tag{86}$$

$$j = 1, 2, \ldots M-1$$

(see Rutherford (1952) for the eigenvalues of P(q) and hence of C) and $\mu_j(M) = \mu_{rj}(r\,M)$ as before. Hence the condition

$$|\mu_j| \leq 1 \qquad j = 1, 2, \ldots M-1$$

is a necessary condition for stability in the maximum norm in this case also. Since μ_j can never exceed unity this reduces to

$$1-4\lambda(1-\theta)\sin^2(\pi j/2M) \geq -1 - 4\lambda\theta\sin^2(\pi j/2M)$$

$$2\lambda(1-2\theta)\sin^2(\pi j/2M) \leq 1$$

for $j = 1, 2, \ldots M-1$. Since this must hold for all M we have

$$\begin{matrix} 2\lambda(1-2\theta) \leq 1 & 0 < \theta < \frac{1}{2} \\ \text{No restriction} & \frac{1}{2} \leq \theta \leq 1 \end{matrix} \tag{87}$$

as in condition (57).

Following similar lines to the discussion in John(1966), Ralston and Wilf(1960) and Godunov and Ryabenki(1964) it can be shown that $2\lambda(1-\theta) \leq 1$ is a sufficient condition for stability in the maximum norm. From equation 81(d)

$$|(1+2\lambda\theta)Z_j^n| \leq 2\lambda\theta \|Z^n\| + |\psi_j^n|$$

(where ψ_j^n denotes the right hand side of equation 81(d)), for $j = 1, 2, \ldots M-1$

and hence taking the value of j for which $|Z_j^n|$ is a maximum on the left hand side we have

$$\|Z^n\| \leq \max_j |\psi_j^n|$$

$$\leq \max_j \left(|\lambda(1-\theta)Z_{j-1}^{n-1} + (1-2\lambda(1-\theta))Z_j^{n-1} + \lambda(1-\theta)Z_{j+1}^{n-1}| + k|\Lambda_j^{n+1}|\right)$$

If $2\lambda(1-\theta) \leq 1$ then

$$\|Z^n\| \leq \|Z^{n-1}\| + k \max_j |\Lambda_j^n|$$

$$\leq \|Z^0\| + n.k\|\Lambda\| \qquad (88)$$

Since this is true for $n = 0, 1, \ldots N$

$$\|Z\| \leq (1+N.k)\|\Lambda\|$$

$$\leq K.\|\Lambda\| \qquad (89)$$

where $K = 1+T$ is independent of M.

Thus sufficient condition for stability in the maximum norm is

$$2\lambda(1-\theta) \leq 1 \qquad (90)$$

which is more restrictive than the comparable condition (80) in the periodic case.

3. Stability Analysis in Partial Differential Equations

3.1 Introduction

This section will consider problems of initial value type, i.e., where the state of the system is given at $t = 0$ and a solution is sought for $0 \leq t \leq T$. In the most general case the solution $y(x,t)$ is a vector of dimension d dependent on the quantity t and an S-dimensional vector x. The simplest type of problem is the pure initial value problem in which no boundary conditions are necessary and the solution is completely defined by the values at $t = 0$. However in other equations, such as the example on the heat conduction equation in Section 2.3, a differential equation is given in an open region D and subsidiary equations describe conditions on the boundary of the region. In the pure initial value problem stability can be determined from the properties of the finite difference scheme which approximates the differential equation. In the case where boundary conditions are also present the problem is complicated by the presence of the finite difference scheme representing the boundary conditions. In some cases the problem with boundary conditions can be shown to be equivalent to a pure initial value problem in which the boundary conditions are replaced by periodicity conditions on the initial function. The problems are equivalent in the sense that the solution of the second problem is identical to the solution of the first problem in the original region and the boundary conditions are also satisfied. Section 3.4 discusses the replacement of boundary conditions by periodicity

conditions and gives examples of problems which can be transformed to equivalent periodic problems. For these periodic problems and for the pure initial value problem the methods of Fourier analysis can be used to investigate stability. It is shown in Section 3.6 that the results traditionally derived by Fourier methods can also be derived by a new approach based on the properties of circulant matrices. If the boundary conditions are to be fully considered the stability of a scheme can sometime be investigated by means of the energy method.

The stability of partial differential equations has been discussed by Forsythe and Wasow(1960), Godunov and Ryabenki(1964), Richtmyer(1957), Richtmyer and Morton(1967) and Wendroff(1966) and the relationship of the general scheme to other frameworks is included in this section. The work of Godunov and Ryabenki (abbreviated to G-R in the later text) was one of the starting points for the development of the general theory of Section 2.1 and therefore fits easily into this framework. The G-R treatment is restricted to linear differential equations but the main difference is that the differential equation and the boundary and initial conditions are treated in parallel in the G-R scheme, but they are formulated as one mapping in the general scheme. A typical differential equation is composed of two parts which we represent in abstract form as

$$(F\,u)_1 = 0 \qquad 1(a)$$

$$(F\,u)_2 = 0 \qquad 1(b)$$

where $(F\,u)_1$ represents the differential equation and $(F\,u)_2$ the initial conditions (and boundary conditions if present).

For example

$$(F\,u)_1 \equiv \frac{\partial u}{\partial t} - \frac{\partial u}{\partial x} \qquad 0 \leq t \leq T \qquad 2(a)$$

$$(F\,u)_2 \equiv u(x,0) - f(x) \qquad 0 \leq x \leq 1 \qquad 2(b)$$

The finite difference scheme can similarly be represented by

$$(F^h \eta)_1 = 0 \qquad 3(a)$$

$$(F^h \eta)_2 = 0 \qquad 3(b)$$

where in the example above this could take the form

$$(F^h \eta)_1^{n+1} \equiv \eta_j^{n+1} - \eta_j^n - \frac{\lambda}{2}(\eta_{j+1}^n - \eta_{j-1}^n)$$

$$n = 0, 1, \ldots N-1 \qquad 4(a)$$

$$(F^h \eta)_2 = \eta_j^0 - f(x_j)$$

$$Nk = T,\ \lambda = k/h,\ h = 1/M,\ x_j = hj \qquad 4(b)$$

Godunov and Ryabenki study non-homogeneous equations which have the general form

$$(L\,u)_1 = g(x) \qquad 5(a)$$

$$(L\,u)_2 = \emptyset \qquad 5(b)$$

where the left hand sides are linear operators and $\emptyset$ consists of those terms in the boundary and initial conditions which cannot be included in the linear operator, e.g., f(x) in equation 2(b). The discussion of consistency then involves the consistency of the four components of equations (5).

It is clear that equations (5) can be formulated in the form (1) by absorbing the non-homogeneous terms into Fu.

Since L is a linear operator if we form the equation

$$G^h(Z) = \Lambda$$ from equations (5) we have

$$(L\,Z)_1 = \Lambda_1$$

$$(L\,Z)_2 = \Lambda_2$$

which are of the same form as (5). Godunov and Ryabenki define stability if the solutions of these equations satisfy

$$\| Z \| \leq N_1 \| \Lambda_1 \| + N_2 \| \Lambda_2 \|$$

If we define $\| \Lambda \| = \max \left[\max_n \| \Lambda_1^n \|, \| \Lambda_2 \| \right]$ then the G-R definition is equivalent to the condition for S-stability, $\| Z \| \leq N \| \Lambda \|$, $N = \max(N_1, N_2)$ where the stability regions are the entire spaces E_h and E_h^o.

Sections 3.2 and 3.3 give a brief survey of the analysis given by Richtmyer and by Wendroff respectively, and the relationship of each to the general scheme. This is followed by a discussion of the problems which can be made amenable to Fourier analysis. Richtmyer and Wendroff have shown different ways in which Fourier analysis can be used in the study of stability and in both cases the condition for stability is that the powers of the amplification matrix should be uniformly bounded. The two alternative Fourier approaches are presented in Section 3.5 and it is shown in Section 3.6 that, for the

problems where Fourier analysis can be applied, the same result can be derived directly. It is desirable that the condition of uniform boundedness of the powers of the amplification matrix should be expressible in terms of conditions on the amplification matrix itself. A survey of the various conditions which have been formulated is given in Section 3.7 leading to the necessary and sufficient conditions of Kreiss and similar necessary and sufficient conditions due to Buchanan.

3.2 The Lax-Richtmyer Theory

3.2.1 Introduction

In the first and second edition of his book on the initial value problem (Richtmyer (1957), Richtmyer and Morton (1967)), Richtmyer formulates an abstract problem which embraces many problems in partial differential equations. The fundamental properties involved in the convergence of the finite difference scheme are then discussed in relation to the abstract problem and the ideas are applied to a variety of practical examples in the later chapters. The important ideas of the Lax-Richtmyer approach (abbreviated to L-R in the later text) are described briefly here so that the relationship with the general scheme can be discussed.

The type of problem involved is one in which the value of the function is defined at the time $t = 0$ and an equation for the derivative is given. Discussion is restricted to those

problems where these two conditions, together with any boundary conditions, define a unique solution in some domain for $0 \leq t \leq T$. For example a simple problem without boundary conditions is defined by

$$\frac{\partial u}{\partial t} = \frac{\partial u}{\partial x} \qquad 0 \leq t \leq T \qquad 6(a)$$

$$u(x,0) = \sin x \qquad 0 \leq x \leq 1 \qquad 6(b)$$

or as an example of a problem with boundary conditions the following equation could be considered

$$\frac{\partial u}{\partial t} = \frac{\partial^2 u}{\partial x^2} \qquad \begin{matrix} 0 < t < T \\ 0 < x < 1 \end{matrix} \qquad 7(a)$$

$$u(x,0) = f(x) \qquad 0 \leq x \leq 1 \qquad 7(b)$$

$$u(0,t) = p_1(t) \qquad 0 \leq t \leq T \qquad 7(c)$$

$$u(1,t) = p_2(t) \qquad 0 \leq t \leq T \qquad 7(d)$$

It should be noted that boundary conditions should be compatible with the initial conditions at the points on the boundaries where $t = 0$. The application of the general theory to the above problem (7) is described in detail in Section 2.3 and the ideas of consistency, stability and convergence are illustrated in a practical context. In order to cover a wide variety of problems the L-R theory is formulated in terms of an abstract problem which embraces many simple problems such as those described above. The solution of the above problems depends upon the one-dimensional variable x but in the L-R theory this

is generalized to vector $x = [x_1, \ldots x_s]$. The solution u may also be a vector quantity and this generalization enables higher order equations to be represented in first order form. In this case there will be a set of differential equations defining the system of dimensionality equal to that of u. For example a pure initial value problem involving the wave equation can be described in this manner by the equations

$$\frac{\partial v}{\partial t} = \frac{\partial w}{\partial x} \qquad 0 \leq t \leq T \qquad 8(a)$$

$$\frac{\partial w}{\partial t} = \frac{\partial v}{\partial x} \qquad 8(b)$$

$$u(x,0) = f(x) \qquad 0 \leq x \leq 1 \qquad 8(c)$$

$$v(x,0) = g(x) \qquad 0 \leq x \leq 1 \qquad 8(d)$$

The following sections show how the concepts of Lax and Richtmyer relate to those of the framework of Section 2.1 and the example on the parabolic equation given in Section 2.3 is used for illustrative purposes where appropriate. A significant difference between the two approaches is that the L-R theory is concerned with functions defined for all values of x whereas the direct approach of Section 3.6 and the example in Section 2.3 are concerned with finite dimensional spaces in which the vector elements have components which correspond to the values calculated in a finite difference scheme. For the L-R theory it is therefore necessary to create extended functions defined for all x from the functions composed of mesh values which a

finite difference scheme would produce. Two methods of forming this extension are demonstrated but the theory is then developed in general terms to keep in line with the L-R theory. It should be noted that although the L-R theory allows any reasonable extension to be used, when stability is being investigated for real examples using Fourier methods, the extension which is implied is always that which is formed by letting the finite difference operator apply at all points including non-mesh points. A series of papers has been written by Eisen investigating some of the differences which arise when a mesh norm is used rather than a norm involving values for all x in a given range. Eisen only consider the extension described above, which is used in the Fourier methods, and some comments are given on use of other extended functions in the example studied by Eisen.

3.2.2 The General Formulation of Lax and Richtmyer

The various problems which the theory describes are included in a general formulation

$$\frac{d}{dt}(y(t)) = A\,y(t) \qquad 0 \leq t \leq T \qquad 9(a)$$

$$y(0) = f \qquad 9(b)$$

where y and f are vector quantities of dimension d. Attention is restricted to the case where the operator A is a linear operator and when boundary conditions are present the domain of A is restricted to those functions satisfying the boundary conditions. In the simple presentation the boundary conditions

must be linear and homogeneous and equation 9(a) does not contain a non-homogeneous term. However, it is also shown in Richtmyer and Morton (1967) that a development of the theory will extend the standard results to the case where a non-homogeneous term is present in 9(a). It is also shown that the problem with non-homogeneous boundary conditions can be transformed to the standard form (9) with a non-homogeneous term present.

The simpler problem (9) will be used in this section for comparison with the scheme of Section 2.1. A further restriction of the L-R scheme is that all problems are assumed properly posed. This property is fully discussed in the two texts previously mentioned; essentially a well posed scheme must satisfy two conditions. Firstly, if there exists an initial function for which a genuine solution does not exist then this initial function can be approximated as close as we please by an initial function which does give rise to a genuine solution. Secondly, the solution must be such that $\|u(t)\| < K \|f\|$ for all t in $0 \leq t \leq T$ and for all f such that a genuine solution exists where K is independent of t, and $\|u(t)\|$, $\|f\|$ are suitably defined norms. If this second condition is not satisfied it can be shown that a consistent scheme cannot be stable.

3.2.3 Relation with the General Scheme

In developing the L-R theory use is made of theorems in functional analysis based on the properties of Banach spaces. The problem is formulated so that the quantities y(t) are elements of a Banach space B and the completed solution y is a family of elements y(t) each of which is a member of the space B. The space E of the general theory would then be formed of elements $u \equiv \{u(t);\ 0 \leq t \leq T\}$ where $u(t) \in B$ with a suitable norm in E such as

$$||u|| = \max_t ||u(t)||$$

Thus the problem (9) can be represented as a single equation $F(y) = 0$ where 0 is the null element of E^o and $F: E \to E^o$ is a mapping defined by

$$(F\,u)_1 \equiv \frac{du}{dt} - Au \qquad 0 \leq t \leq T \qquad 10(a)$$

$$(F\,u)_2 \equiv u(0) - f \qquad 10(b)$$

$$u \in E$$

The finite dimensional spaces E_h and E_h^o are avoided in the L-R theory by forming extended functions from the mesh values so that the extended functions, defined for all x, are elements of the space B. The extended functions could be formed, for example, by linear interpolation or by applying the finite difference scheme at all points x; the latter method is restricted to problems which do not involve boundary conditions.

Formulation in Terms of Finite Dimensional Spaces

In this section a one-step method with homogeneous boundary conditions will be formulated firstly in terms of mesh functions and secondly in terms of the L-R extended functions, and the example of Section 2.3 will be used to illustrated the discussion. The finite difference equations for the above type of problem relating the values at time level t^{n+1} to those at time level t^n can be written

$$B_1 Y^{n+1} = B_o Y^n \qquad n = 0, 1, \ldots N-1 \qquad 11(a)$$

$$Y^o = \Delta h f \qquad 11(b)$$

where Y^n is a component of the finite difference solution which has dimension $\tilde{M}$ and the finite difference solution $Y \varepsilon E_h$ is given by

$$Y \equiv \{Y^n : n=0,1, \ldots N;\ t^n = nk,\ k=T/N\}$$

Since B_1 must be invertible if the system is solvable we can write this

$$Y^{n+1} = B_1^{-1} B_o Y^n \equiv C Y^n \qquad n = 0, \ldots N-1$$

The dimension of the matrices is dependent on the value of h and in general the elements may also depend upon h and k. If we assume that some fixed relation exists between h and k then the equation above can be written

$$Y^{n+1} = C(k) Y^n \qquad n = 0, 1, \ldots N-1 \qquad (12)$$

These equations can also be represented in the usual way by a mapping

$$(F^h \eta)^{n+1} \equiv \frac{\eta^{n+1} - C(k)\,\eta^n}{k} \qquad n = 0, 1, \ldots N-1 \qquad 13(a)$$

$$(F^h \eta)^0 \equiv \eta^0 - \Delta_h f \qquad 13(b)$$

where $y(x,0) = f(x)$. The element in Eh is a vector

$$\eta = \{ \eta^n : n = 0, 1, \ldots N, \quad t^n = nk, \; k = T/N \}$$

The dimension of the vector Y^n and the form of the matrices B_1 and B_0 depend on the particular problem considered and this is illustrated below by two finite difference schemes, one which is applicable to the solution of the initial value problem defined by equations (6) and the other of which is suitable for the problem with boundary conditions given by equations (7).

The problem defined by equations (6) has a finite difference solution within a triangular region given by the points (x,t) with co-ordinates (0,0), (0,1), $(\frac{1}{2}, \frac{\lambda}{2})$, where $\lambda = k/h$ is a constant and $\lambda < 1$. Thus the vector Y^n will have dimension $\tilde{M} = M+1-2n$, $n = 0, 1, \ldots M/2$, $h = 1/M$ and M is taken to be even for convenience in notation. A finite difference scheme for this problem is defined by

$$Y_j^{n+1} = Y_j^n + \frac{\lambda}{2}\left[Y_{j+1}^n - Y_{j-1}^n\right] + \frac{\lambda^2}{4}\left[Y_{j+1}^n - 2Y_j^n + Y_{j-1}^n\right] \qquad (14)$$

$$n = 0, 1, \ldots N-1, \qquad j = n, n+1, \ldots M+1-n$$

or, in terms of the matrices B_1 and B_0 we have $B_1(M,n) = I$, (I is the identity matrix of dimension M+1-2n) and $B_0(M,n)$

is a (M-1-2n) x (M+1-2n) matrix given by

$$\begin{bmatrix} \frac{-\lambda}{2}+\frac{\lambda^2}{4} & 1-\frac{\lambda^2}{2} & \frac{\lambda}{2}+\frac{\lambda^2}{4} & & 0 \\ & \frac{-\lambda}{2}+\frac{\lambda^2}{4} & 1-\frac{\lambda^2}{2} & \frac{\lambda}{2}+\frac{\lambda^2}{4} & \\ & & \cdots\cdots\cdots & & \\ 0 & & \frac{-\lambda}{2}+\frac{\lambda^2}{4} & 1-\frac{\lambda^2}{2} & \frac{\lambda}{2}+\frac{\lambda^2}{4} \end{bmatrix} \qquad (15)$$

Note that the matrix $B_0(M_1 n) \equiv C(k)$ is not a square matrix but has a simple form since it is a circulant matrix (see Appendix 1) and the elements of every row are given by three expressions the values of which are independent of k.

For the problem defined by equations (7) the finite difference scheme which was used in Section 2.3 and was given by equations (2.40) will be considered. In the simplest case where $p_1(t) \equiv p_2(t) \equiv 0$ equations (2.40) can be represented in the form $B_1 Y^{n+1} = B_0 Y^n$, $n = 0, 1, \ldots N-1$ where Y^n has dimension $\tilde{M} = M+1$, $h = 1/M$ and the $(\tilde{M} \times \tilde{M})$ matrices B_0 and B_1 are given by

$$B_1 = \begin{bmatrix} 1 & 0 & & & \\ -r\theta & 1+2r\theta & -r\theta & & 0 \\ & -r\theta & 1+2r\theta & -r\theta & \\ & & \ddots & \ddots & \ddots \\ & & & -r\theta & 1+2r\theta & -r\theta \\ 0 & & & & 0 & 1 \end{bmatrix}$$

16(a)

$$B_0 = \begin{bmatrix} 0 & 0 & & & \\ r(1-\theta) & 1-2r(1-\theta) & r(1-\theta) & & 0 \\ & r(1-\theta) & 1-2r(1-\theta) & r(1-\theta) & \\ & & \ddots & \ddots & \ddots \\ & & & r(1-\theta) & 1-2r(1-\theta) & r(1-\theta) \\ 0 & & & & 0 & 0 \end{bmatrix}$$

16(b)

The matrices no longer have a simple form since they contain rows which represent the approximation to the differential equation and also rows which represent the approximation to the boundary conditions. It will be shown in Section 3.6 that the Fourier methods of stability analysis cannot be applied to such a problem since these methods can only be used when the finite difference equations have the same form at all points.

Formulation in Terms of L-R Extended Functions

The finite difference equations associated with equation (9) can also be used as a basis for the formulation of the L-R extended functions and two possible methods of extension are discussed below for the finite difference scheme defined by equations (12).

The extended function can be formed by linear interpolation between two calculated mesh values by the formula

$$Y^{n+1}(x_{j+p}) \equiv (1-p)Y_j^{n+1} + pY_{j+1}^{n+1} \qquad (17)$$

$x_s \equiv s.h$, $0 \leq p \leq 1$, and if values of x exist at the ends of the range which do not lie between two mesh points then linear extrapolation must be used in equation (17), i.e., $-1 < p \leq 0$ at the lower end of the range and $1 \leq p < 2$ at the upper end.

In certain problems such as that described by equations (14) the finite difference scheme is defined by a set of equations which have the same form at all mesh points except those on the initial line. For these problems the extended operator can be formed by extending this finite difference equation to apply at all points (x,t) within the domain of determination D^h of the difference scheme. For example, in the case of equations (14) the extended function $Y^{n+1}(x)$ is given by the equations

$$Y^{n+1}(x) = Y^n(x) + \frac{\lambda}{2}[Y^n(x+h) - Y^n(x-h)] + \frac{\lambda^2}{4}[Y^n(x+h) - 2Y^n(x) + Y^n(x-h)] \qquad (18)$$

This type of problem can be expressed in general form as follows. The finite difference equations can be represented by a mapping with two components which represent the approximation to the initial conditions and the approximation to the differential equation at the mesh points. It is the form of the approximation to the differential equation which is of interest.

A typical approximation to the differential equation relates to the values of a function at a set of points N_1 in the neighbourhood of the point (x_j, t^{n+1}) to a set of points N_2 in the neighbourhood of the point (x_j, t^n) giving an equation

$$(F^h \Delta_h u)_j^n = \frac{1}{k} \left[\sum_{N_1} b_{1p} \, u(x_{j+p}, t^{n+1}) - \sum_{N_2} b_{0p} \, u(x_{j+p}, t^n) \right] \tag{19}$$

and the extended function which we denote by $\bar{F}^h$ can be defined be extending this relation to all values of x and t within the domain D^h, i.e.,

$$\bar{F}^h \, u(x,t) \equiv \frac{1}{k} \left[\sum_{N_1} b_{1p} \, u(x+hp, t+k) - \sum_{N_2} b_{0p} \, u(x+hp, t) \right] \tag{20}$$

Using an extension such as that described by equations (18) or (20) to define extended operators B_1^* and B_0^* gives an equation similar in form to (11), i.e.,

$$B_1^* \, Y^{n+1}(x) = B_0^* \, Y^n(x) \qquad n = 0, 1, \ldots N-1 \tag{21(a)}$$

$$Y^0 = \Delta_h f \tag{21(b)}$$

Assuming $(B_1^*)^{-1}$ exists this gives rise to a bounded linear transformation which may depend on k

$$C^*(k) = B_1^* B_0^*$$

and the equation (21) can be written

$$Y^{n+1}(x) = C^*(k)\, Y^n(x) \qquad (22)$$

$$n = 0, 1, \ldots N-1$$

$$Y_0 = \Delta_h f$$

3.2.4 Consistency

As shown in the examples of Section 2 consistency is usually established by assuming the functions concerned have continuous derivatives up to a certain order. However, in the L-R theory Banach spaces are used and the elements involved may be the limit points of the space and therefore may not possess the required differentiability properties. This necessitates a more complex definition of consistency as follows.

Definition

The family $C^*(k)$ of operators provides a consistent approximation to the initial value problem (9) if for every u(t) in some class U of genuine solutions whose initial elements u(o) are dense in B

$$\lim_{k \to o} \left\| \left(\frac{C^*(k) - I}{k} - A \right) u(t) \right\| = 0$$

$$0 \leq t \leq T \qquad (23)$$

For problems without boundary conditions it will be shown that the condition of u-consistency is essentially the same as the L-R condition of consistency. However, if the problem includes boundary conditions the formulation of Section 2.1 is more satisfactory.

In most finite difference approximations to a differential equation a formula is used which will be consistent at all points (x,t) for any function $u(x,t)$ which possesses suitable differentiability properties. This can be expressed more formally by means of a simple condition which will be sufficient for u-consistency or for L-R consistency. The formulation of the problem given in equations (10) is considered and we let $F\,u(x,t)$ equal the value of $\frac{du}{dt} - A\,u$ at the point (x,t). An extension of equation (13a) is formed in the manner described by equations (19) and (20) so that $\bar{F}^{h}\,u(x,t)$ equals the value of

$$\frac{u(t+k) - C^{*}(k)\,u(t)}{k}$$

at the point (x,t).

A Sufficient Condition for Consistency

$$\text{If} \quad \lim_{k \to 0} \left| \bar{F}^{h}\,u(x,t) - Fu(x,t) \right| = 0 \qquad (24)$$

for all (x,t) within the domain of the differential equation D then the finite difference scheme is L-R consistent and u-consistent.

In the finite dimensional case the consistency error on the initial line is zero and therefore the discussion of consistency can be restricted to mesh points with co-ordinates

x_j, t^n within the domain D^h and using condition (24) we have, at every mesh point,

$$\lim_{k \to o} \left| (F^h \Delta_h u)_j^n - (\Delta_h^o F u)_j^n \right| = 0$$

where Δ_h and Δ_h^o are defined by

$$(\Delta_h u)_j^n = u(x_j, t^n) \quad , \quad (\Delta_h^o w)_j^n = w(x_j, t^n).$$

Hence using the p-norm or the maximum norm defined in Section 2.1.2 and letting $\| (F^h \Delta_h u - \Delta_h^o F u)^n \| = \phi^n$ we have

$$\lim_{k \to o} \phi^n = 0 \qquad \text{for } n = 0, 1, \ldots N$$

and $\lim_{k \to o} \| F^h \Delta_h u - \Delta_h^o F u \| = \lim_{k \to o} \max_n \phi^n = 0$

Thus condition (24) is sufficient for u-consistency in the finite dimensional case.

In order to establish the L-R consistency we consider

$$\left\| \left(\frac{C^*(k) - I}{k} - A \right) u(t) \right\| =$$

$$\left\| \frac{u(t+k) - C^*(k)u(t)}{k} - \left(\frac{du}{dt} - Au(t) \right) - \frac{u(t+k) - u(t)}{k} + \frac{du}{dt} \right\|$$

$$\leq \left\| \frac{u(t+k) - C^*(k)u(t)}{k} - \left(\frac{du}{dt} - A u(t) \right) \right\| + \left\| \frac{u(t+k) - u(t)}{k} - \frac{du}{dt} \right\|$$

(25)

The first term is equivalent to

$$\| \bar{F}^h u(t) - F u(t) \|$$

and using condition (24) it is seen that this term has limit zero as $k \to o$ for all reasonable norms and for all values of t, $0 \leq t \leq T$. The second term has limit zero as $k \to o$ since the functions u under consideration are continuously differentiable. Hence condition (24) is sufficient to ensure that a scheme is L-R consistent.

The finite dimensional formulation can easily be extended to the case where boundary conditions are present. In this case the mapping F would be composed of four parts

$$(F\,u)_1 \equiv \frac{du}{dt} - Au \qquad 0 \leq t \leq T \qquad 26(a)$$

$$(F\,u)_2 \equiv u(o) - f \qquad 26(b)$$

and components representing the boundary conditions for example

$$(F\,u)_3 \equiv u(o,t) - p_1(t) \qquad 0 \leq t \leq T \qquad 27(a)$$

$$(F\,u)_4 \equiv u(1,t) - p_2(t) \qquad 0 \leq t \leq T \qquad 27(b)$$

and the mapping F^h would have four corresponding parts $(F^h \eta)_i$, $i = 1, 2, 3, 4$. The element of E_h^o formed by $F^h \Delta_h u - \Delta_h^o F\,u$ would then have four components χ_i $i = 1, 2, 3, 4$ and the norm in E_h^o could be defined by

$$\|\chi\| = \max_i \|\chi_i\| \qquad (28)$$

A sufficient condition for consistency would then be

$$\lim_{k \to o} \|(F^h \Delta_h u - \Delta_h^o F\,u)_i\| = 0 \qquad i = 1, 2, 3, 4 \qquad (29)$$

It can be seen that if condition (24) holds at all mesh points within the region and on the boundary then for all reasonable norms condition (29) will hold and the scheme will be u-consistent. These ideas have been used in the example of Section 2.3 on the parabolic equation.

3.2.5 Stability

The L-R definition of stability is formulated in terms of the operator $C^*(k)$ of equation (22) and states that the scheme is stable if there exists $\tau > 0$ such that the infinite set of operators

$$C^*(k)^n \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \qquad (30)$$

is uniformly bounded. A condition can be formulated for the finite dimensional problem which has similar form and is necessary and sufficient for S-stability where S is the entire space E_h, i.e.,

$$\| C^n(k) \| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \qquad (31)$$

This has been established in Section 2.3 when discussing the example of the parabolic equation but the argument is repeated here for convenience.

The sufficiency is established first. From equations (17)

$$Z^{n+1} = C(k)Z^n + k\,\Lambda^{n+1} \qquad n = 1, 2, \ldots N-1 \qquad 32(a)$$

$$Z^o = \Lambda^o \qquad 32(b)$$

and this gives $$Z^{n+1} = C^{n+1}(k)Z^{o} + k \sum_{j=o}^{n} C^{n-j} \Lambda^{j+1} \qquad (33)$$

Hence $$|| Z^{n+1} || \leq K|| \Lambda^{o} || + n.k.K.\max_{n} || \Lambda^{n} ||$$

$$\leq K[1+nk] || \Lambda ||$$

Since this is true for all n we have

$$|| Z || \leq K[1+T] || \Lambda || \qquad (34)$$

so that (31) is a sufficient condition for convergence.

The necessity of condition (31) is established by considering the situation where $\Lambda^{i} = 0$, $i \neq 0$ in which case

$$Z^{n} = C^{n}(k) \Lambda^{o} \qquad (35)$$

We assume that condition (31) is not satisfied and show that this leads to a contradiction. If condition (31) is not satisfied then there exists a sequence S of $C^{n}(k)$ such that $||C^{n}(k) || \geq M$ where M can be as large as we please. Now for any $|| C^{n}(k) || = Q$ there exists a Λ^{o} such that $|| C^{n}(k) \Lambda^{o} || = Q || \Lambda^{o} ||$ and a sequence of Λ^{o} is chosen corresponding to the sequence S of $C^{n}(k)$ so that

$$|| Z^{n} || = || C^{n}(k) || . ||\Lambda^{o} || \geq M || \Lambda || \text{ for any } M$$

which contradicts condition (31). Hence condition (31) is necessary for stability.

3.2.6 Convergence

In the L-R definition of convergence all values of t are considered rather than only the points on the sequence of meshes. The scheme is said to be convergent if there exists a series of increments k_j tending to zero and a series of mesh points $n_j k_j$ such that

$n_j k_j \to t$ as $j \to \infty$ and

$$\lim_{j \to \infty} \| C^*(k_j)^{n_j} u^o - u(t) \|_B = 0 \qquad (36)$$

for all t, $0 \leq t \leq T$ and for every u^o in B. The corresponding definition of Section 2.1 has a very similar form, i.e.,

$$\lim_{k \to o} \| \Delta_h y - Y \|_{E_h} = 0 \qquad (37)$$

and if the norm $\|\eta\|_{E_h} = \max_n \| \eta^n \|$ is used equation (37) becomes

$$\lim_{k \to o} \max_n \| C^n(k) \eta^o - y(t^n) \| = 0 \qquad (38)$$

Thus condition (36) is more demanding than (38) in two ways. It requires convergence for any point t and it requires convergence of the extended operator rather than the mesh operator.

3.2.7 Discussion on Papers by Eisen

In a series of papers Eisen (1966), (1967a) and (1967b) has shown that in certain cases schemes which are unstable and non-convergent in the L-R sense can be shown to be stable and convergent if the norms considered are defined in terms of mesh values only. However, Eisen only considers the extension to the mesh operator which is formed by taking the operator which is applied at the point (x_j, t^n) and extending this to any point (x,t). He does not consider other extensions such as linear interpolation or similar methods. It must be emphasised that the method of extending the function from its mesh values determines whether the results in one framework are equivalent to those in the other. If the example from Eisen (1967a) is taken it is easy to show that a simple modification of the extended operator will give stability and convergence in the L-R scheme. Since the extension of the mesh function by different methods can lead to differing results, the discussion of convergence in terms of a mesh norm which can be calculated from the computational values has much to recommend it.

Eisen's example can be considered as a mapping defined by

$$(F\,u)_1 \equiv \frac{\partial u}{\partial t} \qquad 0 \leq t \leq T,\ 0 \leq x \leq 1 \qquad 39(a)$$

$$(F\,u)_2 \equiv u(x,0) - f(x) \qquad 0 \leq x \leq 1 \qquad 39(b)$$

and a corresponding finite difference scheme

$$(F^h \eta)_j^{n+1} = \frac{\eta_j^{n+1} - \eta_j^n}{k} - \frac{h\,\eta_j^n}{x_j} \qquad n = 0, \ldots N-1 \quad j = \tfrac{1}{2}, \tfrac{3}{2} \ldots M-\tfrac{1}{2} \qquad 40(a)$$

$$(F^h \eta)_j^o = \eta_j^o - f(x_j) \qquad j = 0, \tfrac{1}{2}, \tfrac{3}{2}, \ldots M-\tfrac{1}{2} \qquad 40(b)$$

$$(F^h \eta)_o^{n+1} = \eta_o^{n+1} - \eta_o^n \qquad n = 0, \ldots N-1 \qquad 40(c)$$

where $x_j = h.j$; $j = 0, \frac{1}{2}, \ldots M-\frac{1}{2}$; $h = 1/M$; and for simplicity $k = h$. The operator $C(k)$ on the finite mesh is represented by the diagonal matrix with elements

$$[1, \quad 1+2k, \quad 1+\tfrac{2}{3}k \quad \ldots \quad 1 + {}^{2k}/_{2M-1}]$$

and Eisen takes $C^*(k)$ to be defined as $(1 + \frac{kh}{x})$ at any point $x \neq 0$. Using the p norm with p = 1 and testing stability

$$\| Z^1 \| = \operatorname*{Lim}_{\varepsilon \to o} \int_{\varepsilon}^{1} \left| (1 + \tfrac{kh}{x})\, Z^o(x) \right| dx$$

$$\geq \min_x \left| Z^o(x) \right| \operatorname*{Lim}_{\varepsilon \to o} \int_{\varepsilon}^{1} \left| (1 + \tfrac{kh}{x}) \right| dx$$

$$= K_1 \operatorname*{Lim}_{\varepsilon \to o} \left| [x + kh \log x]_{\varepsilon}^{1} \right| \qquad (41)$$

Assuming $K_1 = \min_x \left| Z^o(x) \right| > 0$ we can see that for any $k \neq 0$ $\| Z^1 \|$ is unbounded. Since for a typical Λ $\| \Lambda^o \|$ is bounded then $C^*(k)$ as defined above gives an unstable scheme.

It should be noted that Lax and Richtmyer assume that the operator $C^*(k)$ is bounded and so Eisen is considering a case

which the L-R theory does not attempt to cover. However, it is clear that if $C^*(k)$ is unbounded then the scheme is unstable and hence non-convergent. In fact the norm of the error after one step is unbounded.

However, if we choose a different extension

i.e., $$C^*(k) = (1+kh/x), \quad x \geq h/2$$

$$C^*(k) = 0 \qquad 0 \leq x < h/2 \tag{42}$$

which gives the same operator C(k) at the mesh points it is shown below that this gives a stable and hence convergent scheme.

Assume that $\delta(c)$ is calculated exactly and gives no consistency error.

$$Z^{n+1} = C^*(k)Z^n + k \Lambda^{n+1}$$

$$= (C^*(k))^{n+1} Z^o + k \sum_{j=o}^{n} (C^*(k))^j \Lambda^{n+1-j}$$

$$Z^o = \Lambda^o$$

$$\| Z^{n+1} \| \leq \int_0^1 \left| C^*(k)^{n+1} \Lambda^o(x) + k \sum_{j=o}^{n} C^*(k)^j \Lambda^{n+1-j}(x) \right| dx$$

$$\leq \int_{h/2}^1 (1+\frac{kh}{x})^{n+1} \left| \Lambda^o(x) \right| dx + k \sum_{j=o}^{n} (1+\frac{kh}{x})^j \left| \Lambda^{n+1-j}(x) \right| dx$$

$$+ \int_0^1 \left| \Lambda^1(x) \right| dx$$

$$\leq \left(1+\frac{2kh}{h}\right)^N \int_{h/2}^{1} | \Lambda^o(x) | dx + kN \max_n | \Lambda^n(x) | dx$$

$$+ \int_0^1 | \Lambda^1(x) | dx$$

$$\leq e^{2T} (2+T) \| \Lambda \| \qquad \text{since } N.k = T \qquad (43)$$

which gives a stable scheme.

The stability in the L-R sense of a scheme which is known to be stable in terms of a mesh norm is thus shown to be dependent on the method of forming the extended operator.

3.3 The Wendroff Theory

3.3.1 The Formulation of the Problem

Another approach is presented by Wendroff (1966) in which the emphasis is on the discrete space. Since the definition of convergence used in section 2.1 is concerned with norms in the space E_h the discussion of convergence by Wendroff is much closer to the ideas of section 2.1 than the L - R approach. However since Wendroff does not define an equivalent of the spaces E and E^o it is not possible to compare the Wendroff scheme and the general scheme in this respect.

The general form of the problem discussed is

$$\frac{\partial u}{\partial t} = f(x,t,u,u_x) \qquad 44(a)$$

$$u(x,o) = \phi(x) \qquad 0 \leq x \leq 1 \qquad 44(b)$$

Note that this is a pure initial value problem without boundary conditions. The finite difference scheme proposed can be written as a mapping as follows

$$(F^h \eta)^{n+1} \equiv \frac{\eta^{n+1} - Sn\,\eta^n - k\,Gn(\eta^n)}{k} \qquad n = 0, 1, \ldots N-1 \qquad 45(a)$$

$$(F^h \eta)^o_j \equiv \eta^o_j - \phi(x_j) \qquad j = 0, 1, \ldots M \qquad 45(b)$$

$$x_j = jh, \; h = 1/M, \quad k = T/N$$

Sn is a linear operator whose coefficients may depend upon the values of x_j and t^n corresponding to the mesh point η^n_j and the form of the operator considered is

$$Sn = \sum_{p=-J}^{+J} C_p(x,t^n)T^p \qquad (46)$$

where T is the translation operator defined by $Tu_j = u_{j+1}$ and G_n is a non-linear operator which must be such that there exists a bounded

linear operator $\delta Gn(u,U)$ satisfying

$$Gn(u) - Gn(U) = \delta Gn(u,U)(u-U) \tag{47}$$

i.e. for all u and U

$$\| \delta Gn(u,U)v \| \leq K \| v \|$$

where K is a constant independent of u and U. The norm is defined in the usual way

$$\| \delta Gn(u,U) \| = \sup_{v \neq o} \frac{\| \delta Gn(u,U)v \|}{\| v \|} \tag{48}$$

The discretization used by Wendroff is

$$u_j(t) = u(x_j,t) , \quad x_j = x' + h_j , \quad j = 0, \pm 1, \pm 2, \ldots$$

and this can be extended to form Δ_h by defining

$$(\Delta_h u)^n_j = u(x_j,t^n), \quad t^n = nk , \quad k = T/N$$

Wendroff considers a space of functions Ω_h defined as follows. Ω_h is the set of all mesh functions

$$u(t) \equiv \{u_j(t) : j = 0, \pm 1, \pm 2, \ldots \} \tag{49}$$

which satisfy the condition $u_j(t) = 0$ for $| x_j | \leq x_{max}(t)$, where $x_{max}(t)$ depends on the particular problem. In the general case $u_j(t)$ may be a vector of dimension d. A scalar product is defined in Ω_h by

$$(u,v) = \sum_{j=-\infty}^{+\infty} u_j v_j \tag{50}$$

The relation between Wendroff's approach and the general scheme can be demonstrated by introducing spaces E and E_h in the following manner. The elements of the space E are functions u(x, t) possessing suitable differentiability properties and satisfying the following

conditions in order that

$$(\Delta_h u)^n \equiv \{ (\Delta_h u)_j^n : j = 0, \pm 1, \pm 2, \ldots \}$$

shall belong to Ω_h

i.e. $$| u(x,t) | = 0 \quad \text{if} \quad | x | > x_{max}(t)$$

and

$$\int_{-x_{max}(t)}^{x_{max}(t)} | u(x,t) |^2 \, dx < \infty \tag{51}$$

The norm in E can then be defined by

$$\| u \|^2 = \sup_t \int_{-x_{max}(t)}^{x_{max}(t)} | u(x,t) |^2 dx \tag{52}$$

The elements η of the space E_h are defined by

$$\eta = \{ \eta^n : n = 0, 1, \ldots N \}$$

where $$\eta^n = \{ \eta_j^n : j = 0, \pm 1, \pm 2, \ldots \}$$

and $(\eta^n, \eta^n) < \infty$. The norm in E_h can then be defined by

$$|| \eta || = \max_n [h(\eta^n, \eta^n)]^{\frac{1}{2}} \tag{53}$$

which is consistent with the norm defined in E. The framework can easily be extended to the case where x is a vector of dimension S in which case $j = [j_1, j_2, \ldots j_s]$ and each component can range over all integer values. Then $x_j = x^1 + h\, j$ where x, x^1, and j are vectors of dimension S and h is a diagonal matrix with elements $h_1, \ldots h_s$.

3.3.2 Consistency

In the development of the theory Wendroff restricts his attention to the case of an explicit finite difference scheme in which the

solution at (x_j, t^{n+1}) depends only on the values at (x_{j-1}, t^n), (x_j, t^n), (x_{j+1}, t^n). Equation 45(a) is then applicable at all points within the triangle Δ defined by the points (0, 0), (0, 1), $(\frac{1}{2}, \lambda/2)$, $\lambda = k/h$. The application of equation (45) to the true solution gives the equation used by Wendroff to define the consistency error, i.e.

$$F^h y(t^{n+1}) \equiv \frac{y(t^{n+1}) - Sn\, y(t^n) - k\, Gn(y(t^n))}{k} \equiv \frac{E^{n+1}}{k}$$

$$n = 0, 1, \ldots N-1 \qquad 54(a)$$

$$F^h y(o) \equiv 0 \qquad 54(b)$$

Wendroff defines the scheme to be consistent if $E^{n+1} = O(k^{p+1})$ where $p > 0$. Since a sufficient condition for y- consistency is that $E^{n+1} = O(k)$ then if a scheme is consistent by Wendroff's definition this implies the scheme is y- consistent.

3.3.3 Stability and Convergence

In formulating the equations which are used in the discussion of stability and convergence Wendroff combines the equation $F^h(Y) = 0$ derived from equations (45) with the equation $F^h(\Delta_h y) = E/k$ derived from equations (54). The stability equation considered is therefore

$$F^h(\Delta_h y) - F^h(Y) = E/k \qquad (55)$$

or expressed in terms of the Wendroff formulation

$$Z^{n+1} = [Sn + k\, \delta Gn(y(t^n), Y^n)]Z^n + E^{n+1} \qquad 56(a)$$

$$n = 0, 1, \ldots N-1$$

$$Z^o = E^o \qquad 56(b)$$

where $Z^n = y(t^n) - Y^n$. The discussion of stability is therefore restricted to y- stability and the only values of Λ which are considered on the right hand side are those which represent consistency errors

for particular values of N. In fact the stability conditions given by Wendroff enable the argument to be applied to the case of general η and Z as can be seen from the following discussion of the Wendroff approach.

Equations (56) are only defined inside the triangle Δ and if the difference scheme is to produce Z^n which are members of Ω_h the equations must be extended to give

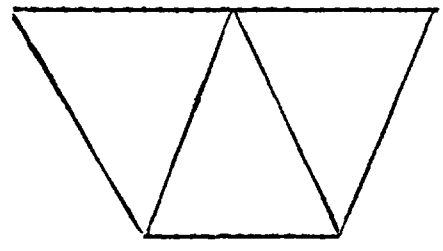
$$Z_j^n \quad \text{for } j = 0, \pm 1, \ldots \qquad \text{with} \quad | Z_j^n | = 0$$

for $| x_j | \geq x_{max}$. E^n is defined by equation (54) within the domain of dependence of the differential scheme. It is assumed that Δ lies within this domain. The extension of E^n is defined by letting E^n be given by (54) inside and on Δ and $E^n = 0$ outside Δ. Since S_n and δG_n may depend on x and t they may be undefined outside Δ but this can be avoided by making the values independent of x outside Δ, e.g. at the point x_j

$$Sn = \sum_{p=-1}^{+1} C_p(x_j', t^n)T^p \qquad (57)$$

where the value of $x_j' = x_j$ if $(x_j, t^n) \in \Delta$ or x_j' is replaced by the value of x at the nearest boundary. The operator δG_n can treated in a similar manner. The quantities on the right of equation (56) are now defined for all x_j and it is easily seen that all Z_j^n are zero outside the region show below and hence $Z^n \in \Omega_h$. Note that $x_{max} = 3/2$.

$(-\frac{1}{2}, \lambda/2)$ $(3/2, \lambda/2)$



$(0,0)$ $(1,0)$

Equation (56) gives

$$Z^{n+1} = \left[\prod_{i=0}^{n} S_i + k\delta G_i(y(t^i), Y^i) \right] Z^0 \qquad 58(a)$$

$$+ \sum_{m=0}^{n} \left[\prod_{i=m+1}^{n} S_i + k\delta G_i(y(t^i), Y^i) \right] E^{m+1}$$

$$Z^0 = E^0 \qquad 58(b)$$

where $\prod_{i=n+1}^{n} S_i + k\delta G_i$ is defined as the identity operator. Wendroff defines a scheme to be stable if there exists a constant K independent of k such that

$$\left\| \prod_{i=m+1}^{n} (S_i + k\delta G_i) \right\| \leq K \qquad -1 \leq m \leq n \leq N \qquad (59)$$

This condition will be satisfied if there exists a uniform bound for $\delta G_i(u, v)$ and for $\prod_{i=m}^{n} S_i$, $m \leq n$ (see Strang 1964). Using condition (59) in equation (58) gives

$$\| Z^{n+1} \|_{\Omega_h} \leq K \| Z^0 \|_{\Omega_h} + \frac{KT}{k} \| E \|_{E_h}$$

and Wendroff shows that this leads to the condition

$$\left(\sum_{(x_j, t^{n+1}) \in \Delta} Z^{n+1}(x_j)^2 \cdot h \right)^{\frac{1}{2}} \leq (K h^{\frac{1}{2}} + KT)\, O(k^p) \qquad (60)$$

where $Z^{n+1} = y(t^{n+1}) - Y^{n+1}$ and hence condition (59) is sufficient for convergence.

Although the Wendroff analysis was carried out for the case $Z = \Delta_h y - Y$ it can easily be shown, by essentially the same argument, that the condition (59) is sufficient for the scheme to possess y-stability. A modified version of equation (55) is considered in which the point Y is replaced by a general point $\Delta_h y + Z$,

i.e., $F^h(\Delta_h y + Z) - F^h(\Delta_h y) = \Lambda$

This gives the following equations which have the same form as equations (56)

$$Z^{n+1} = [Sn + k\,\delta Gn(y(t^n) + Z^n,\ y(t^n))]Z^n + k\ \Lambda^{n+1}$$

$$n = 0,\ 1,\ \ldots\ N-1$$

$$Z^o = \Lambda^o$$

Following the argument used previously and using condition (59) gives

$$\| Z^{n+1}\|_{\Omega_h} \leq K\| Z_o\|_{\Omega_h} + K.k.N \max_n \| \Lambda^n \|_{\Omega_h}$$

which gives

$$\|Z\| \leq K\,[1 + T]\,\|\Lambda\|$$

3.4 Replacement of Boundary Conditions by Periodicity Conditions

3.4.1 Introduction

In the application of the Fourier method for stability analysis the equations describing the finite difference scheme must have the same form at all points. This is the case for a typical scheme applied to the pure initial value problem as exemplified by the Wendroff theory of section 3.3 in which the operator S_j is applied at all points $-\infty \leq x \leq +\infty$. However in the case of a problem which includes boundary conditions, such as that of section 2.3, the operator C(k) will be defined by two sets of equations one of which is applied at internal points and the other at points on or near the boundary. For the study of stability by Fourier methods the problems must be formulated so that the finite difference scheme is completely defined by a single equation which is applied at all points, $t \neq 0$. This section discusses the transformation of the original problem to an equivalent problem in which boundary conditions have been replaced by periodicity conditions and the finite difference scheme is then completely defined by the equation which is applied at the internal points. A problem is said to be equivalent to the stated problem if it has the same solution within the given region and if the boundary and initial conditions are still satisfied. For these problems with periodicity conditions or for the pure initial value problem stability can be investigated by Fourier methods or by the direct method of section 3.6.

The possibility of forming an equivalent problem is demonstrated below by taking a particular example and showing how the equivalent problem can be constructed for different types of boundary conditions. The initial function is taken to be the definition of a function over a half period and a continuation is defined in the other half period so that the boundary conditions will be satisfied. The initial function is then defined periodically for all values of x and the

new problem is effectively a pure initial value problem and the conditions on the boundaries of the original problem are as required in that problem.

The computational scheme can be easily constructed in terms of values inside the region of the original scheme by letting $\eta_j^n = \eta_{j+2M}^n$ whenever j lies outside the range -M+1, -M+2, ...0, ...+M where $x_j = j.h$, $h = \pi/M$. The range of x values in the basic problem is $0 \leq x \leq \pi$ in order to simplify the subsequent analysis. The analysis is first given for the case of homogeneous mixed boundary conditions. However a simple transformation will enable the non-homogeneous case to be considered and an example shows how this can be achieved.

3.4.2 Homogeneous Mixed Boundary Conditions

The existence of an equivalent problem will be shown by considering a simple example on the parabolic equation of section 2.3. The analysis deals with the cases of boundary conditions on the function alone, derivative boundary conditions and a class of mixed boundary conditions. The equations considered are

$$\frac{\partial y}{\partial t} = \frac{\partial^2 y}{\partial x^2} + g(x,t) \qquad \begin{matrix} 0 \leq t \leq T \\ 0 < x < \pi \end{matrix} \qquad 61(a)$$

$$y(x,0) = f(x) \qquad 0 \leq x \leq \pi \qquad 61(b)$$

with mixed boundary conditions

$$\alpha_1 \frac{\partial y(0,t)}{\partial x} + \beta_1 y(0,t) = 0 \qquad 0 \leq t \leq T \qquad 61(c)$$

$$\alpha_2 \frac{\partial y(\pi,t)}{\partial x} + \beta_2 y(\pi,t) = 0 \qquad 61(d)$$

The equivalent problem without boundary conditions is defined by extending g(x, t) and f(x) in some manner into the range $-\pi < x < 0$

and defining values outside that range by $\bar{f}(x+2p\pi) = \bar{f}(x)$, $p=0, \pm1, \ldots, -\pi \leq x \leq \pi$. $\bar{f}(x)$ denotes the extended function which is composed of the given function and a function $p(x), -\pi < x < 0$, which is chosen so that the boundary conditions are satisfied.

The existence of the equivalent problem is most easily demonstrated by considering the class of problems whose solution can be represented as a Fourier series

$$y(x,t) = \left[\sum_{0}^{\infty} a_r \cos rx + \sum_{1}^{\infty} b_r \sin rx \right] e^{-r^2 t} \qquad (62)$$

(see Carslaw and Jaeger 1947). Although the Fourier solution is used to demonstrate the existence of the equivalent problem it should be noted that the stability of the periodic problem can then be analysed without recourse to Fourier methods (see section 3.6). If $y(x, t)$ defined by (62) satisfies (61) then

$$\alpha_1 \sum_{r=1}^{\infty} r\, b_r + \beta_1 \sum_{r=0}^{\infty} a_r = 0 \qquad 63(a)$$

$$\alpha_2 \sum_{r=1}^{\infty} r(-1)^r b_r + \beta_2 \sum_{r=0}^{\infty} (-1)^r a_r = 0 \qquad 63(b)$$

and these equations will be satisfied if $a_o = 0$ and for every $r \geq 1$

$$\alpha_1 r\, b_r + \beta_1 a_r = 0 \qquad 64(a)$$

$$\alpha_2 r\, b_r + \beta_2 a_r = 0 \qquad 64(b)$$

In the case of boundary conditions on the function alone we have $\alpha_1 = \alpha_2 = 0$ and the solution of (64) is $a_r = 0$. Thus the periodic function

$$\bar{y}(x,t) = \left[\sum_{1}^{\infty} b_r \sin rx \right] e^{-r^2 t}$$

will satisfy the boundary conditions and the continuation p(x), $-\pi \leq x \leq 0$ is defined to make the initial function odd, i.e. $p(x) = -f(-x)$, $-\pi < x < 0$. In all cases the function g(x, t) is treated in the same manner.

In the case of boundary conditions on the derivative alone we have $\beta_1 = \beta_2 = 0$ and the solution of (64) is $b_r = 0$. The solution is now defined as a cosine series and f(x) and g(x, t) are defined by even continuation $p(x) = f(-x)$, $-\pi < x < 0$.

In the case of mixed boundary conditions the equations (63) are written in the form

$$\alpha_1 \left[\sum_{r=1}^{\infty} 2r\, b_{2r} + \sum_{r=1}^{\infty} (2r-1) b_{2r-1} \right] + \beta_1 \left[\sum_{r=0}^{\infty} a_{2r} + \sum_{r=1}^{\infty} a_{2r-1} \right] = 0$$

$$\alpha_2 \left[\sum_{r=1}^{\infty} 2r\, b_{2r} - \sum_{r=1}^{\infty} (2r-1) b_{2r-1} \right] + \beta_2 \left[\sum_{r=0}^{\infty} a_{2r} - \sum_{r=1}^{\infty} a_{2r-1} \right] = 0$$

and these equations are satisfied if $a_o = 0$ and for every $r \geq 1$

$$\alpha_1[2r\, b_{2r} + (2r-1) b_{2r-1}] + \beta_1[a_{2r} + a_{2r-1}] = 0 \qquad 65(a)$$

$$\alpha_2[2r\, b_{2r} - (2r-1) b_{2r-1}] + \beta_2[a_{2r} - a_{2r-1}] = 0 \qquad 65(b)$$

It will be assumed that $\beta_1, \beta_2 \neq 0$ in order to simplify the notation but similar arguments can be used if β_1 or β_2 are zero. Let $K_1 = \alpha_1 / \beta_1$, $K_2 = \alpha_2 / \beta_2$ which gives the following equations by addition and subtraction of equations (65)

$$2r(K_1+K_2) b_{2r} + (2r-1)(K_1-K_2) b_{2r-1} + 2a_{2r} = 0 \qquad 66(a)$$

$$2r(K_1-K_2) b_{2r} + (2r-1)(K_1+K_2) b_{2r-1} + 2a_{2r-1} = 0 \qquad 66(b)$$

Therefore the boundary conditions will be satisfied if the initial function can be represented by a Fourier series

$$f(x) = \sum_{r=1}^{\infty} b_r \sin rx - \tfrac{1}{2} \sum_{r=1}^{\infty} [2r(K_1+K_2)b_{2r} + (2r-1)(K_1-K_2)b_{2r-1}]\cos(2rx)$$

$$- \tfrac{1}{2} \sum_{r=1}^{\infty} [2r(K_1-K_2)b_{2r} + (2r-1)(K_1+K_2)b_{2r-1}]\cos((2r-1)x)$$

$$0 \leq x \leq \pi \qquad (67)$$

For a finite Fourier series the extension p(x) can be calculated by $p(x) = f(+x), -\pi < x < 0$ but there is no simple pictorial description of the extension p(x) such as the even or odd continuation of the previous examples.

A simple example is given of a problem for which there exists an equivalent periodic problem. Let the boundary conditions be

$$- \frac{\partial u}{\partial x}(0,t) + u(0,t) = 0$$

$$2 \frac{\partial u}{\partial x}(\pi,t) + u(\pi,t) = 0$$

and let the initial function be

$$f(x) = 8 \cos x + 2 \sin x + 3 \sin 2x$$

$$0 \leq x \leq \pi$$

Since $\frac{\partial f}{\partial x} = - 8 \sin x + 2 \cos x + 6 \cos 2x$

the boundary conditions are satisfied in the equivalent problem with initial function defined by $f(x), -\infty \leq x \leq \infty$ and it can be seen that

the coefficients of the function f(x) satisfy equations 66.

Thus it can be seen that for those problems with an initial function which can be represented in the form (67) then a periodic problem can be formulated for problems with mixed boundary conditions which is equivalent to the original problem.

3.4.3 Non-homogeneous Boundary Conditions

In the case of non-homogeneous boundary conditions the problem can be transformed into a problem with homogeneous boundary conditions. This is most easily demonstrated by means of a simple example. If the boundary conditions are

$$\begin{aligned} y(0,t) &\equiv p_1(t) \\ y(\pi,t) &\equiv p_2(t) \end{aligned} \qquad 0 \leq t \leq T \qquad (68)$$

then the problem can be reduced to the form (61) by defining a new function

$$y^*(x,t) = y(x,t) - (1-x)p_1(t) - xp_2(t) \qquad 69(a)$$

with

$$y^*(x,0) = f^*(x) \equiv f(x) - (1-x)p_1(0) - xp_2(0) \qquad 69(b)$$

$$g^*(x,t) \equiv g(x,t) - (1-x)\frac{\partial p_1}{\partial t} - x\frac{\partial p_2}{\partial t} \qquad 69(c)$$

and this has the required form to satisfy (61). The previous analysis can now be applied to show that the problem is equivalent to a periodic problem.

3.4.4 Comparison of Difference Schemes

From the above discussion it can be seen that the problem defined by equations (61) can be solved by two different approaches. The problem can either be solved directly by setting up difference equations which include approximations to the boundary conditions or the original problem can be transformed to a periodic problem in which case the finite difference scheme only involves the approximations to the differential equation and to the initial conditions. From the results of section 2.3.4 it can be seen that the stability conditions for the two schemes are not necessarily the same. In many cases a difference scheme with boundary conditions is used in practice and strictly speaking the Fourier methods of stability analysis cannot be used in such schemes.

The two alternative schemes are given below for the example of equation (61) with boundary conditions on the function alone, i.e. expressing the equations as a mapping

$$(F\,u)_1 \equiv \frac{\partial u}{\partial t} - \frac{\partial^2 u}{\partial x^2} - g(x,t) \qquad \begin{matrix} 0 \leq t \leq T \\ 0 < x < \pi \end{matrix} \qquad 70(a)$$

$$(F\,u)_2 \equiv u(x,0) - f(x) \qquad 0 \leq x \leq \pi \qquad 70(b)$$

$$(F\,u)_3 \equiv u(0,t) - p_1(t) \qquad 0 \leq t \leq T \qquad 70(c)$$

$$(F\,u)_4 \equiv u(\pi,t) - p_2(t) \qquad 0 \leq t \leq T \qquad 70(d)$$

Firstly a direct difference scheme based on equations 2.40 will be considered which leads to the stability equations

$$A\,Z^{n} = B\,Z^{n-1} + \lambda\theta \begin{bmatrix} \Lambda_0^n \\ 0 \\ \vdots \\ 0 \\ \Lambda_M^n \end{bmatrix} + \lambda(1-\theta) \begin{bmatrix} \Lambda_0^{n-1} \\ 0 \\ \vdots \\ 0 \\ \Lambda_M^{n-1} \end{bmatrix} + k\,\Lambda^{n} \qquad 71(a)$$

$n = 1, 2, \ldots N$

$$Z^{o} = \Lambda^{o} \qquad 71(b)$$

where Z^o and Λ^n are vectors of dimension M-1 corresponding to the internal points j = 1, 2, ... M-1 and the matrices A and B are defined by $A = P(\lambda\theta)$, $B = P(-\lambda(1-\theta))$ where P(q) is the symmetric tridiagonal matrix of dimension M-1 with diagonal elements 1 + 2q and off - diagonal elements -q. The above equations can also be written

$$\begin{bmatrix} 1 & 0 & & & 0 \\ -\lambda\theta & 1+2\lambda\theta & -\lambda\theta & & \\ & \ddots & \ddots & \ddots & \\ & & -\lambda\theta & 1+2\lambda\theta & -\lambda\theta \\ 0 & & & 0 & 1 \end{bmatrix} Z^{n} = \begin{bmatrix} 0 & 0 & & & 0 \\ \lambda(1-\theta) & 1-2\lambda(1-\theta) & \lambda(1-\theta) & & \\ & \ddots & \ddots & \ddots & \\ & & \lambda(1-\theta) & 1-2\lambda(1-\theta) & \lambda(1-\theta) \\ 0 & & & 0 & 0 \end{bmatrix} Z^{n-1}$$

$$+ \begin{bmatrix} \Lambda_0^n \\ 0 \\ \vdots \\ 0 \\ \Lambda_M^n \end{bmatrix} + k \begin{bmatrix} 0 \\ \Lambda_1^n \\ \vdots \\ \Lambda_{M-1}^n \\ 0 \end{bmatrix} \qquad (72)$$

$n = 1, 2, \ldots N$

where the vectors are of dimension M+1. However the form of these matrices makes the evaluation of the norms in equation (72) more difficult and therefore the form given in (71) is normally used.

Alternatively the scheme can be transformed to periodic form in two stages. The boundary conditions can be transformed to homogeneous form as follows

Let $U(x,t) = u(x,t) - (1-x)\, p_1(t) - x\, p_2(t)$

$$G(x,t) = g(x,t) - (1-x)\, \frac{\partial p_1(t)}{\partial t} - x\, \frac{\partial p_2(t)}{\partial t} \qquad (73)$$

$F(x) = f(x) - (1-x)\, p_1(o) - x\, p_2(o)$

Secondly an extension to F(x) is defined in the region $-\pi < x < 0$ by $F(-x) = -F(x)$, $0 < x < \pi$. It is assumed that the boundary and initial conditions are compatible which ensures $F(0) = F(\pi) = 0$.

The definition of the problem is completed by the extension

$$\bar{F}(x+2n\pi) = \bar{F}(x), \quad -\pi \leq x \leq +\pi, \quad n = \pm 1, \pm 2, \ldots$$

This problem has the stability equations

$$-\lambda\theta Z^n_{j-1} + (1+2\lambda\theta) Z^n_j - \lambda\theta Z^n_{j+1} = \lambda(1-\theta) Z^{n-1}_{j-1} + (1-2\lambda(1-\theta)) Z^{n-1}_j + \lambda(1-\theta) Z^{n-1}_{j+1} + k\, \Lambda^n_j \qquad 74(a)$$

$n = 1, 2, \ldots N$, $j = 1, 2, \ldots 2M$ where $x_j = \frac{2\pi j}{2M} - \pi$

and $Z^o_j = \Lambda^o_j$, $\qquad j = 1, 2, \ldots 2M \qquad 74(b)$

For any value of the subscript which is outside the range $j = 1, 2, \ldots 2M$ the subscript is taken modulo 2M giving

$$A\, Z^n = B\, Z^{n-1} + k\, \Lambda^n \qquad (75)$$

where Z^n, Λ^n are vectors of dimension 2M and A and B are circulant matrices with elements defined by the first rows $[1+2\lambda\theta, -\lambda\theta, 0, \ldots 0, -\lambda\theta]$ and $[1-2\lambda(1-\theta), +\lambda(1-\theta), 0, \ldots 0, +\lambda(1-\theta)]$ respectively. The two sets of stability equations to be compared are given by equations (71) and (75). The eigenvalues of the matrix A^{-1} B in equation (71) were given in equation (2.86) and are repeated here for convenience

$$\mu_j = \frac{1-4\lambda(1-\theta)\sin^2(\pi j/2M)}{1+4\lambda\theta\sin^2(\pi j/2M)}, \qquad j = 1, 2, \ldots M-1$$

and the eigenvalues of the corresponding matrix in equation (75) were

given in equation (2.51)

$$\mu_j = \frac{1-4\lambda(1-\theta)\sin^2(\pi j/2M)}{1+4\lambda\theta\sin^2(\pi j/2M)} \quad , \; j = 1, 2, \ldots 2M$$

Thus the eigenvalues of both matrices are provided by mesh values of the function

$$g(x) = \frac{1-4\lambda(1-\theta)\sin^2(\pi x)}{1+4\lambda\theta\sin^2(\pi x)} \qquad 0 \leq x \leq 1$$

and the conditions which would lead to condition (2.46) would be the same for both matrices in the case of the 2 - norm. It was seen previously that for the case of the maximum norm the stability conditions which could be obtained were different. However even if a condition similar to (2.46) could be established the argument of Theorem 2.1 could not be followed due to the difference in the form of equation (71) and (45). Let the terms with coefficients $\lambda\theta$ and $\lambda(1-\theta)$ in equation (71) be A ψ_1^n and A ψ_2^{n-1} respectively which gives

$$Z^n = A^{-1}B\,Z^{n-1} + (\psi_1^n + \psi_2^{n-1}) + k\,\tilde{\Lambda}^n$$

$$= C^n Z^0 + \sum_{j=1}^{n} C^{n-j}(\psi_1^j + \psi_2^{j-1}) + k \sum_{j=1}^{n} C^{n-j}\tilde{\Lambda}^j \qquad (76)$$

If $\quad \| C^n(k) \| \leq K \qquad 0 < k < \tau \qquad 0 \leq nk \leq T \qquad (77)$

is to be a sufficient condition for stability it must be established that

$$\| \sum_{j=1}^{n} C^{n-j}(\psi_1^j + \psi_2^{j-1}) \| \leq K^1 \| \Lambda \| \qquad (78)$$

and the absence of the factor k means that the previous arguments cannot be applied.

Although numerical experiments indicate that the stability criteria derived for the periodic problem are also satisfactory for the problem with boundary conditions, necessary and sufficient conditions for stability are not available for the latter problem and it may be more satisfactory to transform the problem to periodic type so that rigorous conditions for stability can be given.

3.5 Fourier Theory

3.5.1 The Lax-Richtmyer Formulation

This section will discuss the stability of the difference scheme defined by equation (20) in the case of the pure initial value problem or where the boundary conditions can be replaced by periodicity conditions. A wide class of functions have a convergent Fourier series and we consider a member of this class $u \in B$. Expanding u(x) as a Fourier series we have

$$u(x) = \frac{1}{(2\pi)^{s/2}} \sum_{r} V(r)\, e^{ir.x} \tag{79}$$

The summation over r is multiple of dimension s as discussed in A1.11 and r.x is the normal scalar product. The components of r are integers and the summation is over all possible values of these integers. The norm in the space B is defined by

$$\| u \|^2 = \int_0^{2\pi} |u(x)|^2 \, dx$$

where the integral is multiple of order s and the x values are scaled so that $0 \leq x_i \leq 2\pi$. $|u(x)|$ is the normal Euclidean length of a vector. If the norm of V is defined by

$$\| V \|^2 = \sum_{r} |V(r)|^2 \tag{80}$$

then the elements V(r) form the member of a Banach space B' and the two norms are equivalent by Parseval's relation

$$\int_0^{2\pi} |u(x)|^2 \, dx = \sum_r |V(r)|^2 \tag{81}$$

Thus the behaviour of the norm of a function in the space B can be studied by considering the equivalent norm in B!

For the problem under discussion a typical finite difference scheme relates the values of η at a set of points N_1 in the region of x at time level t^{n+1} to a set of points N_o in the region of x at time level t^n. As an example the difference scheme (2.40) of section 2.3 with $s = d = 1$ has a set of three points $x - h$, x, $x + h$ at each time level. Generally the problem is solved on a mesh

$$D^h \equiv \left\{ \begin{array}{l} x_j = hj, \; t^n = nk : j = [j_1, j_2 \cdots j_s]^T, \\ 0 \leq j_1 \leq M_1, \quad n = 0, 1, \ldots N, \; k = T/N \end{array} \right\} \tag{82}$$

h is a diagonal matrix with element $h_1 = 2\pi/M_1$ $1 = 1, 2, \ldots S$ and j is a vector with integer components in the coordinate directions. The form of the difference equations is

$$\sum_{N_1} C_{1p} \, \eta^*(x' + hp, t^{n+1}) = \sum_{N_o} C_{op} \, \eta^*(x' + hp, t^n) \tag{83}$$

where the sets N_1 and N_o define the respective values of p, and C_{ip}, $i = 0, 1$ are d x d matrices which do not depend upon x and t in the simple presentation but may depend upon k and h. Using a Fourier series (79) for $\eta^*(x)$ and applying equation (83) at each point x gives

$$\frac{1}{(2\pi)^{s/2}} \left[\sum_{N_1} C_{1p} \sum_r V_{(r)}^{n+1} \, e^{ir.(x'+hp)} - \sum_{N_o} C_{op} \sum_r V_{(r)}^{n} \, e^{ir.(x'+hp)} \right] = 0 \tag{84}$$

Taking a typical value of r and removing the common factor $e^{ir.x'}/(2\pi)^{s/2}$ gives

$$\sum_{N_1} C_{1p} \overset{n+1}{V}(r) e^{ir.hp} - \sum_{N_0} C_{op} \overset{n}{V}(r) e^{ir.hp} = 0$$

$$\text{Let} \quad H_j = \sum_{N_j} C_{jp} e^{ir.hp} \qquad j = 0, 1 \tag{85}$$

$$\text{then} \quad H_1 V^{n+1}(r) = H_0 V^n(r)$$

Since H_1 must have an inverse if the system is solvable we have

$$\begin{aligned} \overset{n+1}{V}(r) &= H_1^{-1} H_0 V^n(r) \\ &= G(k,r) V^n(r) \\ &= G^{n+1}(k,r) V^0(r) \end{aligned} \tag{86}$$

$$\text{Thus} \quad \| V^n(r) \| \leq \| G^n(k,r) \| \| V^0(r) \|$$

$$\text{and if} \quad \| G^n(k,r) \| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \tag{87}$$

for all r with integer components then the growth of the norm in B' is bounded and by (81) the same is true of the norm in B. Hence as in Theorem 2.1 the scheme is stable according to the definitions of section 2.1. The matrix G(k, r) is known as the amplification matrix. For the example of equation (2.40) the finite difference scheme is defined by

$$-\lambda\theta\, \eta_{j-1}^{n+1} + (1+2\lambda\theta)\eta_j^{n+1} - \lambda\theta\eta_{j+1}^{n+1} = \lambda(1-\theta)\, \eta_{j-1}^{n} + (1-2\lambda(1-\theta))\eta_j^n + \lambda(1-\theta)\eta_{j+1}^n \tag{88}$$

Hence

$$H_1 = (1+2\lambda\theta) - \lambda\theta\left[e^{ir.2\pi/2M} + e^{-2\pi ir/2M}\right]$$

$$H_0 = 1 - 2\lambda(1-\theta) + \lambda(1-\theta)\left[e^{2\pi ir/2M} + e^{-2\pi ir/2M}\right]$$

and G(k, r) is a scalar given by

$$G(k,r) = \frac{1-2\lambda(1-\theta)+2\lambda(1-\theta)\cos 2\pi r/2M}{1+2\lambda\theta-2\lambda\theta\cos 2\pi r/2M}$$

$$= \frac{1-4\lambda(1-\theta)\sin^2(\pi r/2M)}{1+4\lambda\theta\sin^2(\pi r/2M)} \qquad (89)$$

so that the condition (2.50) is necessary and sufficient for stability in the L - R sense as in section 2.3. In considering the necessity of this condition a value of $(\Lambda^o)^*$ is taken corresponding to the maximum value of $\| G(k, r) \|$. However $(\Lambda^o)^*$ is an extended function and the L - R theory does not establish that the mesh function would be subject to the same amplification factor. This topic is discussed in section 3.6 which analyses the problem in terms of a mesh norm.

3.5.2 The Wendroff Formulation

In order to satisfy equation (59) in the case of constant coefficients with no non-linear term it must be established that S^n is uniformly bounded $0 < k < \tau$, $0 \leq nk \leq T$. The Fourier transform $\hat{u}$ is formed as follows

$$\hat{u}(r) = \sum_{j=-\infty}^{+\infty} u_j e^{-ir.j} \qquad (90)$$

where r.j is a scalar product of dimension S. The Fourier transform applied to the operator T of equation (46) gives

$$\widehat{T^\sigma u} = \sum_{j=-\infty}^{+\infty} (T^\sigma u)_j e^{-ir.j}$$

$$= \sum_j u_{j+\sigma} e^{-ir.j}$$

$$= \sum_{k=-\infty}^{+\infty} u_k e^{-ir(k-\sigma)} \qquad k = j + \sigma$$

$$= e^{ir.\sigma} \sum_{k=-\infty}^{+\infty} u_k e^{-ir.k} \qquad (91)$$

Hence if we define $\hat{T}^{\sigma} = e^{ir.\sigma}$ then

$$\hat{T}^{\sigma} u = T^{\sigma} \hat{u}$$

and the operator (46) becomes in the transform space

$$\hat{S} = \sum_{p=-J}^{+J} C_p \, e^{ir.p}$$

Wendroff proves that the norm in the space Ω_h

$$\| u \|^2 = \sum_{j=-\infty}^{+\infty} |u_j|^2 \tag{92}$$

is equivalent to the norm in the transform space

$$\| u \|^2 = \int_0^{2\pi} \cdots \int_0^{2\pi} |\hat{u}|^2 \, dr_1 \cdots dr_s \tag{93}$$

Hence the uniform boundedness of S^n is equivalent to the uniform boundedness of $\hat{S}^n$. Thus the following condition is sufficient for stability of the Wendroff scheme

$$\| \hat{S}^n \| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \tag{94}$$

but Wendroff also shows that if this condition is not satisfied for a single value of r in the transform space then the scheme cannot be stable. Thus

$$\| \left(\sum_{p=-J}^{+J} C_p \, e^{ir.p} \right)^n \| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \tag{95}$$

is also necessary for all r to ensure stability. An alternative approach without recourse to Fourier analysis is presented in section 3.6.

3.6 A Direct Analysis of Stability

3.6.1 Introduction

This section provides a direct approach to the study of stability in partial differential equations for the class of problems which are amenable to Fourier methods; i.e. the method is restricted to either the pure initial value problem or to the case where boundary conditions can be replaced by periodicity conditions as described in section 3.4. In the Fourier analysis it was established that stability was equivalent to the uniform boundedness of the powers of the amplification matrix. It is shown below that the same result can be obtained by using the special properties of the matrix which results from the finite difference equations which arise in the two cases mentioned above. The matrices which are formed are shown to have circulant form and it was necessary to have expressions for the eigenvalues, eigenvectors and certain norms of the circulant matrices. The results available were for the simplest case (Aitken (1958)) but a development of this theory is presented in appendix 1 in which the required results are derived.

3.6.2 A Problem with Boundary Conditions

It is assumed that the transformations described in section 3.4 have been performed so that the finite difference scheme is described by an equation relating values at a set of points in the neighbourhood of x_j at time t^{n+1} to another set of points in the neighbourhood of x_j at time t^n. This equation will be applied at all points x_j and where values are required at points outside the defined region these will be replaced by values inside the region which are equal by the periodicity property. An example has been given in section 2.3 for

the parabolic equation where the values are defined on the mesh D^h (2.38) and the finite difference equation by (2.40.a). By periodicity over the extended range 1, 2M we would replace η_{2M+1} by η_1 etc. and this would give a circulant matrix. For the more general case where the problem is to be solved in the region $0 \leq x^l \leq 1$, $l = 1,2, \ldots s$, $0 \leq t \leq T$ the finite difference solution will be calculated on a mesh such as the following

$$D^h \equiv \left\{ \begin{array}{l} x_j = hj : \quad j = [j_1, j_2, \ldots j_s]^T, \; 1 < j_l < 2M_l, \; h_l = 1/M_l \\ t^n = nk : \quad n = 0, 1, \ldots N, \quad k = T/N \end{array} \right\} \quad (96)$$

h is a diagonal matrix with element h_l, $l = 1,2, \ldots s$ and j is a vector with integer components in the co-ordinate directions. In the general case the finite difference solution will be a vector of dimension d defined for every point of D^h.

The form of matrices which arise is most easily demonstrated by means of practical examples. The example of the heat conduction equation in one and two dimensions is given below. For the case $s = d = 1$ discussed in section 2.3 the stability equation at the point x_j is defined at the point x_j by (2.43.d) and using the periodic replacement of external values the matrix equation is given by

$$\begin{bmatrix} 1+2\lambda\theta & -\lambda\theta & \ldots & 0 & -\lambda\theta \\ -\lambda\theta & 1+2\lambda\theta & -\lambda\theta & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \ldots & -\lambda\theta & 1+2\lambda\theta & -\lambda\theta \\ -\lambda\theta & 0 & \ldots & -\lambda\theta & 1+2\lambda\theta \end{bmatrix} Z^n$$

$$= \begin{bmatrix} 1-2\lambda(1-\theta) & \lambda(1-\theta) & \ldots & 0 & \lambda(1-\theta) \\ \lambda(1-\theta) & 1-2\lambda(1-\theta) & \lambda(1-\theta) & \ldots & 0 \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ 0 & \ldots & \lambda(1-\theta) & 1-2\lambda(1-\theta) & \lambda(1-\theta) \\ \lambda(1-\theta) & 0 & \ldots & \lambda(1-\theta) & 1-2\lambda(1-\theta) \end{bmatrix} Z^{n-1} + k\Lambda^n$$

$$\text{or } A\, Z^n = B\, Z^{n-1} + k\, \Lambda^n \qquad \ldots\ldots\ldots\ldots (97)$$

where A and B are circulant.

This gives $Z^n = C\, Z^{n-1} + k\, \tilde{\Lambda}^n$ $\qquad \ldots\ldots\ldots\ldots (98)$

where $C = A^{-1} B$ is a circulant matrix, $\tilde{\Lambda}^n = A^{-1} \Lambda^n$.

In the two-dimensional case the equations are

$$-\lambda_1 \theta_1 [Z^n_{j_1-1,j_2} + Z^n_{j_1+1,j_2}] + [1 + 2\lambda_1 \theta_1 + 2\lambda_2 \theta_2] Z^n_{j_1,j_2} - \lambda_2 \theta_2 [Z^n_{j_1,j_2-1} + Z^n_{j_1,j_2+1}]$$

$$= \lambda_1(1-\theta_1) [Z^{n-1}_{j_1-1,j_2} + Z^{n-1}_{j_1+1,j_2}] + [1-2\lambda_1(1-\theta_1)-2\lambda_2(1-\theta_2)] Z^{n-1}_{j_1,j_2} + \lambda_2(1-\theta_2)[Z^{n-1}_{j_1,j_2-1} + Z^n_{j_1,j_2+1}] + k\, \Lambda^n_j$$

$\qquad \ldots\ldots\ldots (99)$

If we take the case $M_1 = M_2 = 2$ we have

$$\begin{bmatrix} a^1_1 & a^1_2 & 0 & a^1_2 \\ a^1_2 & a^1_1 & a^1_2 & 0 \\ 0 & a^1_2 & a^1_1 & a^1_2 \\ a^1_2 & 0 & a^1_2 & a^1_1 \end{bmatrix} Z^n = \begin{bmatrix} a^0_1 & a^0_2 & 0 & a^0_2 \\ a^0_2 & a^0_1 & a^0_2 & 0 \\ 0 & a^0_2 & a^0_1 & a^0_2 \\ a^0_2 & 0 & a^0_2 & a^0_1 \end{bmatrix} Z^{n-1} + k\, \Lambda^n$$

$\qquad \ldots\ldots (100)$

$$a^1_1 = \begin{bmatrix} 1 + 2\lambda_1\theta_1 + 2\lambda_2\theta_2 & -\lambda_2\theta_2 & 0 & -\lambda_2\theta_2 \\ -\lambda_2\theta_2 & 1 + 2\lambda_1\theta_1 + 2\lambda_2\theta_2 - \lambda_2\theta_2 & & 0 \\ 0 & -\lambda_2\theta_2 & 1+2\lambda_1\theta_1 + 2\lambda_2\theta_2 & -\lambda_2\theta_2 \\ -\lambda_2\theta_2 & 0 & -\lambda_2\theta_2 & 1+2\lambda_1\theta_1 + 2\lambda_2\theta_2 \end{bmatrix}$$

$$a^1_2 = \begin{bmatrix} -\lambda_1\theta_1 & 0 & 0 & 0 \\ 0 & -\lambda_1\theta_1 & 0 & 0 \\ 0 & 0 & -\lambda_1\theta_1 & 0 \\ 0 & 0 & & -\lambda_1\theta_1 \end{bmatrix}$$

where $Z^T = (Z_1^T, Z_2^T, Z_3^T, Z_4^T)$, $Z_j^T = (Z_{j1}, Z_{j2}, Z_{j3}, Z_{j4})$

Thus the matrices are circulant in form with elements which are themselves circulant matrices. In the most general case the elements of C are d × d matrices a_k with multiple subscripts $k = [k_1, k_2, \ldots k_s]$ and C is circulant as defined by (A1.3).

Following the line of argument of section 3.2 equation (98) gives

$$||Z|| < ||C^n|| \; ||\Lambda|| + k.N. \max_{j=1,..n} ||C^j||. \; ||\Lambda||$$

and if $||C^n||$ is uniformly bounded $0 \leq n.k \leq T$ then

$$||Z|| \leq K \; ||\Lambda|| \qquad \ldots\ldots\ldots(101)$$

which gives stability. The results of appendix 1 will be used to find conditions for the uniform boundedness of $||C^n||$

Theorem 2

If the norm is taken as

$$|Z|^2 = \sum_{j=1}^{2M} |Z_j|^2 \qquad \ldots\ldots\ldots\ldots(102)$$

then the subordinate matrix norm of C^n is given by

$$||C^n|| = \max_r \rho^{\frac{1}{2}} (G^{n*}G^n) = \max_r ||G^n|| \qquad \ldots\ldots(103)$$

where $G = \sum_{k=1}^{2M} a_k \, w_r^{k-1}$ $\qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots$ (104)

In the case where $s > 1$ the same notation can be used if we let

$$\sum_{k=1}^{2M} \equiv \sum_{k_1=1}^{2M_1} \ldots\ldots \sum_{k_s=1}^{2M_s} \quad \text{and} \quad w_r^{k-1} \equiv w_{r_1}^{k_1-1} \cdot w_{r_2}^{k_2-1} \ldots\ldots w_{r_s}^{k_s-1}$$

Proof $\quad \|C^n\|^2 = \rho\,(C^{n*}\,C^n)$

If we define $B^{(n)} = C^{n*}\,C^n$ then we wish to find the spectral radius of $B^{(n)}$. Let $a_k^{(n)}$ be the element in the kth position on the first row of C^n. Using equation (A1.3)

$$b_{pq}^{(n)} = \sum_{k=1}^{2M} a_{p-k+1}^{(n)*}\; a_{q-k+1}^{(n)}$$

In the general case $b_{pq}^{(n)}$ is a square matrix of dimension d. Result 2 of appendix 1 gives the spectral radius of $B^{(n)}$

$$\rho\,(B^{(n)}) = \max_r\; \rho\,\Big(\sum_{k=1}^{2M} b_k^{(n)}\; w_r^{k-1}\Big)$$

$$\sum_{k=1}^{2M} b_k^{(n)}\; w_r^{k-1} = \sum_{k=1}^{2M}\sum_{j=1}^{2M} a_j^{(n)*}\; a_{j+k-1}^{(n)}\; w_r^{k-1}$$

$$= \Big(\sum_{j=1}^{2M} a_j^{(n)*}\; w_r^{-j+1}\Big)\Big(\sum_{i=1}^{2M} a_i^{(n)}\; w_r^{i-1}\Big)$$

$$= \Big(\sum_{j=1}^{2M} a_j^{(n)*}\,(w_r^*)^{j-1}\Big)\Big(\sum_{i=1}^{2M} a_i^{(n)}\; w_r^{i-1}\Big) = Q_n^*(r)\; Q_n(r)$$

.....(105)

Hence it is required to prove that

$$Q_n(r) \equiv \sum_{i=1}^{2M} a_i^{(n)}\; w_r^{i-1} = \Big(\sum_{k=1}^{2M} a_k\; w_r^{k-1}\Big)^n \equiv G^n(r) \quad ....(106)$$

The proof is by induction. For n = 1 the two quantities are identical and equation (106) is assumed for n = 2, ... n. Multiply (106) by $\sum_{i=1}^{2M} a_k\; w_r^{k-1}$. The left hand side becomes

$$\Big(\sum_{i=1}^{2M} a_i\; w_r^{i-1}\Big)\Big(\sum_{j=1}^{2M} a_j^{(n)}\; w_r^{j-1}\Big) = \sum_{k=1}^{2M}\sum_{i=1}^{2M} a_i\; a_{k-i+1}^{(n)}\; w_r^{k-1} \quad ...(107)$$

Now the elements $a_i^{(n+1)}$ are obtained from $C.C^n$.

Hence

$$a_k^{(n+1)} = \sum_{i=1}^{2M} C_{pi}\, C_{iq}^{(n)} = \sum_{i=1}^{2M} a_{i-p+1}\, a_{q-i+1}^{(n)}$$

$$k = (q - p+1 \bmod 2M)$$

Therefore

$$\sum_{k=1}^{2M} a_k^{(n+1)}\, w_r^{k-1} = \sum_{k=1}^{2M}\sum_{i=1}^{2M} a_{i-p+1}\, a_{q-i+1}^{(n)}\, w_r^{k-1}$$

$$= \sum_{k=1}^{2M}\sum_{i=1}^{2M} a_{i-p+1}\, a_{k+p-1-i+1}^{(n)}\, w_r^{k-1}$$

$$= \sum_{k=1}^{2M}\sum_{j=1}^{2M} a_j\, a_{k-j+1}^{(n)}\, w_r^{k-1} \qquad \text{.....(108)}$$

$$j = i-p+1$$

and the right hand side of (108) equals the right hand side of (107) thus proving (106) for value n+1.

Hence $\quad ||C^n||^2 = \rho\,(B^{(n)}) = \max_r \rho\,(G^{n*}\, G^n)$

It is therefore established that the C^n is uniformly bounded if

$$||G^n\,(k,r)|| \leq K \qquad 0 < k < \tau \qquad \text{.....(109)}$$
$$0 \leq nk \leq T$$

for all r. This is the standard condition on the amplification matrix (see equations (87) and (94).

3.6.3 The Pure Initial Value Problem

The discussion of this section is based on the problem formulated in section 3.3, the important features of which are summarised here. The initial values are defined by

$$\eta_j^o = \phi(x_j) \qquad 0 \leq j_l \leq M_l, \quad l = 1,2 \ldots s.$$

$$\eta_j^o = 0 \qquad \text{for other values of } j$$

where j is a vector of S components the values of which range over all integer values. The solution $Y \equiv \{ \eta^n : n = 0,1,\ldots N, T = N.k\}$ is composed of elements $\eta^n = \{ \eta_j^n : j_l = 0, \pm 1, \ldots, l = 1,2,\ldots s\}$ where each η_j^n may be a vector d in the most general case. The finite difference scheme considered is an explicit scheme with constant coefficients

$$\eta^{n+1} = S\,\eta^n \qquad \ldots\ldots\ldots\ldots (110)$$

It is assumed that the operator S is bounded, which will normally be the case if the difference scheme approximates a differential scheme without singularities. It is shown in section 3.3 that if the problem is formulated carefully the mapping S maps the space Ω_h into itself so that all η^{n+1} are members of Ω_h. The complete solution is defined by a collection of elements defined on the mesh

$$D^h \equiv \left\{ \begin{array}{l} x_j = hj : j = [j_1, j_2, \ldots j_s]^T,\ j_l = 0, \pm 1, \ldots; h_l = 1/M_l \\ t^n = nk:\ n = 0,1, \ldots N,\ k = T/N \end{array} \right\}$$

For simplicity of notation the future discussion will be restricted to

the case s = 1 but the results can be extended to the general case quite easily, and the norm is defined by

$$||\eta||^2 = \max_n ||\eta^n||^2 = h \left[\max_n \sum_{j=-\infty}^{+\infty} |\eta_j^n|^2 \right] \quad .....(111)$$

Wendroff defines the space Ω_h such that $\eta_j^n = 0$ for $|x_j| > x_{max}(t)$ and therefore the norm (111) will always exist.

It will be shown in this section that if the ideas of section 3.6.2 are extended to the infinite dimensional case then the Fourier framework occurs quite naturally in discussing the coefficients of a vector η^n when expressed in terms of the eigenvectors of the operator S^n. The condition on the amplification matrix is then found to be a necessary and sufficient condition for stability on the 2-norm. The equation (110) leads to the stability equation

$$Z^{n+1} = S\, Z^n + \Lambda^{n+1}$$

$$= S^{n+1}\, \Lambda^o + \sum_{j=0}^{n} S^j\, \Lambda^{n+1-j} \quad .....(112)$$

and the uniform boundedness of $||S^n||$ is necessary and sufficient for stability. It will first be shown that an element η_j^n can be expressed in terms of the eigenvectors of $(S^n)^* S^n$. The operator S can be represented as an infinite operator of the form

$$\begin{matrix} - & - & - & - & - & - & - & - & - & - & - & - & - & - & - \\ a_{-p} & a_{-p+1} & \cdots\cdots & & a_o & \cdots\cdots & a_p & & & & & & & & \\ & a_{-p} & \cdots\cdots & & a_o & \cdots\cdots & a_p & & & & & & & & \end{matrix}$$

.....(113)

and S^n has a similar form but with elements $a_p^{(n)}$, $p = -nP, \ldots + nP$. It can easily be seen that the elements of S^{n+1} are given by the coefficients of Z^p in the expansion of

$$(a_{-nP}^{(n)} Z^{-nP} + a_{-nP+1}^{(n)} Z^{-nP+1} \ldots\ldots a_{nP} Z^{nP})(a_{-P} Z^{-P} + \ldots\ldots a_P Z^P)$$

and inductively by the coefficient of Z^p in

$$(a_{-P} Z^{-P} + \ldots\ldots a_P Z^P)^{n+1} \qquad \ldots\ldots\ldots\ldots\ldots (114)$$

Hence the elements of $b_p^{(n)}$ of the operator $(S^n)^* S^n$ are given by the coefficients of Z^p in

$$(a_P^* Z^{-P} + \ldots\ldots a_{-P}^* Z^P)^n (a_{-P} Z^{-P} + \ldots\ldots a_P Z^P)^n \ldots\ldots (115)$$

The eigenvalues and eigenvectors of $(S^n)^* S^n$ can now be found from the equation

$$b_{-2nP}^{(n)} v_{j-2nP} + b_{-2nP+1}^{(n)} v_{j-2nP+1} + \ldots\ldots (b_o^{(n)} - \mu) v_j \ldots\ldots + b_{2nP}^{(n)} v_{j+2nP} = 0 \qquad \ldots\ldots\ldots (116)$$

Treating this as a difference equation for the vector components v_j of dimension d a solution can be obtained by assuming a solution $v_j = c^j \nu$ where ν is a constant vector and C is a constant raised to the power j. Equation (116) becomes

$$c^j \left(\sum_{p=-2nP}^{2nP} b_p^{(n)} c^p - \mu I \right) \nu = 0$$

If the components v_j are to be uniformly bounded with respect to n

then since $j = 0, +1, +2, \ldots$ and $j = 0, -1, -2, \ldots$ it follows that $|C| = 1$. Therefore equation (116) becomes

$$\left(\sum_{p=-2nP}^{2nP} b_p^{(n)} e^{ip\theta} - \mu I \right) \nu = 0$$

and using equation (115) this is equivalent to

$$\left[\left(\sum_{p=-P}^{P} a^*_p e^{-ip\theta} \right)^n \left(\sum_{p=-P}^{P} a_p e^{ip\theta} \right)^n - \mu I \right] \nu = 0$$

or $$(G^{n*} G^n - \mu I) \nu = 0 \qquad \ldots\ldots\ldots\ldots (117)$$

where $$G = \sum_{p=-P}^{P} a_p e^{ip\theta} \qquad \ldots\ldots\ldots (118)$$

Thus the eigenvalues of $(S^n)^* S^n$ are given by $\mu_r(\theta)$, $r = 1, 2, \ldots d$, $0 \leq \theta \leq 2\pi$ defined by equation (117) and it can be seen by inspection that the eigenvectors are given by

$$[v_r(\theta)]^T = \ldots\ldots\ldots e^{-i\theta}[\nu_r(\theta)]^T, [\nu_r(\theta)]^T, e^{i\theta}[\nu_r(\theta)]^T, \ldots\ldots (119)$$

for $r = 1, 2, \ldots d$, $0 \leq \theta \leq 2\pi$ where $\nu_r(\theta)$ is an eigenvector of $(G^n)^* G^n$. It will now be shown that

$$||S^n||^2 = \left(\sup_{r,\theta} |\mu_r(\theta)|\right)^2 = \max_{\theta} \rho(G^{n*} G^n) = \max_{\theta} ||G^n||^2 \ldots (120)$$

It will now be shown that any vector $u \in \Omega_h$ can be uniquely expressed in terms of the eigenvectors given by equation (110) and the condition for stability is then seen tc be the standard condition on the amplification matrix. Now for any $u \in \Omega_h$ we have

$$\sum_{j=-\infty}^{+\infty} |u_j|^2 < \infty \qquad \ldots\ldots\ldots (121)$$

and hence a Fourier series can be formed

$$\psi(\theta) = \sum_{j=-\infty}^{+\infty} u_j e^{-ij\theta} \qquad \text{.....(122)}$$

where $\psi(\theta)$ is a d × d matrix

and inverting this relationship gives

$$u_j = \frac{1}{2\pi} \int_0^{2\pi} e^{ij\theta} \psi(\theta)\, d\theta \qquad j = 0, \pm 1, \ldots$$

If the eigenvectors $\nu_r(\theta)$ are linearly independent a unique representation $\psi(\theta) = \nu(\theta)\, w(\theta)$ exists where $\nu(\theta)$ is a matrix whose columns are the eigenvectors $\nu_r(\theta)$ and $w(\theta)$ is a column vector of coefficients. Hence a unique representation exists

$$u_j = \frac{1}{2\pi} \int_0^{2\pi} e^{ij\theta} \nu(\theta)\, w(\theta)\, d\theta \qquad \text{.....(123)}$$

and if the eigenvectors $v_r(\theta)$ are taken as the columns of $v(\theta)$ then the desired representation of u is

$$u = \frac{1}{2\pi} \int_0^{2\pi} v(\theta)\, w(\theta)\, d\theta \qquad \text{.....(124)}$$

If we consider the space B' with elements $\nu(\theta)\, w(\theta)$ then by Parseval's relation

$$||u||^2_{\Omega_h} = ||u||^2_{B'}$$

i.e. $$\sum_{j=-\infty}^{+\infty} |u_j|^2 = \frac{1}{2\pi} \int_0^{2\pi} |\nu(\theta)\, w(\theta)|^2\, d\theta \qquad \text{.....(125)}$$

Hence $$||S^n||^2_{\Omega_h} = \max_u \left(\frac{||S^n u||^2_{B'}}{||u||^2_{B'}} \right.$$

and using the results of equations (115) and (118) the right hand side becomes

$$\max_{u} \left(\frac{\int_0^{2\pi} |G^n(\theta)\, \nu(\theta)\, w(\theta)|^2 \, d\theta}{||u||^2_{B'}} \right)$$

$$= \max_{u} \left(\frac{\int_0^{2\pi} w^*(\theta)\, \nu^*(\theta)\, G^{n^*}(\theta)\, G^n(\theta)\, \nu(\theta)\, w(\theta)\, d\theta}{||u||^2_{B'}} \right)$$

$$\leq \max_{\theta} ||G^{n^*}(\theta)\, G^n(\theta)|| = \left(\sup_{r,\theta} |\mu_r(\theta)| \right)^2 \qquad \ldots\ldots(126)$$

Thus a sufficient condition for stability is that

$$||G^n(\theta)|| \leq K \qquad \begin{matrix} 0 < k < \tau \\ 0 \leq nk \leq T \end{matrix} \qquad \ldots\ldots(127)$$

for all θ.

Wendroff also shows that this condition is necessary for stability and hence (127) is the standard necessary and sufficient condition which the amplification matrix must satisfy to give stability.

3.7 The Amplification Matrix

3.7.1 Introduction

The analysis presented in sections 3.5 and 3.6 established that for a certain class of problems the necessary and sufficient condition for stability is that the powers of the amplification matrix be uniformly bounded, i.e. equations (87), (94) and (109). This section discusses various useful results on the norm of G^n concluding with the necessary and sufficient conditions on G to ensure the uniform boundedness of $||G^n||$. These have been derived in two different forms in the work of Kreiss (1962) and the alternative approach of Buchanan (1963a), (1963b). However, owing to the complexity of the Kreiss and Buchanan conditions the various necessary or sufficient conditions which are presented here are often helpful in analysing particular difference schemes. Firstly some simple results are presented which may be useful in some difference schemes.

The following relation may given an explicit expression for the norm required but generally an analytic form will not be available.

$$||G^n|| = \rho^{\frac{1}{2}} (G^{n*} G^n) \qquad \ldots\ldots\ldots\ldots \quad (128)$$

where $\rho(A) = \max |\lambda_i|$, λ_i are the eigenvalues of A. The following well-known matrix result will give bounds on the norm.

$$\rho(G^n) \leq ||G^n|| \leq ||G||^n \qquad \ldots\ldots\ldots \quad (129)$$

In the case where G is a scalar or more generally where G is a normal matrix the equality holds in (129).

An important necessary condition which follows from (129) is the Von Neumann necessary condition on the eigenvalues of G which states that

$$|\mu_i| \leq 1 + \alpha k \qquad 0 < k < \tau$$

$$i = 1, 2, \ldots d, \ \alpha \text{ constant} \qquad \ldots\ldots (130)$$

where μ_i are the eigenvalues of G. This condition is most easily proved by considering an associated matrix A defined by

$$A = e^{-\alpha k} G \qquad \ldots\ldots\ldots\ldots\ldots\ldots (131)$$

$$\alpha = \log K/T$$

It is first shown that a necessary and sufficient condition for

$$||G^n|| \leq K \qquad 0 < k < \tau$$
$$0 \leq nk \leq T \qquad \ldots\ldots (132)$$

is that $||A^\nu|| \leq K_1$ for all positive integer ν. ..... (133)

For the proof of necessity consider $\nu = r.N+S$, $N = T/k$ where $s < N$

$$||A^\nu|| \leq e^{(-\log K).r} \ (||G^N||)^r. \ e^{-\alpha ks}||G^s||$$

$$\leq K^{-r}. K^r. K = K$$

assuming $K \geq 1$

The condition (133) is also sufficient to establish (132) since

$$||G^n|| = e^{\alpha kn} ||A^n||$$

$$\leq K. K_1$$

It follows from the left in equality of (129) that it is necessary that

$$\rho(A) \leq 1$$

Hence $$\rho(e^{-\alpha k} G) \leq 1$$

$$\rho(G) \leq e^{\alpha k} \leq 1 + \alpha k$$

which establishes condition (130). It was assumed that $K \geq 1$ since if $K < 1$ all the eigenvalues of G have modulus less than one and (130) is trivially satisfied.

If we consider the case where G is a normal matrix then

$$\begin{aligned} ||G^n|| &= \rho^n(G) \\ &\leq (1 + K.k)^n \\ &\leq e^{KT} \end{aligned} \qquad \ldots\ldots\ldots\ldots (134)$$

and the norm is bounded above by a quantity which increases exponentially with T. Since this norm defines the amplification factor for errors introduced in the finite difference scheme the error growth may also be exponential with T. For problems in which the true solution is known to be decreasing such growth would be unacceptable and the Von Neumann condition should be strengthened to

$$|\mu_i| \leq 1 \qquad \ldots\ldots\ldots\ldots (135)$$

Condition (135) is used by Strang (1968) as a definition of strong stability. The idea of strong and weak stability is examined in some depth by Stetter (1965). The definition of strong stability used by Stetter is more demanding but is a more accurate reflection of the property desired in a difference scheme. His definition of

strong stability effectively states that when h tends to zero the finite difference scheme has the same amplification properties as the original differential equation. Thus for a scheme with a strongly decreasing solution the condition (135) would not be strong enough to give strong stability in the sense of Stetter.

3.7.2 Sufficient Conditions

It is convenient to have conditions on G rather than G^n for determining stability. However the necessary and sufficient conditions of Kreiss and Buchanan are often complicated to apply in practice and it is useful to have some simpler sufficient conditions. The conditions given below were discussed by Richtmyer (1957) Richtmyer and Morton (1967) and conditions (d) and (e) are special cases of the result of Kato (1960). The conditions are summarised below and the detailed discussion is available in the texts cited. Any of the following conditions is sufficient for the uniform boundedness of G^n and hence by (109) is sufficient for stability.

a) The Von Neumann condition (130) is both necessary and sufficient in the case where G is a normal matrix by the equality in (129)

b) $$||G||^2 = \max_{|\nu|=1} |G\nu|^2 = \max_{|\nu=1|} \nu^* G^* G \nu$$

Hence $||G|| = \rho^{\frac{1}{2}}(G^*G)$ and a sufficient condition is

$$\rho^{\frac{1}{2}}(G^*G) \leq 1 + K.k \qquad 0 < k < \tau \quad .....(136)$$

c) If all elements of G(o,r) are bounded for all r and if G is Lipschitz continuous at k = 0

i.e. $$G(k,r) \leq G(o,r) + M.k \qquad 0 < k \leq \tau$$

where M is a matrix with bounded coefficients independent of k, and if $\|G(o,r)\| \leq 1$ then condition (109) holds. (137)

d) If G has a complete set of eigenvectors and there exists a constant δ such that $\Delta \geq \delta > 0$ where Δ^2 is the Gram determinant of the normalized eigenvectors then the Von Neumann condition is sufficient as well as necessary. (138)

e) If the elements of G(k,r) are bounded $0 < k \leq \tau$ for all r and if all the eigenvalues of G with the possible of one satisfy

$$|\lambda_i| \leq \gamma < 1$$

then the Von Neumann condition is sufficient as well as necessary. (139)

f) The Lax-Wendroff condition states that if G(k,r) is such that $|V^* GV| \leq (1 + Kk)|V|^2$ $\quad 0 < k \leq \tau$ for any vector V then the difference scheme is stable. (140)

3.7.3 Necessary and Sufficient Conditions

Alternative versions of the necessary and sufficient conditions on G for (109) to hold are given by Kreiss (1962) and Buchanan (1963a) and (1963b). The work of Kreiss will be presented first. The Kreiss matrix theorem proves that four conditions on a family of matrices Q are equivalent so that if Q satisfies one condition then it satisfies the other three also. Using $A = e^{-\alpha k} G$ as in equation (131) it can be seen that the first condition is equivalent to condition (109) which is necessary and sufficient for stability. Hence any of the other three is also necessary and sufficient for stability.

A) There exists a constant C such that for all $A \in Q$ and all positive integers ν

$$\|A^\nu\| \leq C \qquad \ldots\ldots\ldots\ldots (141)$$

R) There exists a constant C_R such that for all complex numbers Z with $|Z| < 1$ and for all $A \in Q$ the matrix $(ZA - I)^{-1}$ exists and

$$||(ZA - I)^{-1}|| \leq \frac{C_R}{1 - |Z|} \qquad \ldots\ldots (142)$$

S) There exist constants C_s and C_B and for each matrix $A \in Q$ a matrix S exists for which

i) $\quad \text{Max}\,(||S||, ||S^{-1}||) \leq C_s \quad$ and

ii) $B = SAS^{-1} =$

$$\begin{bmatrix} \lambda_1 & b_{12} & \ldots\ldots & b_{1d} \\ & \lambda_2 & b_{23} \;\; .. & b_{2d} \\ 0 & & & \lambda_d \end{bmatrix}$$

is upper triangular, the eigenvalues are such that

$|\lambda_j| \leq |\lambda_i| \leq 1 \quad i \leq j$ and the off-diagonal elements satisfy

$$|b_{ij}| \leq C_B(1 - |\lambda_i|) \qquad \ldots\ldots (143)$$

H) There exists a constant $C_H > 0$ and for each $A \in Q$ a positive definite hermitian matrix H for which

$$\max\,(||H||, ||H^{-1}||) \leq C_H \quad \text{and}$$

$$A^* HA \leq H \qquad \ldots\ldots (144)$$

A necessary and sufficient condition has also been provided by Buchanan (1963a) (1963b). We first define the property 'nesting' of a set of complex numbers.

Definition The set of complex numbers $\lambda_1 \ldots\ldots \lambda_d$ is said to be nested with nesting constant R if for all i,j such that $1 \leq i \leq j \leq d$.

$$|\lambda_i - \lambda_j| \leq R|\lambda_r - \lambda_s| \qquad r \leq i \quad j \leq s \quad \ldots(145)$$

It is easily verified that any set of d complex numbers can be nested with constant 2^{d+1}.

Theorem If there is a family of matrices Q of d × d matrices A which have been reduced to upper triangular form with the eigenvalues nested with constant R then the family is stable if, and only if, the Von Neumann condition is satisfied $|\lambda_i| \leq 1$ and the off-diagonal elements satisfy

$$|a_{ij}| \leq K_1 \max (1 - |\lambda_i|, 1 - |\lambda_j|, |\lambda_i - \lambda_j|) \quad .....(146)$$

where K_1 is a constant independent of A.

Note that $A = e^{-\alpha k} G$ as before.

Details of the above theory can be found in Richtmyer and Morton (1967) as well as in the texts quoted.

4. Summary and Conclusions

The main contributions to stability theory which have been made in this thesis are summarised below, together with comments on the work and suggestions for further developments.

One of the initial aims of the study was to provide an abstract framework in which a wide variety of problems involving finite difference approximations could readily be discussed. The starting point for this development was the theory on consistency, stability and convergence due to Dahlquist and Henrici for ordinary differential equations, the parallel theory for partial differential equations presented by Lax and Richtmyer and the rather different approach due to Godunov and Ryabenki. The abstract framework which was developed (see Section 2.1) enables the convergence of many types of discretization scheme to be discussed but the investigation was restricted to the topics of ordinary and partial differential equations. The work on the initial value problem for partial differential equations gave several opportunities for the development of the theory and therefore boundary value problems for ordinary or partial differential equations were not considered. Unfortunately the theory cannot now be presented as a new development since it was found that the theory has many similarities to the framework used by Stetter. However, it is hoped that the presentation of this alternative development in detail, together with the

examples of Section 2.2 and 2.3, will still provide a useful contribution to the unification of stability theory for differential equations. It will be noted that the regions S and T are the entire spaces E_h and E_h^0 for those examples discussed here. In the case of non-linear problems the region T may be dependent upon h as discussed in Stetter (1966).

In the example on multi-step methods for ordinary differential equations it is shown that the general definition of stability leads to the 'root condition' of Dahlquist and Henrici as a necessary and sufficient condition for stability. Dahlquist and Henrici define a multi-step method to be stable if it satisfies the 'root condition' but this conditions seems more appropriate as the necessary and sufficient condition to ensure the more fundamental property of stability as defined here. The later version of Henrici's work (1963) is in fact much closer to the ideas of Section 2.1.

The example on the heat conduction equation and subsequent discussion on periodic problems shows one of the gaps in the present theory. The equivalent problem which is analysed for stability does not correspond to the method of computation for a problem with boundary conditions. It may well be that a method of computation based on solving an equivalent problem with periodic replacement of external values, as suggested in Section 3.4, could be advantageous in that the stability properties could be stated with confidence. However, practical experience has shown that the stability results for the

equivalent problem seems to be closely related to the stability results in the problem with boundary conditions.

For the application of Fourier methods of stability analysis and the equivalent matrix analysis of Section 3.6 it is necessary to restrict consideration to certain classes of problem. In the works on Fourier methods it is stated that the problems considered must be such that boundary conditions are absent or can be replaced by periodicity conditions on the initial function. However, a discussion of the problems which are covered by these conditions was not available in the literature and the analysis of Section 3.4 provides a class of problems for which this replacement by an equivalent periodic problem is possible.

In attempting to derive an alternative approach to the L-R theory which was based on a mesh norm rather than the L-R extended function it was found that the results derived by using Fourier methods of stability analysis could also be obtained by methods based on matrix analysis. For the class of problems which are traditionally investigated by Fourier stability methods it is shown that the matrices relating values at t^{n+1} to those at t^{n} have circulant form. The eigenvalues and norms of matrices of this type are derived in Appendix 1 and are used in Section 3.6 to show that the stability condition of uniform boundedness on powers of the amplification matrix can be derived directly without recourse to Fourier analysis. The theory is used for the analysis of stability in terms of a mesh norm, and the extension of a mesh function to form a

function defined for all x, which is necessary for the L-R approach, is not required in this matrix approach. The section on the amplification matrix is of a survey nature and is included because of the importance and frequent use of the stability condition on the amplification matrix.

A problem which has not been examined here but which would prove a useful field for further study is the question of weakly stable schemes. This is defined by Stetter (1965a) in mathematical terms but can be loosely described as the situation where a scheme is stable and convergent but the finite difference scheme is considerably more sensitive to perturbations than the original differential equation. For example a typical stability equation might give a condition

$$\| Z^n \| \leq (1+Kk)^n \| \Lambda \|$$

$$\leq e^{KT} \| \Lambda \|$$

The quantity Λ represents the error such as truncation error and round-off error and according to the above equation this could be subject to a magnification by a factor e^{KT}. Such an error growth would be unacceptable in examples where the true solution is a decreasing quantity such as the heat conduction equation. The bound on the error growth for finite h is the quantity which would give the most valuable information. The work of Stetter shows that a framework such as that given in Section 2.1 is very suitable for developing the ideas

of weak stability and relative stability and it is felt that this field would provide interesting possibilities for further advance in the theory.

APPENDIX 1

A.1 Circulant Matrices

A.1.1 Introduction

In certain partial differential equation problems of initial value type the boundary conditions can be replaced by periodicity conditions. The set of equations for the determination of the values at the mesh points can then be represented by the equations

$$C_1 u^{n+1} = C_0 u^n \qquad \ldots\ldots\ldots\ldots (1)$$

where C_i, $i = 0,1$, have circulant form.

Definition

A matrix C has circulant form if each row is a permutation of the elements $a_1, \ldots a_N$ of the first row so that the elements of row p are

$$C_{pq} = a_k \quad \text{where } k = (q - p+1 \bmod N) \quad \ldots. (2)$$

Only square circulant matrices will be considered in the future discussion. In the simplest case the elements u^n_k of the vector u^n are scalar quantities representing values of the finite difference solution at points of a one-dimensional mesh; the a_k are then scalars. In the case where the solution is a vector quantity each u^n_k is a vector

$$[u^n_{k_1}, \ldots. u^n_{k_i}, \ldots u^n_{k_d}]^T \ ,$$

where $u^n_{k_i}$ is a scalar quantity, and the a_k will then be $d \times d$ matrices. If the solution of the differential equation depends upon an s dimensional variable $x_1, \ldots x_s$, then values of u^n will be

calculated on an s-dimensional mesh. The vector u^n will have $M_1 \times M_2 \ldots M_s$ components u^n_k defined by the vector $k = [k_1, k_2, \ldots k_s]$, $1 \leq k_1 \leq M_1$, $1 = 1,2,\ldots s$. The quantity u^n_k at a particular mesh point will be a scalar or vector as before according to the nature of the solution of the differential equation. The dimension of C in the most complex situation will be $M_1 \times M_2 \ldots M_s \times d$ and the elements of the first row will be a_k, $k = [k_1, k_2, \ldots k_s]$, $1 \leq k_1 \leq M_1$, where each a_k is a $d \times d$ matrix. The elements of the matrix C are defined by the equation $C_{pq} = a_k$ where $p = [p_1, p_2, \ldots p_s]$, $q = [q_1, q_2, \ldots q_s]$ and k defined by

$$k_1 = (q_1 - p_1 + 1 \bmod M_1), \quad 1 = 1,2,\ldots s. \quad \ldots(3)$$

This section will give results on the eigenvalues and eigenvectors or circulant matrices and the norm defined by

$$||C|| = \min_u \left[\frac{||Cu||_2}{||u||_2} \right] \quad \ldots\ldots\ldots\ldots\ldots (4)$$

where $||u||_2$ is the Euclidean vector norm.

A1.2 The Determinant of a Circulant Matrix

The results presented are extensions of the simple result given, for example, by Aitken (1958).

Result 2.1

In the case where the a_k are scalars with a simple subscript (i.e. $d = 1$, $s = 1$) it is easily shown that

$$\det C = \prod_{r=1}^{M_1} \sum_{k=1}^{M_1} a_k w_r^{k-1} \quad \ldots\ldots\ldots\ldots (5)$$

where $w_r = e^{2\pi ri/M_1}$ is a complex M_1^{th} root of unity, and C is of order M_1.

Proof. Let Ω be a diagonal matrix with values $1, w_r, \ldots w_r^{M_1-1}$ on the leading diagonal. The matrix $C' = \Omega^{-1} C \Omega$ has the same determinant as C and the elements of C' are given by

$$C'_{pq} = a_k w_r^{k-1} \qquad k = (q - p+1 \bmod M_1) \quad \ldots\ldots(6)$$

If we form a matrix C'' by adding each subsequent row to the first row then $\det C'' = \det C$ and each element of the first row equals $\sum_{k=1}^{M_1} a_k w_r^{k-1}$ which is therefore a factor of the determinant. Thus for $r = 1, \ldots M_1$, we have the M_1 factors of equation (5).

Result 2.2

Result 1 is easily extended to the case where the quantity u_k^n is a vector of dimension d in which case the a_k are d × d matrices and

$$\det C = \prod_{r=1}^{M_1} D_r$$

where $$D_r = \det \sum_{k=1}^{M_1} a_k w_r^{k-1} \qquad \ldots\ldots\ldots\ldots(7)$$

Proof The argument follows the same lines but the elements of Ω are now defined by the block diagonal matrix with diagonal elements $I_d, w_r I_d, \ldots w_r^{M_1-1} I_d$. The process of adding the rows together is performed as a matrix addition of d × d matrix elements, i.e. the elements of the first row of C'' are given by

$$C''_{1q} = \sum_{k=0}^{M_1-1} C'_{1+kd,q}$$

The matrix C'' has for its first d rows M_1 repetitions of the matrix $\sum_{k=1}^{M_1} a_k w_r^{k-1}$ and using Laplaces expansion along the first row of d × d submatrices gives a factor of $\det(C'')$ (and therefore of $\det(C')$ and $\det(C)$) equal to

$$\det \sum_{k=1}^{M_1} a_k w_r^{k-1}$$

Result 2.3

In the case where the quantity u_j^n is a vector of dimension d and $j = [j_1, j_2, \ldots j_s]$, $1 \leq j_1 \leq M_1$ then we have the following result.

$$\det C = \prod_{r_1=1}^{M_1} \cdots \prod_{r_s=1}^{M_s} D_r$$

$$D_r = \det \sum_{k_1=1}^{M_1} \cdots \sum_{k_s=1}^{M_s} a_k \, w_r^{k-1} \qquad \ldots\ldots\ldots\ldots (9)$$

where for vectors r and k we have $w_r^{k-1} = w_{r_1}^{k_1-1} \cdot w_{r_2}^{k_2-1} \cdots w_{r_s}^{k_s-1}$ ..(10)

To simplify the notation in the vector case let

$$\prod_{r=1}^{M} \equiv \prod_{r_1=1}^{M_1} \cdots \prod_{r_s=1}^{M_s}$$

and $$\sum_{k=1}^{M} \equiv \sum_{k_1=1}^{M_1} \cdots \sum_{k_s=1}^{M_s}$$

Hence the result (9) above can be written

$$\det C = \prod_{r=1}^{M} \sum_{k=1}^{M} a_k \, w_r^{k-1} \qquad \ldots\ldots\ldots (11)$$

Proof. The proof follows the same lines as the proofs for the previous results. A matrix Ω is defined which has only diagonal elements which are defined in block diagonal form by the diagonal elements given by $P_1(r)$ where $P_1(r)$ is defined by

$$P_l(r) = \{ P_{l+1}(r), P_{l+1}(r) \cdot w_{r_l}, \ldots P_{l+1}(r) \, w_{r_l}^{M_l-1} \} \quad \ldots\ldots (12)$$

$l = 1, 2, \ldots s$

and $P_{s+1}(r) \equiv I_d$.

The matrix C' is formed as before, $C' = \Omega^{-1} C \Omega$ and the $M_1 \times M_2 \ldots M_s$ rows of $d \times d$ matrices are added to give the first row of C'' which can be expressed as

$$C''_{1,q} = \sum_{k=1}^{M} C'_{1+k.d,q} \qquad \ldots\ldots\ldots\ldots \quad (13)$$

where the summation is in vector form as defined by equation (11) and $k.d = (k_1 + k_2 \ldots k_s) \times d$. The first d rows of C'' contain $M_1 \times M_2 \ldots M_s$ elements which are identical and equal to

$$\sum_{k=1}^{M} a_k \, w_r^{k-1}$$

and hence by the previous arguments this represents one factor of det(C). The product of such factors over all values of the vector r gives the value of the determinant of C.

$$\det(C) = \prod_{r=1}^{M} \det \left(\sum_{k=1}^{M} a_k \, w_r^{k-1} \right) \qquad \ldots\ldots\ldots\ldots \quad (14)$$

A.1.3 <u>The Eigenvalues and Eigenvectors</u>

We consider a modified matrix

$$C\mu = C - \mu I \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (15)$$

which is of circulant form so that the results of the previous section can be applied. Result 2.3 includes the case where $s = 1$ or $d = 1$ and therefore the three situations described in section A.1.1 can be discussed together.

<u>Result 3.1</u>

The eigenvalues of C are found from the equation $\det C_\mu = 0$ where

$$\det C_\mu = \prod_{r=1}^{M} \left[\det \left(\sum_{k=1}^{M} a_k \, w_r^{k-1} - \mu I_d \right) \right] \qquad \ldots\ldots\ldots (16)$$

There are $M_1 \times \ldots\ldots M_s \times d$ eigenvalues μ_{r_b} defined for $[r_1, r_2, \ldots\ldots r_s]$, $1 \leq r_1 \leq M_1$, $b = 1,2, \ldots\ldots d$. The eigenvectors v_{r_b} can be defined by using the set of $d \times d$ matrices defined by equation (12). Let $R_1(r)$ be a matrix of d rows defined by the following matrix

$$R_1(r) = [\, R_{1+r}(r)\; R_{1+1}(r).\, w_{r_1} \;\ldots\ldots\; R_{1+1}(r).\, w_{r_1}^{M_1-1}\,] \quad \ldots\ldots(17)$$

$1 = 1,2, \ldots\ldots s$

where $R_{s+1}(r) \equiv I_d$

If ν_{rb} is the eigenvector of dimension d defined by the equation

$$\Big(\sum_{k=1}^{M} a_k\, w_r^{k-1} - \mu_{rb}\Big)\, \nu_{rb} = 0 \qquad \ldots\ldots\ldots\ldots\ldots\ldots(18)$$

$r_1 = 1,2,\ldots\ldots M_1$, $1 = 1,2, \ldots\ldots s$; $b = 1,2,\ldots\ldots d$

then the eigenvectors v_{rb} of C are defined by

$$(v_{rb})^T = \nu_{rb}^T\; R_1(r) \qquad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots (19)$$

Proof. Equation (16) follows directly from equation (14) by replacing the diagonal elements of C_μ by $a_{11\ldots\ldots1} - \mu I_d$.

In the case of $s = d = 1$ it can be seen by inspection that the eigenvector is given by

$$v_r^T = [\, 1, w_r, \ldots\ldots w_r^{M_1-1}] \equiv R_1(r)$$

In the case of general s and d we proceed as follows. Let $a[k_1, k_{1+1}, \ldots\ldots k_s]$ be a circulant matrix with vector subscript k and let the matrix of submatrices which is circulant with respect to k_1 be defined by a first row with elements

$$a[k_{1+1}, \ldots\ldots k_s]_1 \quad - - - - - - \quad a[k_{1+1}, \ldots\ldots k_s]_{M_1}$$

In cases where no confusion can arise the matrix is represented as $a[k_1, K]$ with elements

$$a[K]_1 \qquad a[K]_2 \text{ - - - - - } a[K]_{M_1}$$

The matrix C is a circulant matrix $C = a^s[k_1, K_{s-1}]$ which has eigenvalues and eigenvectors given by the equation

$$(a^s[k_1, K_{s-1}] - \mu)\, v^o = 0 \quad \text{.....(20)}$$

i.e.

$$a^s[K_{s-1}]_1 v_1^o + \text{.....} + a^s[K_{s-1}]_{M_1} v_{M_1}^o = \mu v_1^o$$

$$\text{- - - - - - - - - - - - - - - - - - - -}$$

$$a^s[K_{s-1}]_2 v_1^o \text{ .....} + a^s[K_{s-1}]_1 v_{M_1}^o = \mu v_{M_1}^o \text{ ....(21)}$$

It is first shown that an eigenvalue μ_{r_1} is also an eigenvalue of the matrix $a^{s-1}[K_{s-1}] = \sum_{k_1=1}^{M_1} a^s[K_{s-1}]_{k_1} w_{r_1}^{k_1-1}$

i.e. μ_{r_1} is given by the solutions of

$$\left[\sum_{k_1=1}^{M_1} a^s[K_{s-1}]_{k_1} w_{r_1}^{k_1-1} - \mu_{r_1} \right] v_{r_1}^1 = 0 \quad \text{.....(22)}$$

The subscript r_1 on μ and v^1 will be omitted in the future analysis in order to simplify the subscript notation. The eigenvalues of equation (20) are given by the roots of $\det(C_\mu) = 0$ where C_μ can be written

$$a^s[K_{s-1}]_1 - \mu I_{s-1} \text{ .....} a^s[K_{s-1}]_{M_1}$$

$$\text{- - - - - - - - - - - - - - - - - - - -}$$

$$a^s[K_{s-1}]_2 \text{ .....} a^s[K_{s-1}]_1 - \mu I_{s-1} \text{ .....(23)}$$

where I_{s-1} is the unit matrix of dimension $M_2 \times M_3 \times \text{....} M_s \times d$

A matrix Ω is defined which has block diagonal form with elements $[I_{s-1}, w_{r_1} I_{s-1}, \ldots\ldots w_{r_1}^{M_1-1} I_{s-1}]$ and the matrix $C'_\mu = \Omega^{-1} C_\mu \Omega$ is formed. A new matrix C''_μ is now formed by adding to the first row of $M_2 \times M_3$ x ..... $M_s \times d$ submatrices the following $M_1 - 1$ rows of submatrices. By the previous arguments it follows that a factor of $\det (C''_\mu)$ and hence of $\det (C_\mu)$ is

$$\det \left(\sum_{k_1=1}^{M_1} a^s[K_{s-1}]_{k_1} w_{r_1}^{k_1-1} - \mu \right) \qquad \ldots\ldots\ldots(24)$$

and hence the eigenvalue of $a^s[k_1, K_{s-1}]$ is also an eigenvalue of $a^{s-1}[K_{s-1}]$. Note that due to the dependence of the matrix $a^{s-1}[K_{s-1}]$ on r_1 there are M_1 such matrices each of which has a set of $M_2 \times M_3 \ldots M_s \times d$ eigenvalues which are a subset of the eigenvalues of $a^s[k_1, K_{s-1}]$.

A solution to equations (21) is assumed of the form

$$v_k^o = v^1 \, Q(r,k)$$

where v^1 has dimension $M_2 \times M_3 \ldots \times M_s \times d$ and $Q(r,k)$ is a scalar. This expression is substituted in equations together with the expression for μv^1 given by equations (22). Thus we have

$$\begin{aligned} a^s[K_{s-1}]_1 v^1 . \, Q(r,1) \ldots + a^s[K_{s-1}]_{M_1} v^1 \, Q(r,M_1) \\ = \left[\sum_{k_1=1}^{M_1} a^s[K_{s-1}]_{k_1} w_{r_1}^{k_1-1} \right] v^1 . \, Q(r,1) \end{aligned}$$

- -

$$\begin{aligned} a^s[K_{s-1}]_2 v^1 . Q(r,1) \ldots + a^s[K_{s-1}]_1 v^1 \, Q(r,M_1) \\ = \left[\sum_{k_1=1}^{M_1} a^s[K_{s-1}]_{k_1} w_{r_1}^{k_1-1} \right] v^1 . \, Q(r,M_1) \end{aligned}$$

$$\ldots\ldots\ldots(25)$$

and these equations are satisfied if

$$Q(r,k_1) = w_{r_1}^{k_1-1} \quad A(r,1)$$

Thus the eigenvector of (20) is given by

$$(v^0)^T = [\ (v^1)^T,\ w_{r_1}(v^1)^T,\ \ldots\ldots\ w_{r_1}^{M_1-1}\ (v^1)^T] \quad \ldots\ldots\ (26)$$

The matrix $a^{s-1}[K_{s-1}] \equiv a^{s-1}[k_2, K_{s-2}]$ can also be written in terms of submatrices and this leads to the equation for the eigenvectors v^1, i.e.

$$\begin{aligned}
&a^{s-1}[K_{s-2}]_1\ v_1^1 + \ldots\ldots\ldots a^{s-1}[K_{s-2}]_{M_2}\ v_{M_2}^1 = \mu v_1^1 \\
&- \\
&a^{s-1}[K_{s-2}]_2\ v_1^1 + \ldots\ldots\ldots a^{s-1}[K_{s-2}]_1\ v_{M_2}^1 = \mu v_{M_2}^1
\end{aligned}$$

$$\ldots\ldots(27)$$

and these equations have the same form as equations (21). Using the previous arguments the vector v^1 can be expressed in terms of the eigenvector v^2 given by equation

$$[\sum_{k_1=1}^{M_1}\ \sum_{k_2=1}^{M_2}\ a^{s-1}[K_{s-2}]_{k_1 k_2}\ w_{r_1}^{k_1-1}\ w_{r_2}^{k_2-1} - \mu]\ v^2 = 0 \quad \ldots\ldots(28)$$

In general form we have

$$(v^{l-1})^T = [(v^l)^T,\ w_{r_l}(v^l)^T,\ \ldots\ldots\ w_{r_l}^{M_l-1}\ (v^l)^T] \quad \ldots\ldots(29)$$

where $v^s = I_d \cdot \nu$ and ν is the eigenvector of the equations

$$(\sum_{k=1}^{M} a_k\ w_r^{k-1} - \mu)\ \nu = 0$$

Letting $R_{s+1}(r) = I_d$ and

$$R_1(r) = [R_{1+1}(r),\ w_{r_1} R_{1+1}(r),\ \ldots\ldots\ w_{r_1}^{M_1-1} R_{1+1}(r)]$$

it can be seen that

$$(v^1)^T = \nu^T R_{1+1}(r)$$

and hence $(v^0)^T = \nu^T R_1(r)$ .....(30)

A.1.4 The Norm of a Circulant Matrix

For the norm defined by equation (4)

$$||C|| = \rho^{\frac{1}{2}} (C^*C) \qquad .....(31)$$

where ρ is the maximum modulus of the eigenvalues of C.

Consider the matrices $B = C^*C$. Using equation (3)

$$b_{pq} = \sum_{k=1}^{M} a^*_{p-k+1}\ a_{q-k+1} \qquad ..... (32)$$

where b_{pq} is a square matrix of dimension d. In the case $s > 1$ the summation and subscripts are multiple as before. Since $b_{p+g,\ q+g} = b_{p,q}$ the matrix B has circulant form, hence

$$||C|| = \max_r\ \rho^{\frac{1}{2}} \left(\sum_{k=1}^{M} a^*_k\ w_r^{*k-1} \sum_{k=1}^{M} a_k\ w_r^{k-1} \right)$$

$$= \max_r\ ||G(M,r)||_2 \qquad .....(33)$$

where $G(M,r) = \sum_{k=1}^{M} a_k e^{2\pi r(k-1)i/M}$ .....(34)

where $\frac{rk}{M}$ is a scalar product

$$\left[\frac{r_1}{M_1} \cdot k_1 + \frac{r_2}{M_2} \cdot k_2 \ldots\ldots\ldots \frac{r_s}{M_s} \cdot k_s \right]$$

REFERENCES

AITKEN, A. C., 1958, Determinants and matrices. Oliver and Boyd, Edinburgh and London.

BUCHANAN, M. L., 1963a, A necessary and sufficient condition for stability of difference schemes for initial value problems. J.SIAM., vol.11, 474 - 501.

BUCHANAN, M. L., 1963b, A necessary and sufficient condition for stability of difference schemes for initial value problems. J.SIAM., vol.11, 919 - 935.

DAHLQUIST, G., 1956, Convergence and stability in the numerical integration of ordinary differential equations. Math.Scand., vol.4, 33 - 53.

DAHLQUIST, G., 1959, Stability and error bounds in the numerical integration of ordinary differential equations. Trans.Roy.Inst.Technol., Stockholm, No.130.

EISEN, D., 1966, Stability and convergence of finite difference schemes with singular coefficients. SIAM J.Num.Anal., vol.3, 545 - 552.

EISEN, D., 1967a, The equivalence of stability and convergence for finite difference schemes with singular coefficients. Num.Math., vol.10, 20 - 29.

EISEN, D., 1967b, Stability and convergence of finite difference schemes with singular coefficients. Num.Math., vol.10, 397 - 409.

FOX, L., 1962, Numerical solution of ordinary and partial differential equations. Pergamon Press, Oxford.

GODUNOV, S. K., and RYABENKI, V. S., 1964, Introduction to the theory of difference schemes. North Holland, Amsterdam.

HAMMING, R. W., 1962, Numerical Methods for Scientists and Engineers. McGraw Hill, New York.

HENRICI, P., 1962, Discrete variable methods in ordinary differential equations. Wiley, New York.

HENRICI, P., 1963, Error propagation for difference methods. Wiley, New York.

ISAACSON, E., and KELLER, H. B., 1966, Analysis of numerical methods. Wiley, New York.

JOHN, F., 1967, Lectures on advanced numerical analysis. Gordon and Breach, New York.

KATO, T., 1960, Estimation of iterated matrices with application to the Von Neumann condition. Num.Math., vol.2, 22 - 29.

KREISS, H. O., 1962, Über die Stabilitäts definition für Differenzen gleichungen die partielle Differentialgleichungen approximieren. BIT, vol.2, 153 - 181.

LAX, P. D., and RICHTMYER, R. D., 1956, Survey of the stability of linear finite difference schemes. Comm.Pure Appl.Math., vol.9, 267 - 293.

Modern Computing Methods, 1961, H.M.S.O., London.

RALSTON, A., and WILF, H. S., 1960, Mathematical methods for digital computers. vol.1, Wiley, New York.

RICHTMYER, R. D., 1957, Difference methods for initial value problems. Interscience, New York.

RICHTMYER, R. D., and MORTON, K. W., 1967, Difference methods for initial value problems. Interscience, New York.

RUTHERFORD, D. E., 1951, Some continuant determinants arising in Physics and Chemistry. Proc.Roy.Soc.(Edin), vol.63, 232 - 241.

STETTER, H. J., 1965a, Strong and weak stability in discretization algorithms. SIAM J.Num.Anal., vol.2, 265 - 280.

STETTER, H. J., 1965b, Asympotic expansions for the error of discretization algorithms for non-linear functional equations. Num.Math., vol.7, 18 - 31.

STETTER, H. J., 1966, Stability of non-linear discretization algorithms, in Numerical solution of partial differential equations. ed. Bramble, J.H., Academic Press, New York.

STRANG, G., 1964, Accurate partial difference mathods II. Non-linear problems. Num.Math., vol.6, 37 - 46.

STRANG, G., 1968, On the construction and comparison of difference schemes. SIAM J.Num.Anal., vol.5, 506 - 517.

WENDROFF, B., 1966, Theoretical numerical analysis. Academic Press, New York.