# An Investigation Into The Stability and Accuracy of Boundary Approximations Used In The Numerical Solution Of Hyperbolic Initial-Boundary Value Problems

Scott McMillan Jamieson

A Thesis Submitted In Partial Fulfilment
Of The Requirements For The
Degree Of Doctor of Philosophy

**Department Of Mathematics**
**University Of Strathclyde**
**1984**

*To*

*My Parents*

*Thank You*

# Acknowledgements

My sincere thanks go to my tutor Dr. David M. Sloan for his encouragement and support throughout my time at university and especially in the work of this thesis.

Thanks are due also to Mrs. Lynne Westwood who, with incredible patience, typed the manuscript.

I am indebted to Strathclyde University for the necessary financial support.

*"Aye free, aff han' your story tell,*
*When wi' a bosom crony;*
*But still keep something to yoursel*
*Ye scarely tell to ony..."*

from "Epistle to a Young Friend"
Robert Burns 1759-1796

# CONTENTS

## ABSTRACT

A comparison of boundary approximations used in numerical solution of one-dimensional hyperbolic systems of partial differential equations is undertaken. Stability and accuracy studies of the boundary approximations are conducted for a variety of interior schemes. A new fourth order accurate finite-element scheme is proposed.

## 0.1 INTRODUCTION

In this thesis we examine some of the difficulties inherent in the construction of a numerical approximation to the solution of one-dimensional initial-boundary value hyperbolic problems. The governing system of partial differential equations may be written as

$$\begin{aligned} \frac{\partial}{\partial t}\underset{\sim}{u}(x,t) &= A(\underset{\sim}{u},x,t)\frac{\partial}{\partial x}\underset{\sim}{u}(x,t) &&, (x,t) \in D \times [0,T] \\ \underset{\sim}{u}(x,0) &= \underset{\sim}{f}(x) &&, x \in D \\ B(x,t)\underset{\sim}{u}(x,t) &= \underset{\sim}{g}(x,t) &&, (x,t) \in \partial D \times [0,T] \end{aligned} \tag{0.1}$$

where D is some bounded interval of $\mathbb{R}$ and B is a matrix operator of appropriate dimensions. Multi-dimensional analogues of (0.1) occur frequently in the equations of fluid dynamics, and by studying system (0.1) we seek to develop results that may be extended to practical situations.

We are primarily concerned with the construction and comparison of the approximate boundary values required to obtain the numerical solution of system (0.1) when D is a bounded region. In particular we examine those systems (0.1) that exhibit differing time scales. The physical problems we are concerned with are therefore those which exhibit solution waves, travelling through the physical domain with different speeds and directions and are identified by the matrix A possessing non-equal eigenvalues. A typical example being the equations of the atmosphere in meteorology where there are relatively high speed gravity waves. These waves have a negligible effect on the solution of the governing differential equations but can have a crucial effect on any numerical method of solution. In this thesis we consider the effect the nature of the incident waves on any boundary has on any required boundary approximation.

Many numerical difference schemes can be shown to provide an unstable approximation to the pure initial-value problem given by a linearised version of (0.1) unless an inequality involving the mesh ratio and largest eigenvalue, in modulus , of A is satisfied. This is commonly termed the Courant, Fredrichs, Lewy (CFL) condition. The importance of the concept of stability is encapsulated in the Lax equivalence theorem which guarantees that any numerical solution obtained from a consistent, stable approximation to a linear, constant coefficient difference scheme will converge to the exact differential solution. It should be emphasized that the use of a finite difference scheme that is stable for the pure initial-value problem in no way ensures that the approximation, to a problem defined on a bounded domain, will be stable. Instabilities may be introduced, through the boundary representation, which will be transported along the characteristics and destroy the entire approximation. To ascertain the stability properties of any formulation used on $\partial D$ we invoke the results of Gustafsson, Kreiss and Sundström [1972] when A has distinct real eigenvalues. In this case the stability condition reduces to an algebraic determinant equation, the roots of which may be analysed according to various stability criteria.

In Chapter 1 we review the requirements necessary to guarantee the existence and uniqueness of the solution $\underset{\sim}{u}(x,t)$ of (0.1). The primary pursuit of this thesis is the study of numerical boundary approximations used in the solution of hyperbolic problems exhibiting differing time-scales and to this end the determinant condition of Gustafsson et al [1972] for finite difference approximations is derived. The extension of this result (Strikwerda [1980]) to a method-of-lines approximation is also given. A further approach uses the results of Cronin [1980] and provides a partial stability analysis by examining the eigenvalues of the amplification matrix.

The roots of the algebraic determinant equation derived in Chapter 1 can be obtained analytically only in the simplest of cases. In the majority of situations numerical methods are required and in Chapter 2 a variety of new and existing techniques are presented.

As a test problem for the remaining chapters of this thesis we consider the initial-value problem

$$\begin{aligned} &\frac{\partial}{\partial t}\underset{\sim}{u} = \tfrac{1}{2}\begin{bmatrix} q-1 & q+1 \\ q+1 & q-1 \end{bmatrix}\frac{\partial}{\partial x}\underset{\sim}{u} \qquad (x,t) \in D \times [0,T] \\ &\underset{\sim}{u}(x,0) = \underset{\sim}{f}(x), \quad q(>1) \in \mathbb{R} \qquad\qquad (0.2) \\ &\underset{\sim}{u}(x,t) = (u(x,t),v(x,t))^T \end{aligned}$$

which has characteristic velocities 1 and -q and so exhibits differing time-scales, as required. Any two variable constant-coefficient linear hyperbolic problem (0.1) can be reduced to this form by an appropriate rescaling. We examine the model left boundary problem, where $D = \{x \in \mathbb{R} : x \geqslant 0\}$ and where the fast wave is incident at $x = 0$, separately from the model right boundary problem given by $D = \{x \in \mathbb{R} : x \leqslant 1\}$ which has the slow wave incident at $x = 1$. This separation is performed to allow the computational significance of the fast wave being incident or reflected to be determined. Along the characteristics $\frac{dx}{dt} = 1$ and $\frac{dx}{dt} = -q$ the respective quantities u(x,t) - v(x,t) and u(x,t) + v(x,t) are conserved. These define the so-called Riemann Invariants.

In Chapter 3 the results of Chapters 1 and 2 are invoked when the interior difference scheme applied to (0.2) is that devised by Lax and Wendroff [1960]. A variety of boundary approximations are examined and those that are based on the Riemann Invariants are shown to be the most desirable. For a particular class of boundary approximations describing the extrapolation of a linear combination of u(x,t) and v(x,t) the right boundary is shown to be more sensitive to the choice of the extrapolated quantity. The results of this chapter may be considered as an extension of the work of Sloan [1980].

Whilst the Lax-Wendroff method is simple to apply and provides accurate results its use may, as a result of the associated CFL condition, be computationally expensive. This is due to the restriction on the time increment $\Delta t$, especially when $q >> 1$. This drawback may be avoided by filtering the quantity transported along the fast characteristic out of the initial data using the Bounded Derivative method of Kreiss [1979]. This approach has been applied successfully in meteorological problems by Browning, Kreiss and Kasahara [1980]. For such suitably initialised problems numerical solutions may be obtained using the asymptotic expansion of Gustafsson [1980]. To solve the problem in the original, non-filtered, form the split-explicit approach of Ookochi and Matsumura [1980] could be adopted. This technique involves the partitioning of the differential equations into separate fast and slow wave components. The resulting numerical integration is performed separately on each partition for a different value of $\Delta t$. The authors demonstrate that accurate results can be obtained with a significant reduction in computational cost. We, however, adopt the semi-implicit approach of Kwizak and Robert [1971] where that part of the differential equations contributing the fast wave component is approximated by an implicit technique and all others treated explicitly. In Chapter 4 the standard Leap-frog method is modified using the above approach and is enhanced further through the use of a staggered mesh. Stable and accurate boundary approximations are developed with those relating to the characteristic formulation being among the most appropriate.

In Chapter 5 we present a new finite element method which is designed specially for problems of the form (0.2). A fourth order approximation in both space and time to the solution of the scalar advection equation is developed first. For the appropriate extension to (0.2) accurate and stable boundary techniques are constructed. For symmetric problems, boundary approximations based on a finite element reasoning are the most accurate, however as q increases we again show the desirability of characteristic formulations.

In Chapter 6 we apply the results of the three previous chapters to a simple non-linear system.

## 0.2 Notation and Definitions.

In this thesis we consider various problems defined on the sets of integer, real or complex numbers, denoted by $\mathbb{Z}$, $\mathbb{R}$ and $\mathbb{C}$ respectively, or multidimensional extensions thereof. A vector function u, defined on such a space, will have the specific domain of dependence defined in the text and will be denoted by $\underset{\sim}{u}$.

Consider now the continuous differential problem. To develop the theory in Chapter 1 we require the $L_p(D)$ norm defined by

$$\|\underset{\sim}{u}\|_{p,D} = \left[\int_D |\underset{\sim}{u}(x)|^p dx\right]^{1/p}, \qquad 1 \leq p < \infty \tag{0.3}$$

where D is some set of real numbers. Unless stated otherwise we will consider p = 2 and will drop the first subscript in the left side of (0.3). The continuous Fourier transform of a function $\underset{\sim}{u} \in L_2(\mathbb{R}^m)$ is defined by

$$\hat{\underset{\sim}{u}}(\omega) = \left(\frac{1}{2\pi}\right)^{m/2} \int_{\mathbb{R}^m} u(x) e^{-i\omega x} dx$$

with the associated inverse defined by

$$\underset{\sim}{u}(x) = \left(\frac{1}{2\pi}\right)^{m/2} \int_{\mathbb{R}^m} \hat{\underset{\sim}{u}}(x) e^{i\omega x} d\omega.$$

It is of interest to note that for $\underset{\sim}{u} \in L_2(\mathbb{R}^m)$ we have that $\hat{\underset{\sim}{u}} \in L_2(\mathbb{R}^m)$ and Parseval's equality

$$\|\hat{\underset{\sim}{u}}\|_{\mathbb{R}^m} = \|\underset{\sim}{u}\|_{\mathbb{R}^m}$$

holds.

We also require the Laplace transform of u, defined by

$$\hat{u} = \int_0^\infty e^{-st} u(t) dt, \qquad \mathbb{R}e(s) > \text{some constant } c,$$

to remove the time variable in several situations. We use the symbol ^ arbitrarily to denote a transformed variable. The precise transform being applied will be made clear in the text.

A system of the form (0.1) is said to be hyperbolic if all the eigenvalues of A are real and there exists a continuous transformation S such that $S^{-1}AS$ is diagonal and for some constant k, we have

$$\|S\| < k \quad \text{and} \quad \|S^{-1}\| < k.$$

System (0.1) is strictly hyperbolic if it is hyperbolic with distinct eigenvalues. The definition of hyperbolicity extends to multidimensional problems by effectively requiring that the problem is hyperbolic in each one-dimensional direction. The results reproduced in Chapter 1 are given for such multidimensional problems.

Consider now the discrete problem defined on $\mathbb{R} \times [0,T]$. We construct a two-dimensional mesh with increments $\Delta x$ and $\Delta t$ upon which we approximate $\underset{\sim}{u}(x,t)$ by the grid function $\underset{\sim}{U}_\Delta$. The discrete $L_2$ norm denoted by $L_2(\Delta x)$ is defined

$$\|\underset{\sim}{U}_\Delta\|_h^2 = \sum_j \Delta x |\underset{\sim}{U}(x_j)|^2$$

and where $|.|^2$ is the Euclidean norm of a vector. We reserve the subscript h and the argument $\Delta x$ for discrete norms.

In Chapter 1 we require the one-dimensional discrete Fourier transform defined by

$$\hat{\underset{\sim}{U}}_\Delta(\omega) = \frac{1}{(2\pi)^{\frac{1}{2}}} \sum_j \Delta x \underset{\sim}{U}_\Delta(x_j) e^{-i\omega x_j}, \qquad \omega \in \mathbb{R}$$

where the summation is performed over the spatial grid points.

For convenience we denote by $\utilde{U}_j^n$ the value of $\utilde{U}_\Delta$ at $x = x_j$, $t = t_n$ ($j \in \mathbb{Z}$, $t = \mathbf{Z}^+$) and we define the shift operator $E^\alpha$ by $\utilde{U}_{j+\alpha} = E^\alpha \utilde{U}_j$ ($\alpha \in \mathbf{Z}$). We also define the centred difference operators

$$\delta^2 \utilde{U}_j = \frac{1}{\Delta x^2}(\utilde{U}_{j+1} - 2\utilde{U}_j + \utilde{U}_{j-1}),$$

$$\Delta_o \utilde{U}_j = \frac{1}{2\Delta x}(\utilde{U}_{j+1} - \utilde{U}_{j-1}).$$

# CHAPTER 1

# THEORETICAL RESULTS FOR THE CONTINUOUS AND DISCRETE PROBLEMS

## 1.1 The Differential Problem

In this section we consider what requirements are necessary for the solution of the linear hyperbolic system of partial differential equations

$$\begin{aligned} \underset{\sim}{u}_t &= \sum_{j=1}^{m} A_j(\underset{\sim}{x},t)\underset{\sim}{u}_{x_j} + \underset{\sim}{F}(\underset{\sim}{x},\underset{\sim}{u},t), \quad (\underset{\sim}{x},t) \in D \times [0,T] \\ \underset{\sim}{u}(\underset{\sim}{x},0) &= \underset{\sim}{f}(\underset{\sim}{x}), \qquad \underset{\sim}{x} \in D \end{aligned} \tag{1.1}$$

where $\underset{\sim}{u}(\underset{\sim}{x},t) = (u_1(\underset{\sim}{x},t),\ldots,u_n(\underset{\sim}{x},t))^T$ and $\underset{\sim}{x} = (x_1,\ldots,x_m)$, each $A_j(\underset{\sim}{x},t)$ is a square matrix of order n and D is some arbitrary real smooth bounded region, to exist and be uniquely determined. In other words, what additional information do we require, including boundary conditions on $\partial D$, so that (1.1) defines a well-posed problem. If the problem (1.1) is not well-posed then it will not be possible to obtain a meaningful well-behaved numerical solution to (1.1). The vague adjectives will be given precise definitions later.

### 1.1.1 The Cauchy Problem

Initially we consider D to be the infinite domain $\mathbb{R}^m$ and (1.1) is said to be a pure Cauchy problem. This type of problem is said to be well-posed if the solution, for all time, can be estimated in terms of the initial data. This may be expressed as

Definition 1.1. The Cauchy problem of (1.1) with $\underset{\sim}{F} = \underset{\sim}{0}$ is well-posed for all bounded initial values, if there exists a constant K such that for all solutions $\underset{\sim}{u}(\underset{\sim}{x},t)$ and all time, the estimate

$$\|\underset{\sim}{u}(\underset{\sim}{x},t)\|_{\mathbb{R}^m} \leqslant K\,\|\underset{\sim}{f}(\underset{\sim}{x})\|_{\mathbb{R}^m} \tag{1.2}$$

holds.

The theory necessary to verify (1.2) is only well developed for constant coefficient systems, symmetric or strictly hyperbolic variable coefficient systems. One possible approach is through the method of characteristics. For m > 1 however this is not straightforward and it is easier to use Fourier analysis.

The constant coefficient homogeneous form of (1.1) written as

$$\begin{aligned} \underset{\sim}{u}_t &= P\left(\frac{\partial}{\partial \underset{\sim}{x}}\right)\underset{\sim}{u} \quad , \quad (\underset{\sim}{x},t) \in \mathbb{R}^m \times [0,T] \\ \underset{\sim}{u}(\underset{\sim}{x},0) &= \underset{\sim}{f}(\underset{\sim}{x}), \quad \underset{\sim}{x} \in \mathbb{R}^m \end{aligned} \tag{1.3}$$

where $P(\frac{\partial}{\partial x}) = \sum_{j=1}^{m} A_j \frac{\partial}{\partial x_j}$ , is well-posed. This may be seen by Fourier transforming with respect to the spatial variables $\underset{\sim}{x}$, using the real dual variable $\underset{\sim}{\omega}$, and verifying that there are constants K and $\alpha$ such that

$$\max_{\underset{\approx}{\omega}} |e^{P(i\underset{\sim}{\omega})t}| \leqslant Ke^{\alpha t}. \tag{1.4}$$

Inequality (1.4) is a necessary and sufficient algebraic condition for well-posedness of the Cauchy problem (Kreiss and Oliger [1973]).

For symmetric or strictly hyperbolic variable coefficient problems we have

<u>Theorem 1.1</u> (Kreiss [1979]). Assume for each j that $A_j(\underset{\sim}{x},t) \, \varepsilon \, C^p(\underset{\sim}{x},t)$ and $\underset{\sim}{f}(\underset{\sim}{x}) \, \varepsilon \, C^p(\underset{\sim}{x})$ then (1.1) has a unique solution $\underset{\sim}{u}(\underset{\sim}{x},t)$ and it can, with its first p derivatives, be estimated, in the $L_2$ norm, by $A_j$, f and its derivatives up to order p.

Therefore the continuous dependence of $\underset{\sim}{u}(\underset{\sim}{x},t)$ on the initial data is assured.

### 1.1.2 The Initial Boundary Value Problem

We now consider the initial boundary value problem form of (1.1) where some or all of the spatial variables $x_i$ lie in bounded domains. The appropriate Cauchy problem is always assumed to be well-posed.

Specifically we consider the constant coefficient quarter-plane problem defined in $D = [0,\infty) \times \mathbb{R}^{m-1}$. In other words $x_1 \geqslant 0$ and $x_i \, \varepsilon \, \mathbb{R} \quad i = 2,3,\ldots,m$. Since the problem is hyperbolic the matrix $A_1$ can be diagonalised to

$$\Lambda = \begin{bmatrix} -\Lambda_I & 0 \\ & \\ 0 & \Lambda_{II} \end{bmatrix}$$

where $\Lambda_I = \text{diag}(\lambda_1,\ldots,\lambda_\ell)$ and $\Lambda_{II} = \text{diag}(\lambda_{\ell+1},\ldots,\lambda_n)$ are positive definite matrices and each diagonal element of $\Lambda$ is an eigenvalue of $A_1$. The boundary conditions along $x_1 = 0$ are arbitrary and each set of conditions will define a different solution $\underset{\sim}{u}(\underset{\sim}{x},t)$. From the theory of characteristics we have that the number of solution waves entering the domain at the hyperplane $x_1 = 0$ is given by the number of negative eigenvalues of $A_1$. The boundary conditions define the ingoing solutions in terms of the outgoing solutions and they take the form

$$\underset{\sim}{u}^{I}(0,x_-,t) = S\underset{\sim}{u}^{II}(0,x_-,t) + \underset{\sim}{g}(x_-,t), \quad (\underset{\sim}{x},t) \in \delta D \times [0,T] \qquad (1.5)$$

with

$$\underset{\sim}{u}^{I} = (u_1,\ldots,u_\ell)^T, \quad \underset{\sim}{u}^{II} = (u_{\ell+1},\ldots,u_n)^T$$

$$x_- = (x_2,x_3,\ldots,x_m)^T \quad \text{and} \quad \delta D = (x_1 = 0) \times \mathbb{R}^{m-1} .$$

S is an $\ell \times (n-\ell)$ matrix and can be thought of as the operator defining the reflection of the incident waves against $x_1 = 0$. It is assumed that no discontinuities are introduced at the intersection of the initial and boundary domains.

The strictly hyperbolic problem given above by (1.5),(1.1) is well-posed if for some constant K, possibly dependent on T, we have the estimate

$$\int_0^T \|\underset{\sim}{u}(0,x_-,t)\|_{\delta D}^2 \, dt + \int_0^T \|\underset{\sim}{u}(\underset{\sim}{x},t)\|_D^2 \, dt$$
$$\leqslant K\left\{\int_0^T \|\underset{\sim}{g}(x_-,t)\|_{\delta D}^2 \, dt + \int_0^T \|\underset{\sim}{F}(\underset{\sim}{x},t)\|_D^2 \, dt\right\} . \qquad (1.6)$$

For the one-dimensional case well-posedness is assured if all the boundary expressions are disjoint from the reflected waves (Hersh [1963]). This means that if the boundary conditions at $x = 0$ are written as $B\underset{\sim}{u} = \underset{\sim}{g}$ then well-posedness follows from the disjointness of the null space of B and the negative eigenspace of $A_1$. For higher dimensional problems we can introduce Fourier transforms with respect to $x_-$ and a Laplace transform with respect to time (dual variables $\xi_- \varepsilon \mathbb{R}^{m-1}$ and $s \varepsilon \mathbb{C}$ respectively). If we denote by $\hat{a}$ the transform of a then from (1.1) and (1.5)

$$s\hat{\underset{\sim}{u}} = A_1 \frac{d\hat{\underset{\sim}{u}}}{dx_1} + i \sum_{j=2}^{m} A_j \xi_j \hat{\underset{\sim}{u}} + \hat{\underset{\sim}{F}}, \quad x_1 > 0 \tag{1.7}$$

$$\hat{\underset{\sim}{u}}^I = S\hat{\underset{\sim}{u}}^{II} + \hat{\underset{\sim}{g}} \quad , \quad x_1 = 0. \tag{1.8}$$

If we denote by $\phi \varepsilon L_2(\mathbb{R}^+)$ an eigenfunction corresponding to the eigenvalue s then $\phi$ and s necessarily satisfy the linear differential eigenvalue problem

$$s\phi = A_1 \frac{d\phi}{dx_1} + i \sum_{j=2}^{m} A_j \xi_j \phi \ , \quad x_1 \geqslant 0 \tag{1.9}$$

$$\phi^I = S\phi^{II} \quad , \tag{1.10}$$

with $\phi^I = (\phi_1, \ldots, \phi_\ell)^T$ and $\phi^{II} = (\phi_{\ell+1}, \ldots, \phi_n)^T$.
Note that we are able to drop the inhomogeneous terms $\hat{g}$ and $\hat{F}$ (Kreiss [1970]).

Hersh [1963] quotes an example to indicate that no well-posedness estimate (1.6) can hold if there exists an eigensolution $\phi$ associated with an $s : \text{Re}(s) > 0$. He also shows that for $\text{Re}(s) > 0$ there are exactly $\ell$ solutions $\phi \varepsilon L_2(\mathbb{R}^+)$ . Therefore the general solution of (1.9) can be written

$$\phi = \sum_{j=1}^{\ell} \beta_j \phi_j \quad . \tag{1.11}$$

Clearly then, if there exists a non-trivial set of constants $\{\beta_j\}_{j=1}^{\ell}$ so that (1.11) satisfies (1.10), the problem is not well-posed. This may be expressed algebraically by constructing the order $\ell$ linear system

$$M(s)\underset{\sim}{\beta} = \underset{\sim}{0}$$

by inserting (1.11) into (1.10). Well-posedness then requires that det $M(s) \neq 0$ for any s such that $\text{Re}(s) > 0$. If $\phi$ and s satisfy (1.9) and (1.10) with $\text{Re}(s) = 0$ (we assume $\lim_{\varepsilon \to 0^+} \phi(s+\varepsilon) \in L_2(\mathbb{R}^+)$) then s is said to be a generalised eigenvalue. We then have the main result

Theorem 1.2 (Kreiss [1970]). The problem defined by (1.1) and (1.5) is well-posed if and only if the eigenvalue problem (1.9) and (1.10) has no eigenvalues or generalised eigenvalues with $\text{Re}(s) \geqslant 0$.

The above result was extended to the case of variable complex strictly hyperbolic operators by Ralston [1971]. The variable coefficient problem can be considered by examining all of the linearized constant coefficient problems at every point on the boundary (Coughran [1980]). Non-strictly hyperbolic problems have been considered by Majda and Osher [1975].

Although problems in general will not be of the semi-infinite form considered above, at any particular boundary we can locally consider the domain of definition to be so (Kreiss [1979]), (Oliger and Sundström [1978]).

An older technique for determining well-posedness is the energy method (Richtmyer and Morton [1967]). Although this method requires skill in its use to provide sufficient conditions it can also handle non-linear problems. Elvius and Sundström [1973] proved well-posedness for the shallow-water equations and Oliger and Sundström [1978] handled some complicated systems in fluid dynamics. Griffin and Anderson [1977] studied the influence of boundary conditions on one-dimensional fluid flows through a nozzle.

## 1.2 The Discrete Problem.

### 1.2.1 The Cauchy Problem.

We first approximate the strictly hyperbolic one-dimensional problem

$$\underset{\sim}{u}_t = A\underset{\sim}{u}_x + \underset{\sim}{F}(x,t) \quad , \quad (x,t) \in \mathbb{R} \times [0,\infty) \qquad (1.12)$$

$$\underset{\sim}{u}(x,0) = \underset{\sim}{f}(x) \qquad , \quad x \in \mathbb{R} \qquad (1.13)$$

with A a constant, square matrix of order n and

$$\underset{\sim}{u}(x,t) = (u_1(x,t),\dots,u_n(x,t))^T .$$

We cover the domain with a rectangular mesh or grid where the spatial grid spacing is denoted by $\Delta x > 0$ and the time step by $\Delta t > 0$, assuming a constant mesh ratio of $\lambda = \Delta t/\Delta x$. At each knot we define $\underset{\sim}{U}_j^n$ to be the finite difference approximation of $\underset{\sim}{u}(j\Delta x, n\Delta t)$, $j \in \mathbf{Z}$, $n \in \mathbf{Z}^+$. In practice the $\underset{\sim}{U}_j^n$ will be defined by the multi-step method

$$Q_{-1}\underset{\sim}{U}_j^{n+1} = \sum_{\sigma=0}^{r} Q_\sigma \underset{\sim}{U}_j^{n-\sigma} + \Delta t\, \underset{\sim}{F}_j^n, \quad \forall j \in \mathbb{Z};\ n = r, r+1, \ldots \qquad (1.14)$$

where $Q_\sigma = \sum_{\alpha=-p_\ell}^{p_r} A_{\alpha\sigma}(\Delta x, \Delta t) E^\alpha$; $E\underset{\sim}{U}_j = \underset{\sim}{U}_{j+1}$ are difference operators and $Q_{-1}$ is non-singular.

To initialise the method the following (r+1) levels of data are required

$$\underset{\sim}{U}_j^\sigma = \underset{\sim}{f}_j^\sigma \quad \sigma = 0,1,\ldots,r; \quad \forall j \in \mathbb{Z} \qquad (1.15)$$

The values in (1.15) will usually be supplied by Taylor expansions about $\underset{\sim}{U}^o$, or by using other finite difference schemes that involve fewer time levels. We assume that the initial data will not introduce any loss of accuracy or instability.

As in the previous section we consider first the Cauchy problem approximation (1.14). It will become clear that many of the stability results and methods for the finite difference approximations are synonymous with those of well-posedness in the differential problem.

We initially make clear definitions of some of the properties of a difference approximation.

Definition 1.2. The difference scheme given by (1.14),(1.15) is accurate of order $(q_1,q_2)$ for the particular solution u(x,t) if there is a function c(t) and constants $c_\sigma \geq 0$, bounded on every finite interval $[0,T]$, such that for all sufficiently small $\Delta x$, $\Delta t$

$$\|Q_{-1}\underset{\sim}{U}_j^{n+1} - \sum_{\sigma=0}^{r} Q_\sigma \underset{\sim}{U}_j^{n-\sigma}\|_h \leq \Delta t\, C(t)\, \{(\Delta x)^{q_1} + (\Delta t)^{q_2}\}, \qquad (1.16)$$

$$\|f^\sigma - \underset{\sim}{U}^\sigma\|_h \leq c_\sigma\{(\Delta x)^{q_1} + (\Delta t)^{q_2}\}, \quad \sigma = 0,1,\ldots,r.$$

In many cases we say that the order of accuracy is given by min $(q_1,q_2)$.

Whilst it may seem desirable to maximise the order of accuracy possible, in practice methods more than sixth order accurate become computationally too expensive (Kreiss and Oliger [1972]).

Using (1.16) we say

Definition 1.3. The difference scheme (1.14), (1.15) is said to be consistent if it is accurate at least of order (1,1). We assume all methods considered here are consistent.

An essential property of any approximation to a well-posed problem is that it be stable, that is, it does not grow unboundedly in time. The importance of stability is illustrated by

Theorem 1.3 (Lax's Equivalence Theorem). Given a well-posed linear differential problem and a finite difference approximation to it that is consistent then the approximation is convergent if and only if it is stable.

Accordingly we make the following

Definition 1.4. A finite difference method is stable for the Cauchy problem if there are constants $K > 0$ and $\alpha \in \mathbb{R}^+$ such that the following estimate holds

$$\|\underset{\sim}{U}^n\|_h \leq Ke^{\alpha t} \sum_{\sigma=0}^{r} \|\underset{\sim}{U}^\sigma\|_h . \tag{1.17}$$

Therefore the numerical solution at any point in time must remain bounded in terms of the initial data.

As before we require implementable algebraic conditions for the estimate(1.17) to hold. To do this we first Laplace transform the time variable using the dual variable s, which is akin to seeking a solution of the form

$$\underset{\sim}{U}_j^n = z^{\frac{t}{\Delta t}} \hat{\underset{\sim}{U}}_j, \ \forall j \in \mathbb{Z}, \quad z = e^{s\Delta t} \in C, \quad \hat{\underset{\sim}{U}} \in L_2(\Delta x), \ t = n\Delta t. \quad (1.18)$$

Inserting this in the homogeneous form of (1.14) yields the resolvent equation

$$zQ_{-1}\hat{\underset{\sim}{U}}_j - \sum_{\sigma=0}^{r} z^{-\sigma} Q_\sigma \underset{\sim}{U}_j = 0, \ \forall j \in \mathbb{Z}. \qquad (1.19)$$

Equation (1.19) has the associated eigenvalue problem

$$\left[zQ_{-1} - \sum_{\sigma=0}^{r} z^{-\sigma} Q_\sigma\right] \underset{\sim}{\phi} = 0, \quad \underset{\sim}{\phi} \in L_2(\Delta x). \qquad (1.20)$$

Again it is clear that there cannot be any eigenvalues z with $|z| > 1$ associated with an eigenfunction $\underset{\sim}{\phi} \in L_2(\Delta x)$. If this were so then a solution (1.18) would grow unboundedly in time. We have

<u>Lemma 1.1</u> (Gustafsson et al [1972]). For $|z| > 1$ the number of linearly independent solutions $\underset{\sim}{\phi}$ in (1.20) is $np_\ell$. We shall see later that this lemma is important when dealing with the stability of the initial and boundary value problem .

If we seek a solution in (1.19) of the form $\hat{\underset{\sim}{U}}_j = \kappa^j \underset{\sim}{\phi} \quad \kappa \in \mathbb{C}$ (1.21) we obtain the characteristic polynomial

$$\det\left[zQ_{-1}\kappa^j - \sum_{\sigma=0}^{r} z^{-\sigma} Q_\sigma \kappa^j\right] = 0. \qquad (1.22)$$

Polynomial equation (1.22) is the requirement for the existence of non-trivial $\underset{\sim}{\phi}$ in (1.21) and will invariably decouple into separate characteristic equations involving z and $\kappa$. From this we get

Theorem 13a (Von-Neumann Theorem). The difference scheme (1.14), (1.15) has no exponentially growing solutions if all the solutions in (1.22) satisfy $|z| \leq 1$.

Although this is a necessary condition it happens also to be sufficient in the case of two-level schemes where the amplification matrix $Q_{-1}^{-1}Q_o$ is normal. As all multi-level schemes can be written as two-level schemes this result is important however $Q_{-1}^{-1}Q_o$ may not always be normal (Richtmyer and Morton [1967]).

Difference schemes can be classified by

Definition 1.5. The approximation (1.14) is said to be dissipative of order $2\alpha$ if there exists, for the solutions z of (1.22), an estimate

$$|z| \leq 1 - \delta|\omega\Delta x|^{2\alpha} \; ; \quad \forall z, \quad |\omega\Delta x| < \pi$$

for some positive constant $\delta$ and whole number $\alpha$.

The solution of a dissipative scheme will tend to zero as $t \to \infty$ whereas for a non-dissipative method each mode in the initial data will be transported, through time, with its amplitude conserved. Clearly the choice of difference method, in any particular case will depend on apriori knowledge of the unknown analytic solution. A dissipative method for example, will not provide an acceptable approximation to a long term steady state solution. A non-dissipative method on the other hand would be expected to do so, however rounding errors may cause some of the eigenvalues z to exceed one in modulus. Thus, at best, non-dissipative methods are marginally stable in practice.

The physical interpretation of dissipation is that the high frequency Fourier components associated with large values of $\omega\Delta x$ are damped and the difference solutions will compose mainly of smooth low frequency components. See Plot 2, (page 158) where the dissipative Lax Wendroff method is used in conjunction with an unstable boundary approximation. Although serious inaccuracies are occurring at the boundaries the interior method prevents this transfer of energy throughout the domain. We have

Theorem 1.4. If A in (1.12) is hermitian and the approximation is also dissipative of order $2\alpha$ and accurate of order $2\alpha-1$, (or order $2(\alpha-1)$ for strictly hyperbolic problems) then the difference method is stable.

Theorem 1.4 still applies if the operator A is uniformly bounded and Lipschitz continuous in x.

Dissipation is important when solving non-linear problems in that non-dissipative methods invariably suffer from non-linear instabilities. Fornberg [1973] displayed this behaviour with the Leap-Frog scheme. Garp [1977] gives similar examples with the method-of-lines and Galerkin approximations. Majda and Osher [1978] proposed a method of introducing non-linear dissipation to the standard Lax Wendroff method to avoid this type of instability. As an alternative, it is known that solving the governing equations in their conservation form is less likely to produce non-linear instabilities (Garp [1977]), however they do not guarantee its non-occurrence. The addition of dissipation to a non-dissipative method is popular (Kreiss and Oliger [1973]). The ability to control the degree of dissipation being regarded as more favourable than using a dissipative method initially.

Alternative sufficient stability definitions for the Cauchy problem are given by Richtmyer and Morton [1967] and Kreiss and Oliger [1973].

1.2.2 The Initial Boundary Value Problem.

Consider the mixed initial boundary value quarter-plane problem

$$\begin{aligned} \underset{\sim}{u}_t &= A\underset{\sim}{u}_x + \underset{\sim}{F}(x,t) \quad , \quad (x,t) \in \mathbb{R}^+ \times \mathbb{R}^+ , \\ \underset{\sim}{u}(x,0) &= \underset{\sim}{f}(x) \quad , \quad x \in \mathbb{R}^+ \\ \underset{\sim}{u}^I(0,t) &= S\underset{\sim}{u}^{II}(0,t) + \underset{\sim}{g}(t), \quad t \in \mathbb{R}^+ \end{aligned} \qquad (1.23)$$

where

$$A = \begin{bmatrix} -\Lambda^I & 0 \\ 0 & \Lambda^{II} \end{bmatrix} ,$$

$$\begin{aligned} \Lambda^I &= \mathrm{diag}(\lambda_1,\ldots,\lambda_\ell) > 0, \\ \Lambda^{II} &= \mathrm{diag}(\lambda_{\ell+1},\ldots,\lambda_n) > 0, \\ \underset{\sim}{u}^I(x,t) &= [u_1(x,t),\ldots,u_\ell(x,t)]^T, \\ \underset{\sim}{u}^{II}(x,t) &= [u_{\ell+1}(x,t),\ldots,u_n(x,t)]^T, \end{aligned}$$

$S$ an $\ell \times (n-\ell)$ real reflection matrix.

We solve using finite differences by choosing a scheme that is stable, albeit possibly under certain grid size restrictions, for the Cauchy problem. The introduction of boundary approximations may cause the overall difference scheme to be unconditionally stable, conditionally stable or unstable. We assume homogeneous initial data. We make the

Definition 1.6. A boundary condition is an expression along the boundary required by the differential problem to ensure well-posedness.

A boundary approximation is an additional equation required only by the numerical method.

To solve (1.23) we use the multi-step method (1.14), (1.15) with the boundary information

$$\underset{\sim}{U}_j^{n+1} = \sum_{\sigma=-1}^{r} S_\sigma^{(j)} \underset{\sim}{U}_1^{n-\sigma} + \underset{\sim}{g}_j^n, \quad j = 1-p_\ell, \ldots, 0 \tag{1.24}$$

with $$S_\sigma^{(j)} = \sum_{\mu=0}^{\beta} C_{\mu\sigma}^{(j)} E^\mu . \tag{1.25}$$

Notice that (1.24) incorporates $\ell$ boundary conditions and $(n-\ell)$ boundary approximations on $x = 0$ as well as any approximations required at interior points.

A heuristic stability analysis of (1.14), (1.15) and (1.24) was given by Trapp and Ramshaw [1976] by taking the necessary stability restriction to be the minimum Von-Neumann restriction for each separate difference expression. That is, treating the interior method and each boundary approximation as though they were applied to a separate Cauchy problem. More rigourously we shall consider three methods of analysing the effect of any boundary approximation on stability. Namely

(a) The Normal Mode Analysis of Kreiss;

and if a method of lines approximation is used, then

(b) The approach of Strikwerda, and lastly the

(c) The Matrix Eigenvalue method.

### 1.2.2a Normal Mode Analysis

In 1968 Kreiss gave a stability theory for dissipative approximations to the quarter-plane problem by extending the work of Godunov and Ryabenkii [1963]. This theory was then extended by Gustafsson et al [1972] to either fully dissipative or fully non-dissipative approximations to the twin boundary problem in one dimension. The definition of stability that they based their results upon allows, in certain cases, exponentially growing solutions when $\Delta x$ and $\Delta t$ are not sufficiently small. By modifying the ideas of Gustafsson et al [1972] into a definition of so-called P-stability Beam, Warming and Yee [1981] were able to exclude this possibility.

As in the well-posedness discussion the idea behind the theory is to consider all solutions belonging to $L_2(\Delta x)$ but which are unbounded in time. If any of these unacceptable solutions satisfies the boundary approximations then the overall approximation will be unstable.

Therefore we Laplace transform in time which, as stated before, is akin to seeking a solution of the form $\underset{\sim}{U}_j^n = z^n\hat{\underset{\sim}{U}}_j$ in equations (1.14) and (1.24). Accordingly we obtain the resolvent equations

$$zQ_{-1}\hat{\underset{\sim}{U}}_j - \sum_{\sigma=0}^{r} z^{-\sigma}Q_\sigma\hat{\underset{\sim}{U}}_j = \hat{\underset{\sim}{F}}_j, \qquad \forall j \geqslant 1 \tag{1.26}$$

and
$$z\hat{\underset{\sim}{U}}_j - \sum_{\sigma=-1}^{r} z^{-\sigma}S_\sigma^{(j)}\hat{\underset{\sim}{U}}_j = \hat{\underset{\sim}{g}}_j, \qquad j = 1-p_\ell,\ldots,0. \tag{1.27}$$

Assuming a solution $\hat{\underset{\sim}{U}}_j = \kappa^j\underset{\sim}{\phi} \in L_2(\Delta x)$, $\kappa \in \mathbb{C}$ of the difference equations (1.26) and (1.27) we find that $\underset{\sim}{U}_j^n = z^n\kappa^j\underset{\sim}{\phi}$ is a solution of (1.14) and (1.24) if the following equalities

$$\left[zQ_{-1}\kappa^{j} - \sum_{\sigma=0}^{r} z^{-\sigma}Q_{\sigma}\kappa^{j}\right]\underset{\sim}{\phi} = \hat{\underset{\sim}{F}}_{j}, \qquad \forall j \geqslant 1 \tag{1.28}$$

$$\left[z\kappa^{j} - \sum_{\sigma=0}^{r} z^{-\sigma}S_{\sigma}^{(j)}\kappa\right]\underset{\sim}{\phi} = \hat{\underset{\sim}{g}}_{j}, \quad j = 1-p_{\ell},\ldots,0, \tag{1.29}$$

are satisfied.

We are interested in solutions of (1.28) for which $|z| > 1$, $\hat{\underset{\sim}{U}}_{j} \in L_{2}(\Delta x)$ and the general solution will then be of the form

$$\underset{\sim}{U}_{j}^{n} = z^{n} \sum_{i} P_{i}(j) \left[\kappa_{i}\right]^{j} \underset{\sim}{\phi}_{i}, \quad \forall i : |\kappa_{i}| < 1 \tag{1.30}$$

where the polynomial coefficient will reflect the multiplicity of any $\kappa_i$. Note that stability of the associated Cauchy problem ensures that there are no solutions $\hat{\underset{\sim}{U}}_{j}$ such that $|\kappa_i| = 1$ for $|z| > 1$. From Lemma 1.1 solution (1.30) becomes

$$\underset{\sim}{U}_{j}^{n} = z^{n} \sum_{i=1}^{np_{\ell}} P_{i}(j)\left[\kappa_{i}\right]^{j}\underset{\sim}{\phi}_{i} \tag{1.31}$$

We have now determined all the unacceptable solutions of the interior difference scheme. This leads to

Theorem 1.5 (Godunov-Ryabenkii). If solution (1.31) for $|z| > 1$ while $|\kappa_i| < 1$ satisfies equations (1.28) and (1.29) for some non-trivial vector $\underset{\sim}{\phi}_{i}$ then the approximation (1.14) and (1.24) is unstable for the quarter-plane problem. Such a z is said to be an eigenvalue.

This result is only a necessary stability condition as we have yet to consider the possibility of z lying on the unit circle. This extension was performed by Kreiss [1968] and Gustafsson et al [1972]. The latter authors base their results on

Definition 1.7. The difference approximation to the initial-boundary value problem (1.23) is stable if there are constants $K > 0$ and $\alpha_o \in \mathbb{R}^+$ such that for all $\Delta t > 0$ and $\alpha > \alpha_o$ there exists the estimate

$$\frac{(\alpha-\alpha_o)}{(1+\alpha\Delta t)}\left[\sum_{j=1-p_\ell}^{0}\|e^{-\alpha t}\underset{\sim}{U}_j\|_t^2+\frac{(\alpha-\alpha_o)}{(1-\alpha\Delta t)}\|e^{-\alpha t}\underset{\sim}{U}\|_{x,t}^2\right] \tag{1.32}$$

$$\leqslant K\left[\frac{(\alpha-\alpha_o)}{(1+\alpha\Delta t)}\sum_{j=1-p_\ell}^{0}\|e^{-\alpha(t+\Delta t)}\underset{\sim}{g}_j\|_t^2+\|e^{-\alpha(t+\Delta t)}\underset{\sim}{F}\|_{x,t}^2\right].$$

In general (1.32) allows for the existence of exponentially growing solutions unless $\alpha_o = 0$. The form of (1.32) allowed the authors to construct a straightforward algebraic stability condition for all $|z| \geqslant 1$.

Definition 1.8. If $|z| = 1$ and there is an associated value of $\kappa_i$ such that $|\kappa_i| = 1$ and $z = (1+\delta)e^{i\theta}$ implies that $|\kappa_i| < 1$ for some $\delta$, $0 < \delta << 1$, and $(z,\kappa_i)$ satisfy (1.28) and (1.29) then $z$ is said to be a generalised eigenvalue. We then have

Theorem 1.6 (Gustafsson et al [1972]). The difference scheme (1.14) and (1.24) is a stable approximation of the quarter plane problem (1.23) if and only if (1.28) and (1.29) have no eigenvalues or generalised eigenvalues for $|z| \geqslant 1$.

The familiar algebraic condition is obtained, by inserting (1.21) in the homogeneous form of (1.28) to form the linear system

$$M(z,\kappa)\underset{\sim}{\phi} = 0. \tag{1.33}$$

Theorem 1.6 is satisfied if no eigenvalue or generalised eigenvalue satisfies det $M(z,\kappa) = 0$ for $|z| \geq 1$.

This linear system (1.33) is simple in practice to construct however it often proves difficult to check Theorem 1.6. Possible procedures for doing so are given in Chapter 2.

Kreiss [1968] proved that it was sufficient, when the interior approximation is dissipative, to consider the possible existence of a generalised eigenvalue at $z = 1$. However for non-dissipative methods the entire unit circle has to be examined (see for example the SILF method at $z = i$ as described in Chapter 4 of this thesis).

Gustafsson et al [1972] prove that the two-boundary problem is stable if each separate quarter-plane problem, obtained by extending the other boundary to $\pm \infty$ as applicable, is stable. In practice it is necessary to check both quarter-plane problems (Jamieson and Sloan [1983] and Chapter 3).

For scalar equations it is usually possible to verify the conditions of Theorem 1.6 easily. This is particularly relevant if, in the constant coefficient case of (1.14), the equations and boundary approximations can be written in characteristic form. Gottlieb et al [1978] proved that the stability of each separate scalar characteristic problem was sufficient to ensure stability of the complete problem. This, however, is only relevant in the constant coefficient case, although (linearised) characteristics can play an important role in the construction of boundary approximations.

Trefethen [1983] interpreted instability, as defined above, as amounting to the spontaneous radiation of energy from the boundary into the interior. In an unstable situation a rounding error may be transported back into the computational domain subject to an unbounded amplification factor. The possibility of a generalised eigenvalue is shown to be equivalent to a positive group velocity directed from the boundary into the domain. Using this analogy Trefethen [1983] was able to verify the instability of many well-known difference approximations. This interpretation holds for both dissipative and non-dissipative schemes. It is not, however, able to guarantee stability in the sense of Definition 1.7.

When applied to boundary approximations, used in conjunction with some implicit schemes, Theorem 1.6 has been shown to allow exponentially growing solutions when $\Delta x$ and $\Delta t$ are not sufficiently small (Beam, Warming and Yee [1981]). These authors introduce the more restrictive

Definition 1.9. The difference scheme for an initial-boundary value problem is said to be P-stable if

(a) it is stable for the Cauchy problem,

(b) it is stable for the left and right quarter-plane problems (in the sense of definition 1.7), and

(c) all the eigenvalues of the characteristic equation, for a finite number of spatial mesh intervals, lie in or on the unit circle.

For definition 1.9 (c) the characteristic equation is constructed from the characteristic polynomial equation (1.22) and from all boundary approximations, both at the left and right spatial boundaries. The analysis involved in including both boundaries is only straightforward for scalar equations and may be very complicated for systems.

The possible existence of growing solutions has been considered by Gustafsson [1981]. He examined the relation between stability, as defined in Definition 1.7, and the eigenvalues of the operator Q in the initial and boundary-value difference scheme $\underset{\sim}{U}^{n+1} = Q\underset{\sim}{U}^{n}$. If all the eigenvalues of Q are strictly inside the unit circle the problem is said to have only decreasing modes. He examined the various conclusions which may be formed concerning decreasing modes from an analysis of the two quarter-plane problems.

In order to introduce Gustafsson's result we return to the resolvent equation (1.26) and we write the homogeneous form as the one-step equation in space

$$\underset{\sim}{U}^{n}_{j+1} = G\underset{\sim}{U}^{n}_{j} \; . \tag{1.34}$$

It is possible to transform G to the block diagonal form

$$\mathrm{diag}(L_1,L_2,N_1,N_2) \tag{1.35}$$

(Gustafsson et al [1972]). The blocks in (1.35) may satisfy

$$\begin{aligned} L_1^*L_1 &\leq (1-\delta)I \quad , \\ L_2^*L_2 &\leq (1-\delta)(|z|-1)I, \\ N_1^*N_1 &\geq (1+\delta)I \quad , \\ N_2^*N_2 &\geq (1+\delta)(|z|-1)I, \end{aligned} \tag{1.36}$$

for some $\delta > 0$ and identity operator I, in some neighbourhood of $z_o$ on the unit circle. We then have

Theorem 1.7 (Gustafsson [1981]). If both quarter-plane problems are stable (in the sense of (1.32)) and inequalities (1.36) are satisfied then if either $L_2$ or $N_2$ are empty then there are only decreasing solutions to the twin boundary problem for a sufficiently fine mesh.

The theory of Gustafsson et al [1972] can be applied to variable coefficient problems by invoking the results of Lax and Nirenberg [1966]. For sufficiently smooth variable coefficient problems then stability of the associated left and right quarter plane problems linearized about the boundaries implies stability of the original problem (Kreiss and Oliger [1973]).

The extension of this theory to multi-dimensional problems is not clear although for the Leap-Frog method in two dimensions Abarbanel and Gottlieb [1979] considered the stability of various boundary approximations. For the strictly hyperbolic case Coughran [1980] and Michelson [1981] considered a dissipative multi-dimensional interior difference scheme.

1.2.2b The method of Strikwerda.

Consider the quarter-plane constant coefficient problem

$$\begin{aligned} \underset{\sim}{u}_t &= A\underset{\sim}{u}_x + \underset{\sim}{F}(x,t) , && (x,t) \in \mathbb{R}^+ \times \mathbb{R}^+ \\ \underset{\sim}{u}(x,0) &= \underset{\sim}{f}(x) , && x \in \mathbb{R}^+ \\ \underset{\sim}{u}(0,t) &= \underset{\sim}{g}(t) , && t \in \mathbb{R}^+ . \end{aligned} \tag{1.37}$$

If $\underset{\sim}{u}_j(t)$, regarded as a semi-discrete approximation to $\underset{\sim}{u}(j\Delta x,t)$, satisfies the differential difference method

$$\frac{d}{dt}\underset{\sim}{U}_j = AD\underset{\sim}{U}_j + \underset{\sim}{F}(j\Delta x,t), \quad \forall j \geqslant r \tag{1.38}$$

$$\frac{d}{dt}\underset{\sim}{U}_j = AD_j\underset{\sim}{U}_j + \underset{\sim}{F}(j\Delta x,t), \quad j = 0,1,\ldots,r \tag{1.39}$$

and

$$\frac{d}{dt}\underset{\sim}{U}_o = AD_o\underset{\sim}{U}_o + \underset{\sim}{h}(t) \text{ with } \underset{\sim}{U}_j = 0, \ \forall j \tag{1.40}$$

it is said to be a method-of-lines approximation to $u(j\Delta x,t)$. In the above

$$DU_j = \sum_{\sigma=-r}^{r} d_\sigma U_{j+\sigma} \quad \text{and} \quad D_j \underset{\sim}{U}_j = \sum_{\sigma=-j}^{2r-j} d_{j\sigma} U_{j+\sigma},$$

where $d_\sigma$ and $d_{j\sigma}$ are constants. Equation (1.40) incorporates the boundary conditions and boundary approximations in differential form. We have

Definition 1.10. If there is a function $\underset{\sim}{u}(x,s)$ such that

(a) $s\underset{\sim}{u} = A\underset{\sim}{u}_x$ on $x \geq 0$ ,

(b) $\text{Re } s \geq 0$ ,

(c) for $\text{Re } s > 0$, $\underset{\sim}{u}(x,s)$ is bounded as $x \to \infty$,

(d) for $\text{Re } s = 0$, $\underset{\sim}{u}(x,s) = \lim_{\varepsilon \to 0^+} \underset{\sim}{u}(x,s+\varepsilon)$ satisfying (a) and (c),

(e) $\underset{\sim}{u}(0,s) = 0$

then $\underset{\sim}{u}(x,s)$ is said to be an eigensolution of (1.37).

The definition corresponds to the construction of eigenvalues and generalised eigenvalues of the previous well-posedness analysis. Parts (c) and (d) define what may be regarded as unacceptable solutions and (e) verifies that the boundary conditions are satisfied by such solutions. This leads to

Theorem 1.8 (Strikwerda [1980]). The quarter-plane initial-boundary value problem is well-posed if and only if it has no eigensolutions.

An analogous approach allows us to determine the conditions necessary for stability of the method-of-lines approximation. We seek to construct all solutions which are bounded in space but allow unbounded temporal growth. If any such solution satisfies the boundary approximations then instability follows.

Laplace transformation in time yields the resolvent equations

$$s\hat{\underset{\sim}{U}}_j = AD\hat{\underset{\sim}{U}}_j \tag{1.41}$$

and if any root $\hat{\underset{\sim}{U}}_j \in L_2(\Delta x)$ satisfies, for Re $s > 0$, the transformed boundary equacions then the approximation is unstable. This result can be compared with Theorem 1.5. We also consider the existence of generalised eigensolutions (Strikwerda [1980]).

<u>Definition 1.11.</u> A generalised eigensolution, $\underset{\sim}{U}_j(i\alpha)$ with $\alpha \in \mathbb{R}$, is a solution given by $\underset{\sim}{U}_j(i\alpha) = \lim_{\varepsilon\to 0^+} \underset{\sim}{U}_j(\varepsilon+i\alpha)$ where $\underset{\sim}{U}_j(\varepsilon+i\alpha) \in L_2(\Delta x)$ and satisfies (1.41).

<u>Theorem 1.9</u> (Strikwerda [1980]). The method of lines approximation to the quarter-plane initial-boundary value problem is stable, if and only if it has no eigensolutions or generalised eigensolutions.

A determinant condition, synonymous with that of Theorem 1.6, can be constructed as an algebraic condition for stability. The characteristic equations and the secular equation for the boundary approximations will involve fewer terms but they are not necessarily any easier to solve.

The application of the method-of-lines approximation requires the use of a stable ordinary differential equation solver or a stable time-stepping rule (Vichnevetsky and Bowles [1982]). The previous results concerning sufficiently smooth variable coefficient problems and twin boundary problems still apply.

The final method of analysing stability is given as

### 1.2.2c The Eigenvalue Method on the Method-Of-Lines.

This approach has been used frequently (see, for example, Gunzburger [1977], Yee [1981], Gustafsson [1981]) and studies the method-of-lines approximation (1.38)-(1.40). We use the following Definition 1.12 (Cronin [1980]). For the system

$$\frac{d}{dt}\underset{\sim}{v} = f(t,\underset{\sim}{v}) \qquad (1.42)$$

in some real domain that includes positive time, let $\underset{\sim}{v}(t)$ denote the solution $\forall t > \tau > 0$. Then $v(t)$ is stable if there exists a $t_o > \tau$ and positive real constant b and $\varepsilon$ such that, if $\underset{\sim}{v}(t_o) = \underset{\sim}{v}^o$ and where $\underset{\sim}{v}(t) \equiv \underset{\sim}{v}(t,t_o,\underset{\sim}{v}^o)$, the following hold

(a) If $|\underset{\sim}{v}'-\underset{\sim}{v}^o| < b$ then $\underset{\sim}{v}(t,t_o,\underset{\sim}{v}')$ is a solution of (1.42) and is defined for all $t \geqslant t_o$,

(b) There exists $\delta = \delta(\varepsilon,\underset{\sim}{f},t_o,\underset{\sim}{v}^o) > 0$ such that $\delta \leqslant b$ and if $|\underset{\sim}{v}'-\underset{\sim}{v}^o| < \delta$ then $|\underset{\sim}{v}(t,t_o,\underset{\sim}{v}') - \underset{\sim}{v}(t,t_o,\underset{\sim}{v}^o)| < \varepsilon$ for all $t \geqslant t_o$.

If (a) and (b) hold and if

(c) there exists $\alpha = \alpha(f,t_o\underset{\sim}{v}^o)$ such that $\alpha < b$ and $|\underset{\sim}{v}'-\underset{\sim}{v}^o| < \alpha$ then

$$\lim_{t\to\infty} |\underset{\sim}{v}(t,t_o,\underset{\sim}{v}') - \underset{\sim}{v}(t,t_o,\underset{\sim}{v}^o)| = 0$$

and $\underset{\sim}{v}(t)$ is asymptotically stable.

Definition 1.12 states that given an initial solution $\underset{\sim}{v}(t_o)$ the general solution $\underset{\sim}{v}(t)$ will be stable if, once in a neighbourhood of $\underset{\sim}{v}(t_o)$, it stays in that neighbourhood and will be well-defined. The solution $\underset{\sim}{v}(t)$ will be asymptotically stable if it is convergent in that neighbourhood. In comparison with the theory of Gustafsson et al

[1972] stability, in terms of Definition 1.12, is equivalent to large-time or asymptotic stability as implied by Definition 1.7. The above definition therefore places a less stringent restriction upon the numerical approximation.

In general the differential-difference system will be of the form

$$\frac{d}{dt}\underset{\sim}{v} = A\underset{\sim}{v} \tag{1.43}$$

where, for constant coefficient hyperbolic problems, A is a constant square matrix. If $\lambda$ denotes an eigenvalue of A then we have

Theorem 1.10 (Cronin [1980]). Let $\underset{\sim}{v}$ denote the solution of the homogeneous system (1.43). Then $\underset{\sim}{v}$ is stable if, for all $\lambda$, $\mathrm{Re}(\lambda) \leq 0$ and all imaginary eigenvalues are distinct. Furthermore, $\underset{\sim}{v}$ is asymptotically stable if all the eigenvalues have real parts that are negative.

Theorem 1.10 is readily applicable although the evaluation of the eigenvalues, for large systems, can be computationally expensive.

# CHAPTER 2

# METHODS FOR VERIFYING STABILITY RESULTS

In the previous chapter algebraic stability conditions were derived that involved the roots of a system of non-linear complex polynomial equations. Whilst it may be possible to solve this system analytically (see Chapter 3), in general, numerical solution techniques are required. Problems of this type have been examined by Allgower and Georg [1980], Ypma [1982] and Hirsch and Smale [1979]. Allgower and Georg reviewed various simplicial and continuation techniques. Ypma considered the use of Newton methods where the Jacobian is replaced by a difference approximation. This modified Newton method is then incorporated into a continuation algorithm. Hirsch and Smale proposed several algorithms based on Newton's method, including a "sure-fire" method. This method is, however, slow to converge and very costly computationally. Practical algorithms are also presented which are easier to apply though not guaranteed to converge. In this chapter we shall consider the use of polynomial resultants and various forms of continuation. Both methods work well in practice and are used to analyse all the boundary approximations considered in Chapters 3,4 and 5. Since all applications in those chapters involve systems of the form $F_\nu : \mathbb{C}^n \to \mathbb{C}$ where

$$F_\nu(z_1,z_2,\ldots,z_n) = 0 \qquad \nu = 1,2,\ldots,n \tag{2.1}$$

we will assume that the numerical methods are to be applied to systems of the form (2.1). In the later chapters $n = 3$ and the variables $z_1,z_2$ and $z_3$ may be written as $\kappa_1,\kappa_3$ and $z$. $\kappa_1$ and $\kappa_3$ are the only roots of the homogeneous form of (1.28) which we need to consider. The equations $F_1(\kappa_1,\kappa_3,z) = 0$ and $F_2(\kappa_1,\kappa_3,z) = 0$ are the characteristic equations associated with the interior method of integration and $F_3(\kappa_1,\kappa_3,z) = 0$ is the determinant condition obtained from (1.33). The expressions $F_i$, $i = 1,2,3$, may be written as polynomials in $z$ with coefficients which are polynomials in $\kappa_1$ and $\kappa_3$.

## 2.1 Polynomial Resultant Approach

This method was described by Collins [1971] and involves the evaluation of a succession of matrix determinants.

Definition 2.1. Let $A(x) = \sum_{i=0}^{m} a_i x^i$ and $B(x) = \sum_{i=0}^{n} b_i x^i$ be polynomials in x of degree m and n respectively. The Sylvester matrix of A and B is the square matrix of order (m+n)

$$S = \left|\begin{array}{ccccccc}
a_m & a_{m-1} & \cdots & a_o & 0 & \cdots & 0\\
0 & a_m & \cdots & a_1 & a_o & & \vdots\\
\vdots & 0 & \ddots & \vdots & \vdots & \ddots & 0\\
0\cdots 0 & & & a_m & a_{m-1} & \cdots & a_o\\
b_n & b_{n-1} & \cdots & \cdots & b_o & 0\cdots & 0\\
0 & b_n & & & b_1 & b_o 0 \cdots & 0\\
\vdots & 0 & \ddots & & \vdots & \ddots & \vdots\\
\vdots & \vdots & & \ddots & \vdots & & 0\\
0 & \cdots & \cdots & 0 & b_n & b_{n-1} \cdots & b_o
\end{array}\right|$$

The resultant of A and B, res(A,B), is the determinant of S.

Clearly res(A,B) involves only the coefficients of A and B. The method suggested by Collins [1971] to solve (2.1) is to construct polynomials

$$P_1(\kappa_1,\kappa_3) = \text{res}(F_1,F_3) = 0; \quad P_2(\kappa_1,\kappa_3) = \text{res}(F_2,F_3) = 0$$

thus eliminating z. If now $P_1$ and $P_2$ are written as polynomials in $\kappa_3$ with coefficients which are polynomials in $\kappa_1$ then $P_3(\kappa_1) = \text{res}(P_1,P_2) = 0$ defines one polynomial, albeit of high degree, in one complex unknown, $\kappa_1$. The solution of $P_3 = 0$ using a standard library routine and subsequent solution for $\kappa_3$ and z in $P_1 = 0$ and $F_3 = 0$, respectively, will yield a collection of triples $(\kappa_1,\kappa_3,z)$. The evaluation of $F_1,F_2,F_3$ for each triple will identify the true roots of (2.1). We illustrate by the example:

Find the roots $x,y,z \in \mathbb{C}$ satisfying

$$F_1 : xz - y = 0 \quad , \qquad (2.2)$$

$$F_2 : xy - 1 = 0 \quad , \qquad (2.3)$$

$$F_3 : x + y + z = 0 \quad . \qquad (2.4)$$

We then have $P_1(x,z) = \text{res}(F_1,F_2) = 1-x^2z$ and $P_2(x,z) = \text{res}(F_1,F_3) = -(x+z+xz)$. Combining $P_1$ and $P_2$ we have

$$P_3(x) = \text{res}(P_1,P_2) = x^3+x+1 = 0$$

Using the roots of $P_3 = 0$ we obtain three values of y from (2.3) and then three values of z from (2.4). Only those triples (x,y,z) which satisfy (2.2) are roots of the system.

This example is trivial. In practice the construction of $P_i$ $i = 1,2,3$, requires the construction of polynomial multiplication and determinant routines. In all applications in this thesis the highest degree of $P_3 = 0$ was eight. A more detailed example is given in Appendix I.

## 2.2 Continuation

The basis of continuation is to split the polynomial system (2.1) into a simple part and a remainder. The simple part is constructed in such a way that the problem it defines on its own may be solved analytically and then numerical solutions are obtained for a sequence of problems by successively adding increments of the aforementioned remainder to the simple part. When the complete remainder has been added a solution of (2.1) is then obtained. To be precise, system (2.1) is imbedded into a family of equations given by $H_\nu : \mathbb{C}^{n+1} \to \mathbb{C}$ where

$$H_\nu(\underset{\sim}{z},t) = \eta_\nu(\underset{\sim}{z}) + t\varepsilon_\nu(\underset{\sim}{z}) \qquad \nu = 1,2,\ldots,n \qquad (2.6)$$

with $\underset{\sim}{z} = (z_1,\ldots,z_n)^T$, $t \in \mathbb{C}$, such that $H_\nu(\underset{\sim}{z},0) = \eta_\nu(\underset{\sim}{z}) = 0$ has simple roots that are readily found and $H_\nu(\underset{\sim}{z},1) = F_\nu(\underset{\sim}{z})$. In order that no roots be lost it is necessary that the degree of $\eta_\nu$ should not be less than that of $F_\nu$.

There are various ways of continuing t in (2.6). Firstly, in "Standard Continuation", t may be regarded as an independent variable and moved from 0 to 1 along a path in the complex plane. This approach has been used by Wasserstrom [1973] and more extensively by

Drexler [1978]. Secondly we can introduce a real parameter s so that $\underset{\sim}{z} = \underset{\sim}{z}(s)$ and $t = t(s)$. The path in $\mathbb{C}^n \times \mathbb{R}$ defined by $(\underset{\sim}{z}(s), t(s))$ is the so-called homotopy path. If system (2.6) is differentiated with respect to s an initial-value problem is formed which may be integrated from the initial roots of $\eta_\nu(\underset{\sim}{z}) = 0$ to the required solution triple at $t = 1$. This has been the more recent approach examined by Garcia and Zangwill [1980] in connection with polynomial systems and by Watson [1979, 1980a,b,c, 1981] in a variety of situations. Finally we shall consider a hybrid method of the above techniques.

Faced with the problem of solving system (2.1), extensive numerical experiments were performed in order that some insight might be obtained into the relative merits and demerits of the different approaches. For the purpose of comparison the following two problems were used,

(a) Drexler's First Problem:

$$x^2-1+xy = 0; \quad y^2-1-yz = 0; \quad z^2-1+2xz = 0 \qquad x,y,z \in \mathbb{C} \quad (2.7)$$

which has six real and two complex roots, and

(b) Drexler's Second Problem:

$$x^2+xy+y^2+1 = 0; \quad y^3+xy(x+y)+1 = 0 \qquad x,y \in \mathbb{C} \qquad (2.8)$$

where there are two complex roots only.

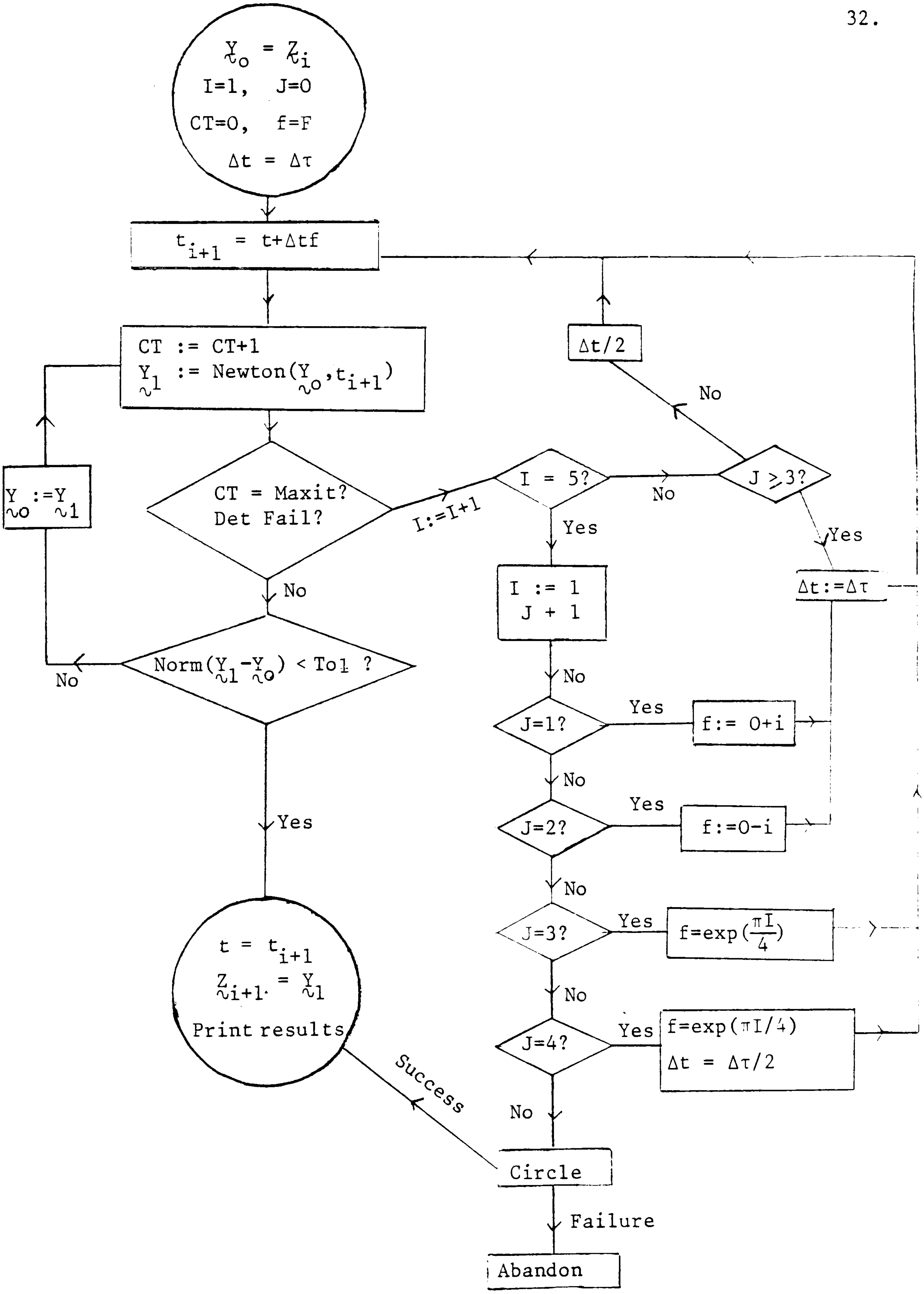
Experience gained on the above problems led to changes in existing algorithms. Extensions were made to the analysis which permitted simplifications to be introduced.

### 2.2.1 Standard Continuation

The basic algorithm for this approach has been described by Wasserstrom [1973] and developed and implemented by Drexler [1978] and is of a predicror-corrector type.

Differentiation of $H(\underset{\sim}{z},t) = 0$ with respect to t, together with the roots of $H(\underset{\sim}{z},0) = 0$, forms an initial-value problem, which when integrated from $t = 0$ to $t = 1$, yields the desired roots. To perform this integration the interval $[0,1]$ is subdivided and in each subdivision the integral is evaluated numerically. In $[t_i,t_{i+1}]$ typically we have the solution $\underset{\sim}{z}(t_i)$ to $H(\underset{\sim}{z},t_i) = 0$. The crudest predictor delivers $\underset{\sim}{z}(t_i)$ as an initial estimate of $\underset{\sim}{z}(t)$ at $t = t_{i+1}$. Newton's method is then used to find an accurate value of $\underset{\sim}{z}(t_{i+1})$. The union of all such subintervals forms the solution curve W associated with the given initial vector. To justify the use of this method Drexler proves that the solution curve W-{1} may always be chosen such that the solutions along the curve are unique, bounded, continuous and many times differentiable. He does so by showing that there are only a finite number of points where these properties do not hold. The point $t = 1$ is excluded since, if this solution path does not correspond to a zero of $H(\underset{\sim}{z},1) = 0$, then a pole will occur. Also the presence of multiple roots will be indicated by two solution curves coalescing at $t = 1$. As stated above, it is the choice of W in practice which dictates the effectiveness of the technique.

A program was written which used standard continuation to produce the curve W for a system such as (2.1). The program which emerged from the numerical experiments is represented by the following flow chart. This chart describes a method for finding $\underset{\sim}{z}_{i+1}$ in the interval $[t_i,t_{i+1}]$.


$\underset{\sim}{Y}_o = \underset{\sim}{Z}_i$
I=1, J=0
CT=0, f=F
$\Delta t = \Delta\tau$
$t_{i+1} = t+\Delta tf$
CT := CT+1
$\underset{\sim}{Y}_1 := \text{Newton}(\underset{\sim}{Y}_o, t_{i+1})$
$\underset{\sim}{Y}_o := \underset{\sim}{Y}_1$
CT = Maxit?
Det Fail?
I:=I+1
No
$\text{Norm}(\underset{\sim}{Y}_1 - \underset{\sim}{Y}_o) < \text{Tol}$ ?
No
Yes
$t = t_{i+1}$
$\underset{\sim}{Z}_{i+1} = \underset{\sim}{Y}_1$
Print results
Success
$\Delta t/2$
No
I = 5?
No
J ≥ 3?
Yes
Yes
$\Delta t := \Delta\tau$
I := 1
J + 1
No
J=1?
Yes
f:= 0+i
No
J=2?
Yes
f:=0-i
No
J=3?
Yes
$f=\exp(\frac{\pi I}{4})$
No
J=4?
Yes
$f=\exp(\pi I/4)$
$\Delta t = \Delta\tau/2$
No
Circle
Failure
Abandon

The Standard Continuation Flow Chart

To explain the previous flow chart, assume a point of success $\underset{\sim}{z}_i$ has been reached at $t = t_i$ along W. The next point $t_{i+1}$ is estimated by advancing t in a straight line to t = 1 a distance $\Delta t$, performed by the function $F(\mu_1,\mu_2)$ with $\mu_1$ and $\mu_2$ representing movements in the real and imaginary directions respectively. Using $\underset{\sim}{z}_i$ as the predicted solution at $t = t_{i+1}$ we correct it by using repeated application of Newton's Method. If, in the iteration process, a point of difficulty is encountered then several approaches to adjust $t_{i+1}$ are used. Moving back to the previous solution at $t = t_i$ we adopt the following successive procedure:

(a) divide $\Delta t$ by 2 a maximum number of four times,

(b) move t in the positive imaginary direction a distance $\Delta t$ and reapply (a) if necessary,

(c) move t in the negative imaginary direction a distance $\Delta t$ and reapply (a) if necessary,

(d) move t towards t = 1 distance $\Delta t$ and rotate by $\frac{\pi}{4}$ and rotate further in multiples of $\frac{\pi}{2}$ if necessary after reapplying (a).

If a solution point $\underset{\sim}{z}_{i+1}$ cannot be found then go back to t = 0 and try moving t around the unit circle until t = 1 is reached. A change of direction for W is sought when any of the following occur (Drexler [1978]):

(1) Newton's method fails to converge in a given number of iterations,

(2) the determinant of the Jacobian matrix overflows,

(3) the determinant of the Jacobian matrix increases five-fold from step (k) to (k+1) of the iteration, decreases by half and then oscillates.

As should be apparent, the above listing of decision criteria and responses is far from exhaustive. Clearly the major problem with this method is that it is not possible to even remotely guarantee that a correct solution curve for each set of initial data can be found numerically.

When used on both test problems all the roots were found. However, while the W-curves remained bounded, the decision criteria could not guarantee uniqueness. For this reason, and due to the complexity of the algorithm, another approach was sought.

## 2.2.2 Parameterized Continuation

In the previous section the continuation parameter t was moved along the solution curve by a set of decision criteria that responded to the behaviour of the variables as the curve moved through $\mathbb{C}^{n+1}$. The philosophy behind parameterized continuation is to let t move freely with the other variables by introducing a real parameter s upon which all the variables are assumed to be dependent. As a result it will be seen that many of the problems encountered previously do not occur (Watson [1980]). The parameter t is now constrained to be real and in this case the homotopy path, or solution curve, is a path in $\mathbb{C}^n \times \mathbb{R}$ or $\mathbb{R}^{2n+1}$. The success of the method relies on the assumption that the solution curve is regular and so the difficulties in dealing with bifurcation points are avoided. In the solution of many systems of the form (2.1) this assumption has never been violated.

As before an initial value problem is constructed in which s is increased from its arbitrary initial value that corresponds to $t = 0$ until the unknown zero of $\underset{\sim}{F}(\underset{\sim}{z}) = \underset{\sim}{0}$ is obtained at $t = 1$. It should be noted that the initial and final t points may be interchanged. In other words we may decrease s from t = 1 until t = 0. This of course requires a change in the definition of $\underset{\sim}{H}$.

Consider $W(\underset{\sim}{z}(s),t(s)) : \mathbb{R}^{2n} \times [0,1] \to \mathbb{R}^{2n}$ to be the regular solution curve for a given set of initial data $\underset{\sim}{z}^{(j)}$ with $t(0) = 0$, $\underset{\sim}{z}(0) = \underset{\sim}{z}^{(j)}$. Notice that we refer to any $\alpha \in \mathbb{C}$ as being equivalent to a pair of real numbers. The first defines the real part of $\alpha$ the latter the imaginary part. Thus the solution curve W represents the solution of the system

$$\frac{d}{ds}\, \underset{\sim}{H}(\underset{\sim}{z}(s),t(s)) = \underset{\sim}{H}'(W(s))\dot{W}(s) = 0, \qquad (2.9a)$$

$$t(0) = 0 \; ; \quad \underset{\sim}{z}(0) = \underset{\sim}{z}^{(j)}, \qquad (2.9b)$$

where $\underset{\sim}{H}'$ denotes the Fréchet derivative of H with respect to W. Thus the integration of (2.9) with respect to s until t = 1 is reached will yield the required zero point. In many cases (Watson and Wang [1982], and Chow, Mallet-Parret and Yorke [1978]) the parameter s can be scaled to correspond to the arc length of W by normalising the vector $[\dot{z},\dot{t}]^T$.

This approach was first proposed by Kellog, Li and Yorke [1976] for proving Brouwer's theorem and was adapted by Chow et al [1978] into a root or fixed point finding algorithm. It has subsequently been widely implemented in fixed-point, zero-finding, two point boundary value, optimization and continuum mechanics problems by Watson and co-workers. The method seems particularly suited to fixed point problems where Chow et al [1978] have proved that, under modest assumptions of regularity, it has probability one of being successful. Though for the location of zeros the authors advise caution. Specifically for root-finding in systems of polynomials Garcia and Zangwill [1980] have constructed a system of ordinary differential equations that correspnd to a solution of (2.9a). Garcia and Zangwill do not scale the parameter s.

Due to the nature of system (2.1) the next method adopted will be that of Garcia and Zangwill with s being the arc length of W. Notice the domain of definition of W. In the majority of the papers referenced, the $z_\nu$ $\nu = 1,\ldots,n$ are treated as real variables and even when $\underset{\sim}{z} \in \mathbb{C}^n$ (2.9a) is separated into real and imaginary components as stated earlier (Garcia and Zangwill [1980]). Below we show that z may be retained as a vector in $\mathbb{C}^n$. A transformation may be introduced which enables the system for $\dot{W}(s)$ to be treated in complex form and a proof is given in this chapter that t remains real and strictly monotonic in s. This change to complex form gives an improvement on the Garcia and Zangwill algorithm.

Garcia and Zangwill Algorithm:

To aid the description the following is of use

<u>Definition 2.1.</u> A polynomial f(z), $\underset{\sim}{z} \in C^n$, is said to be in maximal degree form if $\exists$ integers $\nu$ and $\mu$ such that

$$f(\underset{\sim}{z}) = z_\nu^\mu + \sum_k a_k z_1^{r_k^1} z_2^{r_k^2} \cdots z_n^{r_k^n}$$

where $a_k \in \mathbb{R}$, $r_k^i \in \mathbb{Z}^+$ and where $\mu > \max_k \sum_{\sigma=1}^{n} r_k^\sigma$. In other words the polynomial $f(\underset{\sim}{z})$ has a dominant term $z_\nu^\mu$.

In all applications in connection with this thesis n is at most three and so, for notational simplicity, all the following results will be proven for n = 3. Generalisation to arbitrary n is immediate. Consequently the variable $z_4$ will be identified with the variable t. Garcia and Zangwill differ from Watson, and others, in that their homotopies have the initial point at t = 1.

Consider the original system(2.1)imbedded in the homotopies $H_\nu : \mathbb{C}^3 \times [0,1] \to \mathbb{C}$ where

$$H_\nu = H_\nu(z_1(s),z_2(s),z_3(s),z_4(s)). \qquad \nu = 1,2,3 \qquad (2.10)$$

It is at this point that Garcia and Zangwill would have formed the real system $G_\nu : \mathbb{R}^6 \times [0,1] \to \mathbb{R} \quad \nu = 1,2,\ldots,6$ by defining

$$G_{2\nu-1}(\underset{\sim}{\omega}) = \mathrm{Re}(H_\nu), G_{2\nu}(\underset{\sim}{\omega}) = \mathrm{Im}(H_\nu) \qquad \nu = 1,2,3$$

and $\omega_{2\nu-1} = \mathrm{Re}(z_\nu), \quad \omega_{2\nu} = \mathrm{Im}(z_\nu) \qquad \nu = 1,2,3 \; ; \; \omega_7 = t.$
However, for ease of explanation we shall remain in complex form for the present and return to the system $\underset{\sim}{G}$ to show compatibility of results later.

Clearly, if (2.10) holds for all s then $\frac{d}{ds} H_\nu = 0$ yielding

$$\sum_{\sigma=1}^{4} \frac{\partial}{\partial z_\sigma} H_\nu \cdot \frac{d}{ds} z_\sigma = 0 \qquad \nu = 1,2,3. \qquad (2.11)$$

Introducing the notation $\dot{\alpha} = \frac{d}{ds} \alpha$, $\quad H_{\nu\sigma} = \frac{\partial}{\partial z_\sigma} H_\nu$ system (2.11) may be written

$$H_{11}\dot{z}_1 + H_{12}\dot{z}_2 + H_{13}\dot{z}_3 + H_{14}\dot{z}_4 = 0, \tag{2.12a}$$

$$H_{21}\dot{z}_1 + H_{22}\dot{z}_2 + H_{23}\dot{z}_3 + H_{24}\dot{z}_4 = 0, \tag{2.12b}$$

$$H_{31}\dot{z}_1 + H_{32}\dot{z}_2 + H_{33}\dot{z}_3 + H_{34}\dot{z}_4 = 0, \tag{2.12c}$$

or

$$\underset{\sim}{H}'\underset{\sim}{\dot{z}} = 0.$$

A solution of (2.12) is readily shown to be

$$\dot{z}_1 = \begin{vmatrix} H_{12} & H_{13} & H_{14} \\ H_{22} & H_{23} & H_{24} \\ H_{32} & H_{33} & H_{34} \end{vmatrix}, \quad \dot{z}_2 = (-1) \begin{vmatrix} H_{11} & H_{13} & H_{14} \\ H_{21} & H_{23} & H_{24} \\ H_{31} & H_{33} & H_{34} \end{vmatrix},$$

$$\dot{z}_3 = \begin{vmatrix} H_{11} & H_{12} & H_{14} \\ H_{12} & H_{22} & H_{24} \\ H_{31} & H_{32} & H_{34} \end{vmatrix}, \quad \dot{z}_4 = (-1) \begin{vmatrix} H_{11} & H_{12} & H_{13} \\ H_{21} & H_{22} & H_{23} \\ H_{31} & H_{32} & H_{33} \end{vmatrix}. \tag{2.13}$$

If the solutions (2.13) are inserted into the left hand side of (2.12a) then the result is equivalent to evaluating the determinant of a matrix having linearly dependent rows. The solutions (2.13) may be written as

$$\dot{z}_\nu = (-1)^{\nu+1} \det H'_\nu \; , \; \nu = 1,2,3,4 \tag{2.14}$$

where $H'_\nu$ is the 3 × 3 matrix obtained by removing column $\nu$ from $\underset{\sim}{H}'$. Since rank($\underset{\sim}{H}'$) = 3, from the regurality assumption, $\underset{\sim}{\dot{z}}$ is an element of the null space N($\underset{\sim}{H}'$) having rank unity. Therefore the $\underset{\sim}{\dot{z}}$ in (2.13) are determined to within a constant.

Reconsidering the real representation of Garcia and Zangwill, their determinant in the expression of $\dot{t} = \dot{\omega}_7$ would be

$$\begin{vmatrix} G_{11} & G_{12} \cdots G_{16} \\ G_{21} & G_{22} \cdots G_{26} \\ \vdots & \vdots \\ G_{61} & G_{62} \cdots G_{66} \end{vmatrix} \quad \text{where } G_{\nu\sigma} = \frac{\partial}{\partial\omega_\sigma} G_\nu \qquad (2.15)$$

Using the Cauchy-Riemann equations and some matrix manipulation, determinant (2.15) may be reduced to the block determinant

$$- \begin{vmatrix} B & C \\ C & -B \end{vmatrix} \qquad (2.16)$$

where B and C are the real matrices

$$\begin{bmatrix} G_{12} & G_{14} & G_{16} \\ G_{32} & G_{34} & G_{36} \\ G_{52} & G_{54} & G_{56} \end{bmatrix} \text{ and } \begin{bmatrix} G_{11} & G_{13} & G_{15} \\ G_{31} & G_{33} & G_{35} \\ G_{51} & G_{53} & G_{55} \end{bmatrix}, \text{ respectively.}$$

We may also write (2.16) as

$$\begin{aligned} \dot{t} = \dot{\omega}_7 &= - \begin{vmatrix} iB & C \\ C & iB \end{vmatrix} \\ &= - \begin{vmatrix} 0 & C+iB \\ C-iB & iB \end{vmatrix} \\ &= \det(C+iB)\det(C-iB) \end{aligned} \qquad (2.17)$$

It is readily shown that C-iB = $H_4'$, and if the determinant of this complex matrix is denoted by J, then (2.17) takes the form

$$\dot{t} = \bar{J} \det H_4' = |J|^2 .$$

Note that this solution for $\dot{t}$ is obtained from (2.14) if (2.14) is multiplied by $-\bar{J}$. If the real solutions of Garcia and Zangwill are combined in the form $\dot{\omega}_{2\nu-1} + i\dot{\omega}_{2\nu}$ $(\nu = 1,2,3)$ it is readily shown that, using analysis similar to that above, the solutions of (2.14) are obtained in the modified form'

$$\dot{z}_\nu = (-1)^\nu \, \bar{J} \det H_\nu' \quad , \nu = 1,2,3,4 \ . \qquad (2.18)$$

This is a more convenient form than that used by Garcia and Zangwill and the dimensions of all the determinants have been halved. All computations were performed using the complex arithmetic facility of ALGOL 68, so the change to complex form causes no inconvenience. Garcia and Zangwill [1980] proved that, using their determinants, the variable t behaved monotonically between t = 1 and t = 0. This result follows immediately from the complex formulation (2.18).

Note that from the form of (2.18) the occurrence $\dot{t} = \dot{z}_4 = 0$ necessarily implies that $\dot{z}_\nu = 0$ $(\nu = 1,2,3)$. This being the case, then all (3 × 3) minors of the matrix $\underset{\sim}{H}'$ are zero, thus contradicting the regularity assumption that rank $(\underset{\sim}{H}') = 3$. It then follows that at all points on the path $\dot{t}$ the initial sign adopted at t = 0 or 1 is maintained. If the initial point is t = 0 then solution (2.18) is used and $\dot{t} > 0 \ \forall s \in \mathbb{R}$ and if the initial point is t = 1 then

(2.18) is negated and $\dot{t} < 0 \ \forall s \in \mathbb{R}$.

The above result therefore ensures that as we move along the path W the value of t cannot oscillate or turn back. Hence two possibilities remain. The curve can either remain bounded at the final point or diverge to infinity as the final t-point is approached. The latter event indicates that there is no solution vector corresponding to the given initial data. Garcia and Zangwill prove that the number of spurious paths will reduce if more of the $F_\nu(\underset{\sim}{z})$ $(\nu=1,\ldots,n)$ are written in maximal degree form. Since it is not always possible to obtain such a form spurious paths will always exist. The problem is then to construct the homotopies to yield all the desired roots.

Practical experience suggests that the $H_\nu$ should be written in maximal degree form thus greatly increasing the required number of initial vectors. This form proved completely successful for the test problems whereas a reduced degree form, equivalent to the degree of $F_\nu$, missed several vector zeros.

When integrating the differential system it is desirable to avoid standard packages which do not allow the progress to be monitored. Successful curve following involves the implementation of certain test criteria after each integration step from s to $s+\Delta s$. Such criteria will be discussed in detail for the next continuation method.

A third approach to the continuation problem combines features of the two previous methods. Again an independent parameter s is introduced upon which all the other variables are assumed to be dependent. An initial value problem is constructed in a manner similar to that of Garcia and Zangwill and is solved by a predictor-corrector technique. The predictor used is an r-th order Runge-Kutta approximation. Numerical experiments were performed on Drexler's test problems with r = 0,2 and 4. The value r = 4 produces an accurate prediction and reduces the CPU time required by the corrector. For most of the experiments accurate results were obtained most efficiently using r = 0 and r = 2. This accords with the experience of Rheinboldt [1981]. Newton's method was used to correct the initial approximation. To minimise the computational effort complex arithmetic was used as far as possible throughout the algorithm.

The above predictor-corrector method was discussed, among others, by Allgower and Georg [1980]. Garcia and Li [1979] describe the algorithm in detail for a general predictor and also prove that the method will converge to a solution point. Li and Yorke [1979] discuss the practical implementation with reference to the decision criteria required to ensure uniqueness of the solution curve. The authors demonstrate the performance of the algorithm by solving Wilkinson's classic unstable polynomial.

### 2.2.3 A Composite Continuation Method

As before we construct the system of homotopies

$$H(\underset{\sim}{z}(s),\ t(s)) = 0 \qquad (2.19)$$

where $H : \mathbb{C}^n \times [0,1] \to C^n$ and define $W \in \mathbb{C}^n \times [0,1]$ to represent the solution curve to (2.19) for any $s \in \mathbb{R}$. The path W is generated as the solution of

$$H'(W)\dot{W} = 0 \tag{2.20}$$

where $\dot{W} = \frac{dW}{ds}$ and H' is the matrix of complex partial derivatives. The parameter s corresponds to arc length on the homotopy path so the normalising condition $||\dot{W}||^2 = 1$ must be used. Here $||\dot{W}||^2$ denotes the squared Euclidean norm $|\dot{\underset{\sim}{z}}|^2 + \dot{t}^2$. Assume we have reached a point $W_{(i)}$ along the solution curve. To advance to $W_{(i+1)}$ we proceed as follows: calculate the unit tangent vector $\underset{\sim}{b}_{(i)} \in C^n \times \mathbb{R}$ such that

$$H'(W_{(i)})\underset{\sim}{b}_{(i)} = 0 \quad \text{and} \quad ||\underset{\sim}{b}_{(i)}|| = 1$$

by forming the determinants (2.18) and normalising.
Advancing along this vector a distance $\Delta s$ from $W_{(i)}$ we obtain the initial approximation, $\underset{\sim}{x}^o = W_{(i)} + \Delta s\, \underset{\sim}{b}_{(i)}$ to $W_{(i+1)}$. We now use $\underset{\sim}{x}^o$ to initiate a Newton iteration for solving equation (2.19). Since (2.19) represents n complex equations in n+1 unknowns we require an additional constraint. This is provided by insisting that all subsequent approximations lie in the hyperplane $\theta$, through $\underset{\sim}{x}^o$, perpendicular to $\underset{\sim}{b}_{(i)}$.

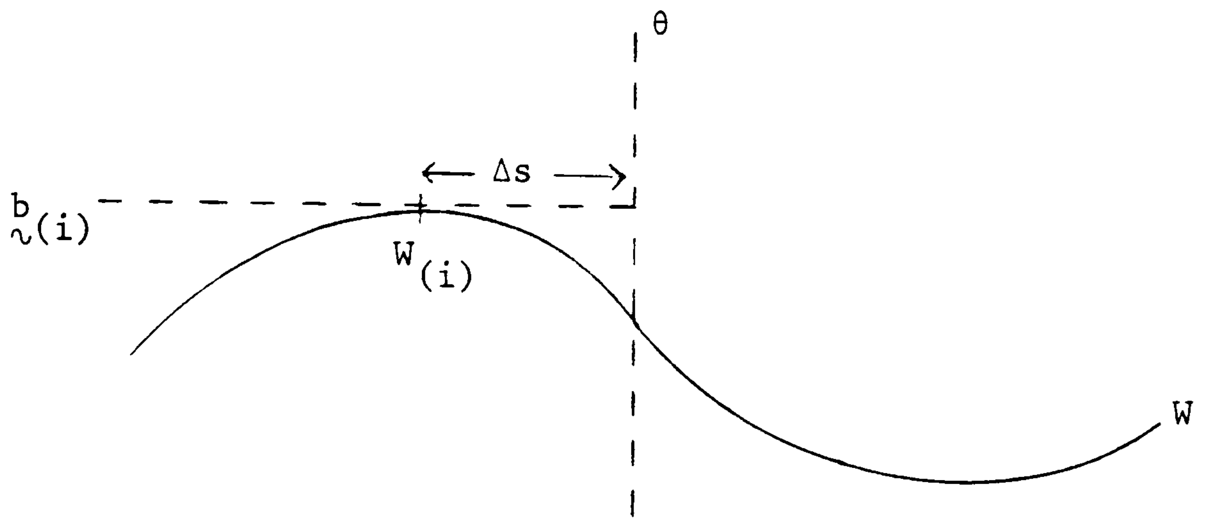


Figure 2.1 : Advancement of $W_{(i)}$ to $W_{(i+1)}$ in continuation method.

This is illustrated in Figure 2.1.

The Newtonian system is therefore

$$H'(\underset{\sim}{x}^k)(\underset{\sim}{x}^k-\underset{\sim}{x}^{k+1}) = H(\underset{\sim}{x}^k) \;, \qquad (2.21a)$$

$$k = 0,1,\dots \; .$$

$$\underset{\sim}{b}^T_{(i)}(\underset{\sim}{x}^k-\underset{\sim}{x}^{k+1}) = 0 \qquad , \quad . \qquad (2.21b)$$

The above explanation has invoked the predictor r = 0. However $\underset{\sim}{b}_{(i)}$ could have been improved to an r-th order Runge-Kutta approximation by evaluating the required number of unit vectors. Clearly the more accurate the prediction the more efficient the Newton iteration becomes. The cost, however, in obtaining even $\underset{\sim}{b}_{(i)}$ (r = 0) is high.

To solve system (2.21) without recourse to a formulation in real arithmetic define $\xi^k_\nu$ $(\nu = 1,2,\dots,n)$ to be the complex increment $(x^{k+1}_\nu-x^k_\nu)$ and $\xi^k_{n+1}$ the real increment $(x^{k+1}_{n+1}-x^k_{n+1})$. Insertion of $\underset{\sim}{\xi}^k$ in (2.21a) results in

$$H'(\underset{\sim}{x}^k)\underset{\sim}{\xi}^k = - \underset{\sim}{H}(\underset{\sim}{x}^k) . \qquad (2.22)$$

If we now remove from H' the last column representing $\frac{\partial H}{\partial t}$, denoted by $\underset{\sim}{H}^{n+1}$, and if we also let $\underset{\sim}{\gamma}^k = \{\xi_i^k : i = 1,\dots,n\}$ then (2.22) becomes, with $H'_{-(n+1)} \equiv H'$ less the column $\underset{\sim}{H}^{n+1}$,

$$H'_{-(n+1)}(\underset{\sim}{x}^k)\underset{\sim}{\gamma}^k = -\underset{\sim}{H}(\underset{\sim}{x}^k) - \xi_{n+1}^k \cdot \underset{\sim}{H}^{n+1}$$

or,

$$\underset{\sim}{\gamma}^k = \underset{\sim}{M} + \underset{\sim}{N}\xi_{n+1}^k \qquad (2.23)$$

where $$\underset{\sim}{M} = -\left[H'_{-(n+1)}(\underset{\sim}{x}^k)\right]^{-1} \underset{\sim}{H}(\underset{\sim}{x}^k),$$

and $$\underset{\sim}{N} = -\left[H'_{-(n+1)}(\underset{\sim}{x}^k)\right]^{-1} \underset{\sim}{H}^{n+1} .$$

Consider equation (2.21b). Transformation of the unit tangent vector $\underset{\sim}{b}_{(i)}$ into a real (2n+1) array $\underset{\sim}{R}^i$ and solving for the real variable $\xi_{n+1}^k$ yields

$$\xi_{n+1}^k = -\frac{\sum_{\sigma=1}^{n} R_{2\sigma-1}^i \cdot \mathrm{Re}(M_\sigma) + \sum_{\sigma=1}^{n} R_{2\sigma}^i \cdot \mathrm{Im}(M_\sigma)}{\left[\sum_{\sigma=1}^{n} R_{2\sigma-1}^i \cdot \mathrm{Re}(N_\sigma) + \sum_{\sigma=1}^{n} R_{2\sigma}^i \cdot \mathrm{Im}(N_\sigma) + R_{2n+1}^i\right]} .$$

Therefore solving for $\xi_\nu^k$ ($\nu = 1,\dots,n$) in (2.23) we obtain the new approximation $\underset{\sim}{x}^{k+1} = \underset{\sim}{x}^k + \underset{\sim}{\xi}^k$.

For small systems (2.1) it is advantageous to calculate the inverse in $\underset{\sim}{M}$ and $\underset{\sim}{N}$ analytically making use of the location of zeros in H'. The use, otherwise, of standard routines greatly increases the cost of each iteration. In practice this increase is of order three. Thus for large systems this method may not be the most appropriate.

The crucial factor in the implementation lies in the choice of $\Delta s$. The homotopy path is guaranteed to be monotonic in t but may turn sharply in any of the other variables. Such erratic behaviour may require $\Delta s$ to be significantly reduced in order that the hyperplane $\theta$ still intersects the intended solution curve and not into the domain of convergence of another path. All experiments undertaken used an upper bound on $\Delta s$ of one. To help follow the path around a bend Li and Yorke [1979] suggest an "angle test". This involves checking the angle between any two tangent vectors at successive points on W. If at $W_{(i+1)}$ the tangent vectors at $W_{(i+1)}$ and $W_{(i)}$ are greater than $\phi^o$ apart then the point $W_{(i+1)}$ should be rejected, $\Delta s$ halved and the process started again at $W_{(i)}$. Li and Yorke suggest $\phi = 18$. If, during the Newton iteration, the inequality

$$||\underset{\sim}{H}(\underset{\sim}{x}^k)|| \leqslant 0.01 \, ||H(\underset{\sim}{x}^{k-1})||$$

for some norm $||.||$, is violated then Li and Yorke consider that $\underset{\sim}{x}^k$ lies outside the domain of quadratic convergence of Newton's method for the path W. Therefore the iteration is stopped, $\Delta s$ halved and the process repeated at the previous point on W. If, in some interval, the curve is smoothly behaved, indicated by three successively equal values of $\Delta s$, then $\Delta s$ may be multiplied by an arbitrary factor to increase the efficiency of the method. In practice a factor 5 was used.

The above decision criteria are identical to those mentioned in the discussion of Garcia and Zangwill approach.

Recently Zirilli (1982) has extended the parameterised approach by constructing the initial-value problem as a second order ordinary differential equation. A continuation algorithm similar to that of this section has been developed by Ypma [1980], where the Jacobian is approximated by a difference method. The author considers the use of highly accurate predictors and adaptive steplength techniques.

In practice the parameterized methods fail when the increment $\Delta s$ falls below the machine accuracy and so a possible solution point is lost. For this reason maximal degree homotopies are more successful since a large number of initial vectors are created. All the zeros of the test problems were found. Examining the different approaches of standard and parameterized continuation we consider the parameterized approach to be superior. This conclusion is reinforced through its widespread application by Watson. The parameterized method has a real and monotonic parameter t and avoids the problems of a zero Jacobian that would cause Drexler's method to fail. The decision criteria are also much more problem dependent for the standard approach. There seems little to choose between the second and third methods though for small systems (2.1) the composite method is more efficient.

The resultant method is guaranteed to work and will find all the zeros simultaneously though the method is algebraically complex. Both the composite continuation and the resultant methods are used successfully to find stability triples $(\kappa_1,\kappa_3,z)$ in connection with various boundary approximations in later chapters. Sloan [1982] has used the above algorithm to this end when the interior method is a fourth order leap-frog scheme.

# CHAPTER 3

## BOUNDARY TECHNIQUES FOR THE LAX-WENDROFF METHOD

The first interior approximation we shall consider is that proposed by Lax and Wendroff [1960]. Their scheme is widely used and, together with the Leap-Frog method (4.1), has probably become the most frequently analysed of all finite difference schemes; see for example Richtmyer [1963], Richtmyer and Morton [1967], Morton [1971].

The test problems, described earlier, have related Cauchy problems of the form

$$\underset{\sim}{u}_t(x,t) = A\underset{\sim}{u}_x(x,t) \qquad x \in \mathbb{R}, \qquad t > 0 \tag{3.1}$$

$$\underset{\sim}{u}(x,0) = \underset{\sim}{f}(x) \qquad x \in \mathbb{R},$$

where $\underset{\sim}{u}(x,t) = (u(x,t)v(x,t))^T$ and $\underset{\sim}{u}$ and $\underset{\sim}{f}$ are real vector functions, with $A$ a constant real square matrix of appropriate dimensions. Approximating $\underset{\sim}{u}$ on a mesh with spacing $\Delta x$ and $\Delta t$ by a mesh function $\underset{\sim}{U}_\Delta$ and letting $\underset{\sim}{U}_j^n = (U_j^n, V_j^n)^T$ denote the value of $\underset{\sim}{U}_\Delta$ at $x_j = j\Delta x$, $t_n = n\Delta t$, the Lax-Wendroff (L-W) approximation to the solution $\underset{\sim}{u}(j\Delta x,(n+1)\Delta t)$ is given by

$$\underset{\sim}{U}_j^{n+1} = (A_{-1}E^{-1} + A_o + A_1E)\underset{\sim}{U}_j^n, \qquad j \in \mathbf{Z} \tag{3.2}$$

with initial data $\underset{\sim}{U}_j^o = \underset{\sim}{f}(j\Delta x)$, $j \in \mathbf{Z}$. In standard notation E is the shift operator $E^{\nu}\underset{\sim}{U}_j^n = \underset{\sim}{U}_{j+\nu}^n$ and, for constant mesh ratio $\lambda = \Delta t/\Delta x$, and the identity matrix I,

$$A_{-1} = \tfrac{1}{2}\lambda A(\lambda A-I), \quad A_o = I-\lambda^2A^2, \quad A_1 = \tfrac{1}{2}\lambda A(\lambda A+I) \quad .$$

The L-W method has second order truncation error in both space and time, is also explicit, dissipative of order 4 and stable if $\lambda\rho(A) \leqslant 1$, where $\rho(A)$ denotes the spectral radius of A.

Gadd [1978] defended approximation (3.2) against "unfavourable appraisals" and attempted to correct the problem of large phase errors by devising a third order variant. Fromm [1968] has also produced an adaptation with very little dispersion. These extensions, and the others that produce fourth order modifications, invariably increase the support of the scheme to at least five points at the lower time level. When solving bounded domain problems these methods then require intermediate boundary approximations in addition to the usual boundary approximations. The construction of these intermediate expressions presents no difficulty (Strang [1980]) and their resulting stability analyses are no more complicated than those of the standard boundary approximations themselves. For simplicity, however, we proceed by considering boundary approximations for the standard Lax-Wendroff method (3.2). A compact fourth order interior approximation is considered in Chapter 5.

Chu and Sereny [1974] considered various boundary approximations in conjunction with the L-W method applied to the one-dimensional equations of gas dynamics. Their numerical experiments suggest that the 'best' boundary approximation would be based on the characteristic variables of the system. Sundström [1975] clarified some errors and misunderstandings on the part of Chu and Sereny and established the stability of all the approximations considered using the theory of Gustafsson et al [1972].

The matrix in (3.1) is of the form

$$A = \begin{bmatrix} \frac{1}{2}(q-1) & \frac{1}{2}(q+1) \\ \frac{1}{2}(q+1) & \frac{1}{2}(q-1) \end{bmatrix}, \quad q \in \mathbb{R}^+, \quad q \geqslant 1$$

and so we shall be extending the results of Sloan [1980] who considered the symmetric case q = 1. We are particularly interested in the effect that the uneven wave speeds have on the stability and accuracy of the boundary approximations. The selection of boundary approximations which we consider is different from that of Sloan; however we include his most successful method which was based on the characteristic variables. May and Morton [1976] considered the same problem as Sloan and, with (3.2), the most successful boundary approximation which they examined was that proposed by Matsuno [1966]. This approximation was also studied by Sloan [1980] and found to be competitive with the best characteristic-based technique and when applied, with (3.2), to system (3.1) a stable approximation resulted. However the approximation of Matsuno is not considered in detail here. Gottlieb and Turkel [1978] studied the differential system $\underset{\sim}{u}_t + A\underset{\sim}{u}_x = 0$ with q = 3. The authors transformed the system into characteristic form and, of all the boundary approximations studied, recommended a one-sided Euler treatment of the outgoing characteristic equation. Coughran [1980] examined the off-diagonally dominant problem given by (3.1) and $A = \begin{pmatrix} \alpha & 1 \\ 1 & \alpha \end{pmatrix}$, $\alpha \in [-\frac{1}{2},\frac{1}{2}]$. When using the L-W scheme in the interior, Coughran concluded that second order extrapolation at the boundary attained the best accuracy. Bramley and Sloan [1977] have considered the two-dimensional analogue of the problem in Sloan [1980]. The results again recommended the incorporation of characteristics into the boundary approximations.

## 3.1 Boundary approximations and Stability Analysis.

In this chapter we consider a variety of boundary approximations required in the implementation of the L-W method to the left and right quarter plane problems of (3.1). Unlike the results of Chapter 4 and Chapter 5 the stability of many of the approximations will be established analytically. This simplicity is due, in part, to the characteristic formulation of some of the approximations and, in part, to the dissipative nature of the L-W scheme. Section [3.3] involves the extrapolation of a quantity intermediate between the ingoing and outgoing characteristic variables and is taken from Jamieson and Sloan [1983].

### 3.1.1 The Left Boundary problem.

We consider first the left quarter-plane problem

$$\begin{aligned} &\underset{\sim}{u}_t = \begin{bmatrix} a & b \\ b & a \end{bmatrix} \underset{\sim}{u}_x, \quad x \in \mathbb{R}^+, \quad t > 0, \\ &\underset{\sim}{u}(x,0) = \underset{\sim}{f}(x), \quad u(0,t) = g(t), \\ &\underset{\sim}{u}(x,t) = [u(x,t),v(x,t)]^T, \quad a = \tfrac{1}{2}(q-1), \ b = \tfrac{1}{2}(q+1). \end{aligned} \tag{3.3}$$

We approximate the solution $\underset{\sim}{u}(j\Delta x, n\Delta t)$ by $\underset{\sim}{U}{}^n_j$ for $n \geqslant 0$, $j \geqslant 1$ using the L-W scheme (3.2). Following the theory of Gustafsson et al [1972] we analyse stability by seeking a solution to (3.2) of the form $\underset{\sim}{U}{}^n_j = z^n \kappa^j \underset{\sim}{d}$, where $\underset{\sim}{d}$ is a vector of complex constants and $z, \kappa \in \mathbb{C}$. Substitution of the desired solution in (3.2) yields

$$[zI - \kappa^{-1}A_{-1} - A_o - \kappa A_1]\underset{\sim}{d} = \underset{\sim}{0}. \qquad (3.4)$$

The condition for a non-trivial solution $\underset{\sim}{d}$ is that $\kappa$ and $z$ satisfy

$$\det[zI - \kappa^{-1}A_{-1} - A_o - \kappa A_1] = 0,$$

which reduces to the characteristic equations

$$\kappa(z-1) + \tfrac{1}{2}\lambda\mu(\kappa^2-1) - \tfrac{1}{2}\lambda^2\mu^2(\kappa-1)^2 = 0; \quad \mu = 1,-q \qquad (3.5a,b).$$

To construct a general solution we require the eigenvectors $\underset{\sim}{d}$ associated with (3.5a,b). It is readily shown that if $\kappa$ and $z$ satisfy (3.5a) then $\underset{\sim}{d}^T = [1,-1]$, and if $\kappa$ and $z$ satisfy (3.5b) then $\underset{\sim}{d}^T = [1,1]$. Denoting by $\kappa_1(z;\lambda)$ and $\kappa_2(z;\lambda)$ the roots of (3.5a) and by $\kappa_3(z;\lambda q)$ and $\kappa_4(z;\lambda q)$ the roots of (3.5b), for a given $z$, the general solution is given by

$$\underset{\sim}{U}{}^n_j = z^n\left[(\eta_1\kappa_1^j + \eta_2\kappa_2^j)\begin{bmatrix}1\\-1\end{bmatrix} + (\eta_3\kappa_3^j + \eta_4\kappa_4^j)\begin{bmatrix}1\\1\end{bmatrix}\right], \qquad (3.6)$$

where $\eta_i$ $(i = 1,2,3,4)$ is an arbitrary scalar. We have dropped the dependence of $\kappa_i$ in (3.6) and will continue to do so when the parameter dependence is clear. It is essential that, for $|z| > 1$, the general solution (3.6) decays as j increases and to this end we quote

<u>Lemma 3.1.</u> (Gustafsson et al [1972]). There is a $\delta > 0$ such that

(a) $|\kappa_1| < 1$ for $|z| \geqslant 1$, $z \neq 1$ and $\kappa_1(1;\lambda) = 1$,
$|\kappa_2| \geqslant 1+\delta$ for $|z| \geqslant 1$, and

(b) $|\kappa_3| \leqslant 1-\delta$ for $|z| \geqslant 1$,
$|\kappa_4| > 1$ for $|z| \geqslant 1$, $z \neq 1$ and $\kappa_4(1;\lambda q) = 1$.

The results of Lemma 3.1 then yield the desired general solution in the space $L_2(\Delta x)$ as

$$\underset{\sim}{U}_j^n = z^n\left[\eta_1\kappa_1^j\begin{bmatrix}1\\-1\end{bmatrix} + \eta_3\kappa_3^j\begin{bmatrix}1\\1\end{bmatrix}\right]. \qquad (3.7)$$

This is a particular example of equation (1.31).
The roots $\kappa_1$ and $\kappa_3$ will be termed the inner roots and $\kappa_2$ and $\kappa_4$ the outer roots. With (3.5a,b), (3.7) and Lemma 3.1 available we are now at liberty to analyse any boundary approximation involved in the L-W integration of the left boundary problem (3.3). The necessity of prescribing $V_o^{n+1}$ is a feature of the numerical problem and not of the differential problem - (3.3) being well-posed. The boundary approximations to $v(0,(n+1)\Delta t)$ to be considered are

$s^{th}$-order extrapolation $(EX_s)$ : $\Delta^{s+1}V_o^{n+1} = 0$, with $\Delta V_j = V_{j+1}-V_j$. (3.8)

Characteristics (C):

$$R_o^{n+1} = \tfrac{1}{2}(\lambda q-1)(\lambda q-2)R_o^n + \lambda q(2-\lambda q)R_1^n + \tfrac{1}{2}\lambda q(\lambda q-1)R_2^n, \qquad (3.9)$$

where R = V + U is the outgoing characteristic variable,

½-Int Conservation (½IC):

$$V_o^{n+1} = V_o^n + \lambda b(U_o^n+U_1^n) + \lambda a(V_1^n-V_o^n) - 2b\lambda U_o^{n+\frac{1}{2}} + \lambda^2 ab(U_1^n-U_o^n) + \lambda^2 b^2(V_1^n-V_o^n), \qquad (3.10)$$

Box (B):

$$(1+a\lambda)V_o^{n+1} + (1-a\lambda)V_1^{n+1} = (1-a\lambda)V_o^n + (1+a\lambda)V_1^n + \lambda b(U_1^{n+1}-U_o^{n+1}+U_1^n-U_o^n), \qquad (3.11)$$

Characteristic Euler (CE):

$$R_o^{n+1} = R_o^n + \lambda q(R_1^n-R_o^n). \qquad (3.12)$$

Sloan [1980] analysed, for q = 1, the boundary approximations $EX_o$, C and ½IC.

Approximation C is obtained by using quadratic interpolation to estimate the intersection of the line $t = t_n$ and the outgoing characteristic $\frac{dx}{dt} = -q$ through the point x = 0, t = (n+1)Δt. ½IC is constructed so that the conservation property of (3.1) is maintained. B is the usual box integration of the second physical equation and approximation CE is derived from a one sided Euler treatment of the characteristic equation $r_t = qr_x$ where $r(x,t) = u(x,t) + v(x,t)$.

The approximation to $u(0,(n+1)\Delta t)$ is given exactly by setting $U_o^{n+1} = g((n+1)\Delta t)$ and the value of $V_o^{n+1}$ is given by any one of the approximations (3.8)-(3.12). Prescribing the values of $U_o^{n+1}$ and $V_o^{n+1}$ is in keeping with the result of Lemma 1.1 in Chapter 1 where it was proved that $np_\ell$ boundary values were required at the boundary. For Lax-Wendroff interior method $np_\ell$ is equal to 2. Suppose $V_o^{n+1}$ is obtained from the extrapolation condition $EX_o$. In this case we substitute the general solution (3.7) into the boundary conditions $U_o^{n+1} = g((n+1)\Delta t)$ and $V_o^{n+1} = V_1^{n+1}$ to obtain

$$\eta_1 + \qquad \eta_3 = g((n+1)\Delta t)/z^{n+1},$$

$$(1-\kappa_1)\eta_1 + (\kappa_3-1)\eta_3 = 0$$

This is the algebraic system to which Theorem 1.6 applies. Therefore we may state, using Theorem 1.6, that $EX_o$ is a stable approximation if $\forall |z| \geqslant 1$ the determinant condition

$$\kappa_1 + \kappa_3 - 2 \neq 0 \qquad (3.13)$$

holds. For all the above boundary approximations an associated determinant equation, given by equality in (3.13) can be constructed in a similar manner. These equations are contained in Table 3.1.

| Boundary Approximation | Determinant Equation | |
|---|---|---|
| $EX_s$ | $(\kappa_1-1)^{s+1}+(\kappa_3-1)^{s+1} = 0$ | (3.14) |
| C | $z = c_o + c_1\kappa_3 + c_2\kappa_3^2$ | (3.15) |
| ½IC | $z = h_o\kappa_1 + h_1\kappa_3 + h_2$ | (3.16) |
| B | $b_oz = b_1$ | (3.17) |
| CE | $z = 1+\lambda q(\kappa_3-1)$ | (3.18) |

Table 3.1: Determinant equation of Boundary Approximations (3.8) - (3.12).

In Table 3.1 we have used the notation

$$c_o = \tfrac{1}{2}(\lambda q-1)(\lambda q-2); \quad c_1 = \lambda q(2-\lambda q); \quad c_2 = \tfrac{1}{2}\lambda q(\lambda q-1);$$

$$h_o = \tfrac{1}{2}\lambda(\lambda b-1); \quad h_1 = \tfrac{1}{2}\lambda q(\lambda b+1); \quad h_2 = 2(1-a\lambda) - 2\lambda^2b^2$$

$$b_o = 2(1+a\lambda) + (1+\lambda)\kappa_1 + (1-\lambda q)\kappa_3, \quad \text{and}$$

$$b_1 = 2(1-a\lambda) + (1-\lambda)\kappa_1 + (1+\lambda q)\kappa_3.$$

To illustrate the stability analysis consider the $EX_s$ approximation. Stability of the initial-boundary value quarter-plane difference problem will be assured if there are no solution triples $(\kappa_1,\kappa_3,z)$, of the system

$$\begin{aligned}
\kappa_1(z-1) + \frac{\lambda}{2}(\kappa_1^2-1) - \frac{\lambda^2}{2}(\kappa_1-1)^2 &= 0\\
\kappa_3(z-1) - \frac{\lambda}{2}\, q(\kappa_3^2-1) - \tfrac{1}{2}\lambda^2q^2(\kappa_3-1)^2 &= 0\\
(\kappa_1-1)^{s+1} + (\kappa_3-1)^{s+1} &= 0,
\end{aligned}$$

for which z is an eigenvalue or generalised eigenvalue. If such a triple did exist then there would be a general solution (3.7) which was bounded in space but unbounded in time, and this is an unstable solution. The above equations constitute the system referred to as (2.1) in Chapter 2. We have

Lemma 3.2. The L-W method and the $EX_s$ boundary approximation define a stable approximation for the left quarter plane problem if $\lambda q \leqslant 1$ and $s < 2$ for $q = 1$ or $s \leqslant 2$ for $q > 1$.

Proof. The restriction $\lambda q \leqslant 1$ is the necessary stability condition for the L-W approximation of the pure Cauchy problem. For the remainder of this chapter we assume that this inequality is satisfied. The result for $q = 1$ can be established analytically, whereas, for general q, numerical methods must be used on a discrete data set $(\lambda,q)$. The details are given in section [3.3] where the $EX_s$ approximation is a special case of the boundary approximation considered therein.

Similarly,

Lemma 3.3. The L-W method together with either the C or the CE boundary approximations define a stable approximation to the left quarter plane problem if $\lambda q \leqslant 1$.

Proof. Substitution of the appropriate determinant equation into the characteristic equation (3.5b) produces a polynomial in $\kappa_3$. In both cases the roots of the polynomial are multiples of the root $\kappa_3$ and this has the value unity. Lemma 3.1 establishes stability.

Finally we can establish

Result 3.1. For the discrete set $\{(\lambda,q) \; ; \; q = 1,2,\ldots,10; \; \lambda q = 0.95\}$ the L-W method with either the ½IC or the B approximation defines a stable approximation of the left quarter plane problem.

For the $\frac{1}{2}$IC approximation the result for q = 1 was established analytically by Sloan [1980]. The remaining conclusions of Result 3.1 were obtained by reducing the corresponding multivariate system to a single polynomial in one complex unknown. Using the algorithm of Grant and Hitchins all the roots that were obtained indicated stability. For the $\frac{1}{2}$IC approximation the reduction to one polynomial was straightforward. However the B technique required the use of the Resultant method of Chapter 2 (an analogous application is given in Appendix I).

In summary, we have established, using the stability theory of Gustafsson et al given in Chapter 1, the stability of the boundary approximations (3.8) - (3.12) when used in conjunction with the L-W method to approximate the left boundary problem of (3.1). For the extrapolation approximations stability was shown to be directly related to the value of q. Whilst stability of the $\frac{1}{2}$IC and B approximations can only be guaranteed for the above data set we have no evidence to suggest that there exist values of $\lambda$ and q that would produce instabilities.

### 3.1.2 The Right Boundary Problem

We now consider the potentially less stable right boundary version of (3.3). The domain is now $\{(x,t) : x \leqslant 1,\ t > 0\}$ and we specify $u(1,t) = h(t)$. If the right boundary is at $x = 1 = J\Delta x$ then approximation (3.2) is applied for all integers $j \leqslant J-1$. The right boundary problem is synonymous with a left boundary problem where the inward and outward characteristics have been interchanged. Therefore, if we seek a general solution of the form $\underset{\sim}{U}^n_{J-j} = z^n \kappa^{-j} \underset{\sim}{d}$

in (3.3), the characteristic equations are those of (3.5) with $\lambda$ negated, namely

$$\kappa(z-1) - \tfrac{1}{2}\lambda\mu(\kappa^2-1) - \tfrac{1}{2}\lambda^2\mu^2(\kappa-1)^2 = 0, \qquad \mu = 1 \text{ or } -q. \quad (3.19a,b)$$

If we denote by $\kappa_2(z;\lambda)$ and $\kappa_4(z;\lambda q)$ the exterior roots of (3.19a) and (3.19b) respectively, then the general solution which decays as j increases is

$$\underset{\sim}{U}^n_{J-j} = z^n\left[d_2\kappa_2^{-j}\begin{bmatrix}1\\-1\end{bmatrix} + d_4\kappa_4^{-j}\begin{bmatrix}1\\1\end{bmatrix}\right]. \qquad (3.20)$$

Examination of (3.19) and (3.5) will show that

$$\kappa_2^{-1}(z;\lambda) = \kappa_1(z;-\lambda), \qquad \kappa_4^{-1}(z;\lambda q) = \kappa_3(z;-\lambda q), \qquad (3.21)$$

where $\kappa_1(z;-\lambda)$ and $\kappa_3(z;-\lambda q)$ are the respective inner roots of (3.19a,b). Solution (3.20) then becomes

$$\underset{\sim}{U}^n_{J-j} = z^n\{d_2\kappa_1^j\begin{bmatrix}1\\-1\end{bmatrix} + d_4\kappa_3^j\begin{bmatrix}1\\1\end{bmatrix}\} \quad . \qquad (3.22)$$

We then require

Lemma 3.4. (Gustafsson et al [1972]). There is a $\delta > 0$ such that, if $\kappa_1$ and $\kappa_3$ are the interior roots of (3.19) then,

$$|\kappa_1| \leqslant 1-\delta \quad \text{for } |z| \geqslant 1 \quad \text{and} \quad |\kappa_3| < 1 \text{ for } |z| \geqslant 1,\ z \neq 1$$

and $\kappa_3 = 1$ for $z = 1$.

The boundary approximations, applied to the right boundary problem, take the form

$$\mathrm{EX}_s \;:\; \nabla^{s+1} V_J^{n+1} = 0, \text{ with } \nabla V_j = V_j - V_{j-1}, \tag{3.23}$$

$$\mathrm{C} \;:\; W_J^{n+1} = \tfrac{1}{2}(\lambda-1)(\lambda-2)W_J^n + \lambda(2-\lambda)W_{J-1}^n + \tfrac{1}{2}\lambda(\lambda-1)W_{J-2}^n;\; W = U-V, \tag{3.24}$$

$$\begin{aligned}\tfrac{1}{2}\mathrm{IC} \;:\; V_J^{n+1} = V_J^n - \lambda b(U_J^n+U_{J-1}^n) - a\lambda(V_{J-1}^n-V_J^n) + 2b\lambda U_J^{n+\frac{1}{2}} + \lambda^2 ab(U_{J-1}^n-U_J^n)\\ + \lambda^2 b^2(V_{J-1}^n-V_J^n),\end{aligned} \tag{3.25}$$

$$\begin{aligned}\mathrm{B} \;:\; (1-a\lambda)V_J^{n+1} + (1+a\lambda)V_{J-1}^{n+1} = (1+a\lambda)V_J^n + (1-a\lambda)V_{J-1}^n\\ - \lambda b\{U_{J-1}^{n+1}-U_J^{n+1}+U_{J-1}^n-U_J^n\},\end{aligned} \tag{3.26}$$

$$\mathrm{CE} \;:\; W_J^{n+1} = W_J^n + \lambda(W_{J-1}^n-W_J^n). \tag{3.27}$$

The determinant equation may be constructed as before and stability analysed analogously. The only result that differs from the left boundary problem is contained in

Lemma 3.5. The L-W method and $\mathrm{EX}_s$ boundary approximation define a stable approximation for the right quarter-plane problem if $\lambda q \leqslant 1$ and $s < 2$ for all q.

Proof. See section[3.3].

Lemma 3.5 illustrates the greater sensitivity, in terms of stability, of approximations applied on the boundary at which the faster wave is entering the computational domain. The other boundary approximations, (3.24)-(3.27), can be shown to define a stable approximation, to the right boundary situation, using similar techniques to those used in Lemma 3.3 and Result 3.1.

Again we have established the effect that the value of q has on the stability properties of the boundary approximations (3.23) - (3.27). Lemma 3.5 illustrates that it is insufficient to consider only one boundary of twin boundary problems of the form (3.1). We have shown that the numerical treatment of the boundary at which the slower wave is incident is more likely to induce computational instabilities than that which reflects the faster wave.

In the next section we illustrate the practical application of the above results.

## 3.2 Numerical Results

To illustrate the stability results of the previous section we consider the numerical solution of

$$\underset{\sim}{u}_t = \begin{bmatrix} a & b \\ b & a \end{bmatrix} \underset{\sim}{u}_x, \quad x \in [0,1], \quad t \geqslant 0,$$

$$\underset{\sim}{u}(x,t) = (u(x,t), v(x,t))^T, \quad a = \tfrac{1}{2}(q-1), \quad b = \tfrac{1}{2}(q+1). \qquad (3.28)$$

In this thesis we are considering the effect that differing wave speeds have on boundary approximations. It is clear that the fastest wave speed has a crucial role to play on the stability of many interior difference schemes, including the Lax-Wendroff method. To control the existence and strength of the faster characteristic waves we introduce the parameter $\varepsilon \in \mathbb{R}$ and define the initial data

$$\underset{\sim}{u}(x,0) = \begin{bmatrix} \varepsilon\ 2\pi \sin 2\pi x + 2\pi \cos 2\pi x \\ \varepsilon\ 2\pi \sin 2\pi x - 2\pi \cos 2\pi x \end{bmatrix} . \qquad (3.28a)$$

If $\varepsilon = 0$ then the effect of the faster wave is filtered out of problem (3.28). The general solution of (3.28) and (3.28a) is given by

$$u(x,t) = \varepsilon\ 2\pi \sin 2\pi(x+qt) + 2\pi \cos 2\pi(x-t)$$

and $\qquad\qquad (3.29)$

$$v(x,t) = \varepsilon\ 2\pi \sin 2\pi(x+qt) - 2\pi \cos 2\pi(x-t)$$

from which we obtain the boundary conditions $u(0,t)$ and $u(1,t)$. To analyse the effect of the left boundary, $v(0,t)$ is approximated by any of (3.8) - (3.12) and $v(1,t)$ is given by (3.29). The treatment is reversed for the right problem. We use the following error measurements

a) The maximum norm, $\max_j |\underset{\sim}{u}(j\Delta x,t_n) - \underset{\sim}{U}_j^n|$ , denoted by $\|\underset{\sim}{U}_\Delta\|_\infty$ , at any given time level $t = n\Delta t$.

b) If we define, for the exact solution $\underset{\sim}{u}(x,t_n)$, the quantity

$$\|\underset{\sim}{u}(\cdot,t_n)\|^2_{E_{q,\varepsilon}} = \int_0^1 [\underset{\sim}{u}(x,t_n]^T E_q [\underset{\sim}{u}(x,t_n)] dx \qquad (3.30)$$

where $E_q$ is the positive definite matrix $\begin{bmatrix} b & -a \\ -a & b \end{bmatrix}$ , then we can show

$\frac{\partial}{\partial t} \|\underset{\sim}{u}(\cdot,t_n)\|^2_{E_{q,\varepsilon}} = 0$. Therefore (3.30) is invariant in time and the value of (3.30) will be $\|\underset{\sim}{u}(\cdot,0)\|^2_{E_{q,\varepsilon}}$ for all time.

For the initial data (3.28)

$$||\underset{\sim}{u}(\cdot,0)||^2_{E_{q,\varepsilon}} = 4\pi^2(q+\varepsilon^2).$$

The analogous norm for the discrete problem is denoted by $||\underset{\sim}{U}_\Delta||_{E_{q,\varepsilon}}$ at any given time level and is given by a trapezoidal integration of right side of (3.30). This norm could have been constructed via the results of Gunzburger and Plemons [1979]. The results are summarized in Tables 3.2 - 3.6.

| | $q = 1$ | $q = 2$ | $q = 5$ |
|---|---|---|---|
| $\varepsilon = 0.01$ | 39.482 | 78.960 | 197.39 |
| $\varepsilon = 0.5$ | 49.349 | 88.826 | 207.26 |

Table 3.2: $||\underset{\sim}{u}(x,0)||^2_{E_{q,\varepsilon}}$ for various $q,\varepsilon$.

All the integrations were performed over three fast wave cycles. The numerical results indicate a degree of insensitivity of approximation $EX_2$, with $q = 1$, for smooth initial data. Even when the initial data were corrupted by a random number generator, adding a 25% variation, stability remained.

It is clear that there is a degree of choice available among both the left and right boundary approximations. Also the 'good' boundary approximations are, for this smooth problem at least, insensitive to whether the fast characteristic is ingoing or not. The tables also illustrate the results of Lemmas 3.2 and 3.5. Tables 3.3 and 3.4 indicate the increased sensitivity of the right boundary

| Boundary Approximation | q = 1 | | q = 2 | | | | q = 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Left | | Right | | Left | | Right | |
| Exact | 39.485 | 0.003 | 79.011 | 0.023 | 79.011 | 0.023 | 197.41 | 0.023 | 197.41 | 0.023 |
| $EX_o$ | 38.995 | 0.481 | 78.013 | 0.496 | 79.121 | 0.338 | 197.28 | 0.490 | 197.41 | 0.348 |
| $EX_1$ | 39.954 | 0.071 | 79.939 | 0.073 | 79.015 | 0.044 | 198.96 | 0.064 | 197.41 | 0.046 |
| $EX_2$ | 39.490 | 0.002 | 79.020 | 0.021 | 1.2(16) | 5.8(8) | 197.41 | 0.023 | 1.1(6) | 5.5(3) |
| C | 39.485 | 0.003 | 79.012 | 0.023 | 79.013 | 0.047 | 197.41 | 0.023 | 197.41 | 0.032 |
| ½IC | 39.551 | 0.010 | 79.094 | 0.025 | 79.013 | 0.053 | 197.41 | 0.023 | 197.41 | 0.027 |
| B | 39.486 | 0.003 | 79.013 | 0.023 | 79.013 | 0.047 | 197.41 | 0.023 | 197.41 | 0.032 |
| CE | 39.485 | 0.003 | 79.012 | 0.024 | 79.013 | 0.052 | 197.41 | 0.023 | 197.41 | 0.027 |
| $EX_2(\alpha = 1.0)$ | 39.486 | 0.003 | 79.012 | 0.024 | 79.013 | 0.048 | 197.41 | 0.023 | 197.41 | 0.033 |

Table 3.3: $\|\underset{\sim}{U}_\Delta\|_{E_{q,\varepsilon}}$ | $\|\underset{\sim}{U}_\Delta\|_\infty$ for $\lambda q = 0.95$, $\Delta x = \frac{1}{80}$, $\varepsilon = 0.01$.

| Boundary Approximation | q = 1 | | q = 2 | | | | q = 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Left | | Right | | Left | | Right | |
| Exact | 49.349 | 0.003 | 88.877 | 0.024 | 88.877 | 0.024 | 207.27 | 0.024 | 207.27 | 0.024 |
| $EX_o$ | 46.038 | 0.625 | 88.082 | 0.718 | 89.037 | 0.518 | 207.12 | 0.854 | 208.02 | 0.042 |
| $EX_1$ | 49.815 | 0.076 | 89.804 | 0.071 | 88.841 | 0.044 | 208.70 | 0.057 | 207.28 | 0.049 |
| $EX_2$ | 49.395 | 0.008 | 88.885 | 0.024 | 1.1(17) | 2.2(9) | 207.27 | 0.023 | 1.2(7) | 1.8(4) |
| C | 49.373 | 0.004 | 88.876 | 0.026 | 88.883 | 0.047 | 207.27 | 0.024 | 207.23 | 0.032 |
| ½IC | 49.438 | 0.014 | 88.958 | 0.026 | 88.977 | 0.060 | 207.34 | 0.024 | 207.36 | 0.045 |
| B | 49.373 | 0.004 | 88.877 | 0.025 | 88.883 | 0.047 | 207.27 | 0.024 | 207.23 | 0.032 |
| CE | 39.373 | 0.004 | 88.876 | 0.026 | 88.922 | 0.052 | 207.27 | 0.024 | 207.23 | 0.026 |
| $EX_2(\alpha = 1.0)$ | 49.391 | 0.007 | 88.876 | 0.027 | 88.882 | 0.048 | 207.27 | 0.024 | 207.22 | 0.033 |

Table 3.4: $\|\underset{\sim}{U}_\Delta\|_{E_{q,\varepsilon}} \mid \|\underset{\sim}{U}_\Delta\|_\infty$ for $\lambda q = 0.95$, $\Delta x = \frac{1}{80}$, $\varepsilon = 0.5$.

problem. On pages 158-160 the evolution of the initial data is depicted using the graphic facilities of an ICL 1904S. Plot 2 illustrates the dissipative effect of the Lax-Wendroff interior method when an unstable boundary approximation is applied. Typical exact solutions are shown in Plots 3 and 4.

## 3.3 Extrapolation of Characteristic Variables

The work in this section has published in the Internatioanl Journal for Numerical Methods in Engineering (Jamieson and Sloan [1983]).

In this section we consider the boundary approximations

$$\Delta^{s+1}(V_o^{n+1} + \alpha U_o^{n+1}) = 0, \ \alpha \in [-1,1], \qquad (3.31)$$

and

$$\nabla^{s+1}(V_J^{n+1} - \alpha V_J^{n+1}) = 0, \ \alpha \in [-1,1]. \qquad (3.32)$$

When $\alpha = 1$ it can be seen that the extrapolated quantity in (3.31) and (3.32) corresponds to the outgoing Riemann invariant and $\alpha = -1$ corresponds to the ingoing Riemann invariant. The value $\alpha = 0$ reduces (3.31) and (3.32) to (3.8) and (3.23), respectively.

Gottlieb and Turkel [1978] show that the zero-order extrapolation of a linear combination of U and V may be unstable when the variables involved in boundary conditions and boundary approximations are nearly linearly dependent. Sloan [1980] has shown that, for q = 1, (3.31) and (3.32) are stable for all $s \geqslant 0$ if $\alpha = 1$ and unstable for all $s \geqslant 0$ if $\alpha = -1$. In this section we wish to make precise the relationship between $q,\alpha$ and s that guarantees stability. Clearly $\alpha = 1$ is the optimal choice, however the nature of the physical problem may, in some cases, require a different quantity to be extrapolated.

Following the earlier approach we construct the determinant equations associated with the left and right quarter plane problems as

$$(1-\alpha)\{\kappa_1(z;\lambda)-1\}^{s+1} + (1+\alpha)\ \{\kappa_3(z;\lambda q)-1\}^{s+1} = 0 \qquad (3.33)$$

and

$$(1+\alpha)\{\kappa_1(z;-\lambda)-1\}^{s+1} + (1-\alpha)\{\kappa_3(z;-\lambda q)-1\}^{s+1} = 0 \qquad (3.34)$$

respectively.

For the symmetric problem q = 1, we have $\kappa_1(z;-\lambda) = \kappa_3(z;\lambda)$ and $\kappa_3(z;-\lambda) = \kappa_1(z;\lambda)$ and so (3.33) and (3.34) are, in fact, the same equations. Hence it is sufficient to consider only the left boundary problem with approximation (3.31). For q > 1, however, this symmetry is lost and both problems must be examined separately.

q = 1 Stability Analysis:

As before we seek to determine whether or not there are any solution triples $(\kappa_1,\kappa_3,z)$ of the system

$$\kappa_1(z-1) + \frac{\lambda}{2}(\kappa_1^2-1) - \tfrac{1}{2}\lambda^2(\kappa_1-1)^2 = 0, \qquad (3.35)$$

$$\kappa_3(z-1) - \frac{\lambda}{2}(\kappa_3^2-1) - \tfrac{1}{2}\lambda^2(\kappa_3-1)^2 = 0, \qquad (3.36)$$

$$(1-\alpha)(\kappa_1-1)^{s+1} + (1+\alpha)(\kappa_3-1)^{s+1} = 0 \qquad (3.33)$$

for which z is an eigenvalue or generalised eigenvalue. The determinant equation (3.33) can be written as

$$w_s(\kappa_1-1) = \rho_s(\kappa_3-1), \tag{3.37}$$

where $\rho_s = ((1+\alpha)/(1-\alpha))^{1/s+1}$ and $w_s$ is any root of the equation $w^{s+1} = -1$. Examination of (3.35) and (3.36) yields the relation

$$\kappa_3 = g\kappa_1, \quad g = (\lambda-1)/(\lambda+1) \in (-1,0). \tag{3.38}$$

We may combine (3.38) and (3.37) to obtain

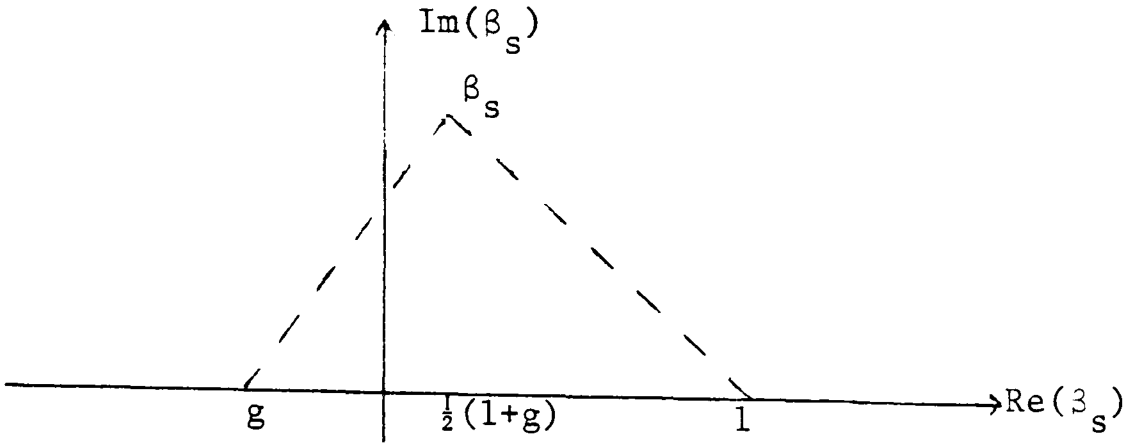
$$\kappa_1 = \frac{w_s-\rho_s}{w_s-g\rho_s} \tag{3.39}$$

which involves only the inner root of the characteristic equation (3.35) for any prescribed $\alpha$, $\lambda$ and s. The difference approximation involving (3.31) will be stable if the right hand side of (3.39) cannot be identified with the inner root of (3.35). Recourse to Lemma 3.1 indicates that $\kappa_1 = 1$ at $z = 1$ and $|\kappa_1| < 1$ for all $|z| \geqslant 1$. Since $g \neq 1$ for $\lambda \in (0,1)$ we can therefore dispense with the unit circle and conclude the sufficient stability requirement of

$$\left|\frac{w_s-\rho_s}{w_s-g\rho_s}\right| > 1 \quad \text{for all } w_s. \tag{3.40}$$

If we use the notation $\beta_s = w_s\rho_s^{-1}$ then (3.40) may be expressed as

$$|\beta_s-1| > |\beta_s-g|, \text{ for all } w_s. \tag{3.41}$$



The above diagram illustrates that (3.41) will be satisfied if the distance, in the complex plane, from $\beta_s$ to 1 exceeds that from $\beta_s$ to $g$. This is equivalent to requiring that

$$\mathrm{Re}(\beta_s) < \tfrac{1}{2}(1+g) = \frac{\lambda}{\lambda+1}\ , \text{ for all } \beta_s. \qquad (3.42)$$

Examination of the roots $w_s$ reveals that the largest value of $\mathrm{Re}(\beta_s)$ will be associated with $w_s = \exp\left[\pm\, i\pi/(s+1)\right]$ and so (3.42) will be satisfied if and only if

$$\cos\left[\pi/(s+1)\right] < \frac{\lambda\rho_s}{(\lambda+1)}\ . \qquad (3.43)$$

For $s = 0$ or $s = 1$ (3.43) is satisfied for any $\alpha \in (-1,1)$ and $\lambda \in (0,1)$. As we observed previously the limit values $\alpha = -1$ and $\alpha = 1$ yield unstable and stable difference approximations, respectively, regardless of the value of $s$. For all $s \geqslant 2$ the sufficient stability condition (3.43) may be rearranged to yield the restriction

$$\alpha > \frac{(\lambda+1)^{s+1} - \left[\lambda \sec(\pi/(s+1))\right]^{s+1}}{(\lambda+1)^{s+1} + \left[\lambda \sec(\pi/(s+1))\right]^{s+1}}\ . \qquad (3.44)$$

We therefore obtain

Lemma 3.6. For $\lambda \in (0,1)$ and $s \geqslant 2$ then a sufficient condition for stability of the L-W and (3.31) difference scheme is that $\alpha$ satisfies (3.44). If we denote the right-hand side of (3.44) by $\alpha_s^{(1)}$ for $s \geqslant 2$ and let $\alpha_o^{(1)} = \alpha_1^{(1)} = -1$, then the difference scheme is stable if $\alpha \in (\alpha_s^{(1)},1]$.

It can be seen that for $s > 0$ and for any $\lambda \in (0,1)$, the sequence $\{\alpha_s^{(1)}\}_{s=1}^{\infty}$ is monotonic, strictly increasing.

We have therefore obtained, by requiring that $|\kappa| < 1$, a sufficient interval relating to the degree of extrapolation in approximation (3.31) for any $\alpha \in (-1,1]$. However numerical experiments indicate that Lemma 3.6 is over-restrictive. Namely, solution triples $(\kappa_1,\kappa_3,z)$ were found which suggested that the true stability interval is of the form $(\alpha_s^{(1)} - \delta_s,1]$ for some $\delta_s > 0$. We calculated the $\alpha_s^{(1)}$. as the point corresponding to $\kappa$ moving inside the unit circle. Violation of the interval in Lemma 3.6 therefore allows the possibility of $|\kappa| < 1$; however, the approximation will only be unstable if $|z| \geqslant 1$. We proceed by obtaining the $\alpha$-interval such that for all $w_s$, the inequality $|z| < 1$ be satisfied. From (3.35) we have

$$z = \frac{1}{\kappa_1}\{\frac{\lambda^2}{2}(\kappa_1-1)^2 - \frac{\lambda}{2}(\kappa_1^2-1) + \kappa_1\}, \qquad (3.45)$$

where $\kappa_1$ is given by (3.39). Equation (3.45) may be reduced to

$$z = \frac{\beta_s^2 + g}{\beta_s^2 - (1+g)\beta_s + g} \qquad (3.46)$$

$$= \frac{\xi_s}{\xi_s - (1+g)} \ ,$$

where $\xi_s = \beta_s + a\beta_s^{-1}$.

Following the previous geometric argument in the complex plane, we obtain that $|z| < 1$ is equivalent to

$$\mathrm{Re}(\xi_s) < \tfrac{1}{2}(1+g) = \frac{\lambda}{\lambda+1} \quad . \qquad (3.47)$$

From lemma 3.6 $\alpha_s^{(1)} > -1$ for any s, so we need only consider $s > 2$ in (3.47). If we consider $\alpha = 0$ and $w_s = \exp(i\pi/(s+1))$ in (3.47) we find that $|z| < 1$ reduces to $\cos(\pi/(s+1)) < \tfrac{1}{2}$. This condition is violated if $s \geqslant 2$. Since (3.43) also fails we have that extrapolation of order $s \geqslant 2$ when $\alpha = 0$ is unstable for the model problem with $q = 1$. This proves the first part of Lemma 3.2 and confirms the result of Sloan [1980]. The search for an extension of the stability interval $(\alpha_s^{(1)}, 1|$ $(s \geqslant 2)$ may therefore be terminated on the left at $\alpha = 0$.

If (3.47) is written as

$$(g\rho_s + \rho_s^{-1}) r_s < \frac{\lambda}{\lambda+1} \ , \quad r_s = \mathrm{Re}(w_s),$$

we obtain the equivalent inequality

$$(\rho_s - \psi_s)(\rho_s - \gamma_s) > 0, \qquad (3.48)$$

where

$$\psi_s, \gamma_s = -\frac{1}{2g}\left[-L_s \mp \sqrt{\{L_s^2 - 4g\}}\right] \qquad (3.49)$$

and $L_s = \lambda/(r_s(1+\lambda))$.

Since $\rho_s > 0$, $\gamma_s > 0$ and $\psi_s < 0$ inequality (3.48) will be satisfied if $\rho_s > \gamma_s$. This is represented as a condition on $\alpha$ as

$$\alpha > \frac{\gamma_s^{1+s} - 1}{\gamma_s^{1+s} + 1} \tag{3.50}$$

In (3.50) we need only consider those roots $w_s$ for which the inequality (3.47) is violated. Denote the quotient in (3.50) by $\alpha_s$. From this we obtain

$$\frac{\partial}{\partial L_s} \alpha_s = -\gamma_s (L_s^2 - 4g)^{-\frac{1}{2}} < 0$$

and so $\alpha_s$ is a monotonic, strictly decreasing function of $L_s$. Therefore we need only consider that root $w_s$, in (3.49), corresponding to the largest value of $r_s$. This implies that $r_s = \cos(\pi/(s+1)) \equiv r_s^*$. If $\alpha_s^{(2)} = \max\limits_{w_s} \alpha_s$, then $\alpha_s^{(2)}$ is obtained from (3.50) with $r_s = r_s^*$ in defining relations (3.49). If $\alpha \in (\alpha_s^{(2)}, 1]$, then $|z| < 1$ for all $w_s$ and the difference approximation is stable. If $\alpha \leqslant \alpha_s^{(2)}$ then $|z| \geqslant 1$ and since $\alpha_s^{(2)} < \alpha_s^{(1)}$ then $|\kappa_1| < 1$, for $w_s = \exp(i\pi/(s+1))$, and the system is unstable. Defining $\alpha_0^{(2)} = \alpha_1^{(2)} = -1$ we have proved

Lemma 3.7. If $q = 1$, $\lambda \in (0,1)$ and $\alpha$ confined to the interval $[-1,1]$ then a necessary and sufficient condition for the approximation, defined by the L-W method and (3.31), to the left quarter-plane problem to be stable is that $\alpha \in (\alpha_s^{(2)}, 1]$.

For any $\lambda \in (0,1)$ we have $\alpha_2^{(2)} = 0$ and, for $\lambda = 0.95$, the values of $\alpha_3^{(2)}$ and $\alpha_4^{(2)}$ are 0.568 and 0.803, respectively, to three decimal places.

q > 1 Stability Analysis:

The previous results were established analytically due to the symmetry inherent in the problem. For general q however, the solution triples of the system composed of the two characteristic and the determinant equations must be found explicitly, and stability results obtained thereafter. Whilst the triples must be found using numerical techniques, recourse need not be made to the methods of Chapter 2. In fact, for both the left and right quarter plane problems the appropriate 'stability' system may be reduced to a single polynomial equation in one complex variable. We consider first the left-hand boundary at $x = 0$.

The analysis of the left quarter-plane problem involves the simultaneous solution of (3.5a,b) and (3.33) for the triple $(\kappa_1(z;\lambda),\ \kappa_3(z;\lambda q),z)$. By combining (3.5a) and (3.5b) to eliminate z and using the equivalent form of (3.33), equation (3.37), we obtain a real coefficient cubic in $\kappa_3$. For any $\lambda, q, \alpha$ and s the cubic has a root $\kappa_3 = 1$ leading to, after back substitution, $\kappa_1 = 1$ and $z = 1$. Lemma 3.1 indicates that the root $\kappa_3 = 1$ is actually the limit point, as $z \to 1$, of the outer root $\kappa_4(z;\lambda q)$, and thus the (1,1,1) triple may be regarded as 'stable'. Removing, from the cubic, the root $\kappa_3 = 1$ by deflation we obtain the quadratic

$$\mathcal{A}\kappa_3^2 + \mathcal{B}\kappa_3 + \mathcal{C} = 0 \tag{3.51}$$

where

$$\mathcal{A} = 1 - \lambda + q\beta_s(1+\lambda q)$$

$$\mathcal{B} = \lambda - 1 + 2\beta_s(1-\lambda q^2) + q\beta_s^2(1+\lambda q), \quad \text{and}$$

$$\mathcal{C} = q\beta_s(\lambda q - 1) + q\beta_s^2(1-\lambda q) \ .$$

For any values of $\lambda, q, s$ and $\alpha$, equation (3.51) is solved for $\kappa_3$, and $\kappa_1$ and z are obtained by back substitution. We note that if $\kappa_3$ is a root corresponding to $w_s$ then $\bar{\kappa}_3$ is a root associated with $\bar{w}_s$. Thus we need only consider $w_s$ such that $\mathrm{Im}(w_s) \geq 0$. As before, we consider the roots of (3.51) associated with $w_s = \exp(i\pi/(s+1))$. The stability boundary is then the least value of $\alpha$ for which the solution triple satisfies $|\kappa_1| < 1$, $|\kappa_3| < 1$, $|z| \geq 1$. Notice that any lower bound will only be approximate, however it is not expected that the value of $\alpha_s^{(2)}$ would be required to more than one decimal place.

As discussed earlier the right quarter plane problem may be analysed by negating $\lambda$ in the system of equations. If we denote the roots $\kappa_1(z;-\lambda)$ and $\kappa_3(z;-\lambda q)$ by $\hat{\kappa}_1$ and $\hat{\kappa}_3$ respectively and we modify the determinant equation to

$$\hat{\kappa}_3 - 1 = \hat{\beta}_s(\hat{\kappa}, -1), \tag{3.52}$$

where $\hat{\beta}_s = w_s \rho_s$, then the system (3.5a,b), (3.31) in $\kappa_1, \kappa_3$ and z transforms to the system (3.19a,b), (3.52) in $\hat{\kappa}_1, \hat{\kappa}_3$ and z if $\lambda$ is replaced by $-\lambda$ and $\alpha$ by $-\alpha$. The left and right boundary problems may therefore be handled by the previously described algorithm.

The stability intervals are given in Table 3.5 for q = 1,2 and s and s = 0,1,2,3 such that $\lambda q = 0.95$.

|  | q = 1 | q = 2 |  | q = 5 |  |
|---|---|---|---|---|---|
|  | Left and Right | Left | Right | Left | Right |
| s = 0 | (-1,1] | (-1,1] | (-1,1] | (-1,1] | (-1,1] |
| s = 1 | (-1,1] | (-1,1] | (-0.7,1] | (-1,1] | (-0.7,1] |
| s = 2 | (0,1] | (-0.2,1] | (0.3,1] | (-0.1,1] | (0.2,1] |
| s = 3 | (0.6,1] | (0.2,1] | (0.7,1] | (0.2,1] | (0.6,1] |

Table 3.5: Stability intervals with $\lambda q = 0.95$.

Examination of the intervals will show that the boundary which has the fast wave outgoing has the wider stability intervals. For $q > 1$ quadratic extrapolation may be stable at $x = 0$. This concludes the proof of Lemma 3.2. Table 3.5 also verifies the result of Lemma 3.5 for the right boundary problem.

To illustrate the results of Table 3.5 we apply the bounday approximations (3.31) and (3.32) to the model problem integrated earlier. The results are given below.

| $(s,\alpha)$ | $\|\underset{\sim}{U}_\Delta\|_E$ | $\|\underset{\sim}{U}_\Delta\|_\infty$ |
|---|---|---|
| (2, 0.1) | 39.489 | 0.0016 |
| (2, -0.1) | 39.380 | 0.530 |

Table 3.6 a: $q = 1, V_J^n = v(1,t_n)$

| $(s,\alpha)$ | $\|\underset{\sim}{U}_\Delta\|_E$ | $\|\underset{\sim}{U}_\Delta\|_\infty$ |
|---|---|---|
| (2, -0.1) | 79.022 | 0.021 |
| (2, -0.4 | 78.825 | 0.632 |

Table 3.6 b: $q = 2, V_J^n = v(1,t_n)$

| $(s,\alpha)$ | $\|\underset{\sim}{U}_{\Delta}\|_E$ | $\|\underset{\sim}{U}_{\Delta}\|_\infty$ |
|---|---|---|
| (1, -0.6) | 79.014 | 0.046 |
| (1, -0.8) | 79.028 | 0.105 |

Table 3.6 c: $q = 2, V_o^n = v(0,t_n)$

| $(s,\alpha)$ | $\|\underset{\sim}{U}_{\Delta}\|_E$ | $\|\underset{\sim}{U}_{\Delta}\|_\infty$ |
|---|---|---|
| (1, -0.6) | 197.41 | 0.046 |
| (1, -0.8) | 226.48 | 20.501 |

Table 3.6 d: $q = 5, V_o^n = v(0,t_n)$

All the above results were obtained for $\lambda q = 0.95$, $\Delta x = \frac{1}{80}$ and $\varepsilon = 0.01$. The loss of accuracy is clear if $\alpha$ is outwith the stability interval. If random 'noise' is introduced into the initial data then instability in the E-norm becomes obvious. Tables 3.3 and 3.4 illustrate the desirability of setting $\alpha = 1$ in the boundary approximation. The overall accuracy attained being comparable with the best alternatives in section 3.2.

## 3.4 Accuracy Analysis

Skollërmo [1975,1978] presented a partial insight into the measurement of the relative effect that any boundary approximation has on overall solution accuracy. In her analysis Sköllermo identified three separate components of the total error. The partition consisted of the error associated with the pure Cauchy problem, that associated with the boundary approximations and a third residual error function. It is the second component we consider here. The results of Sköllermo have been applied by Gottlieb and Turkel [1978] and Coughran [1980] to rank various boundary approximations in terms of the minimum number of mesh intervals per wavelength to attain a prescribed accuracy. Sloan [1980] used both this technique and that of the wave reflection analysis of Chu and Sereny [1974] to a similar end. For the boundary schemes applied by Sloan to the case q = 1, there was no appreciable difference in either approach and broad agreement was found with the numerical integrations. In this thesis we use the method of Sköllermo since the wave reflection analysis cannot be applied to implicit boundary approximations used in conjunction with implicit difference equations. Whilst this does not effect the methods in this chapter, the difference methods of chapters 4 and 5 are both implicit.

We proceed by defining error functions $\underset{\sim}{\varepsilon}_j^n = (\varepsilon_j^n, \theta_j^n)$, where $\varepsilon = u-U$, $\theta = v-V$ represent the approximation error in U and V at the point $(j\Delta x, n\Delta t)$. If the difference scheme is of the form $\underset{\sim}{U}_j^{n+1} = G\underset{\sim}{U}_j^n$, then it follows that $\underset{\sim}{\varepsilon}_j^{n+1} = G\underset{\sim}{\varepsilon}_j^n$. If the time domain is extended to $-\infty$ and if $\hat{\underset{\sim}{\varepsilon}}(x_j,\omega)$ denotes the Fourier transform of $\underset{\sim}{\varepsilon}_j$ then we have

$$z\hat{\underset{\sim}{\varepsilon}}(x_j,\omega) = G\hat{\underset{\sim}{\varepsilon}}(x_j,\omega) \tag{3.53}$$

where $z = \exp(2\pi\omega\Delta t i)$. By assuming a solution $\hat{\underset{\sim}{\varepsilon}}(x_j,\omega) = \kappa^j\Phi(\omega)$ of the difference equation (3.53) we obtain two characteristic equations whose inner roots and $\kappa_\alpha$ and $\kappa_\beta$. Therefore, for related eigenvectors $\underset{\sim}{y}_\alpha$, $\underset{\sim}{y}_\beta$ the general solution of (3.53) which is bounded as $j \to \infty$ is

$$\hat{\underset{\sim}{\varepsilon}}(x_j,\omega) = d_\alpha\kappa_\alpha^j\underset{\sim}{y}_\alpha + d_\beta\kappa_\beta^j\underset{\sim}{y}_\beta, \tag{3.54}$$

where $d_\alpha$ and $d_\beta$ are constants. The above characteristic equations will usually be the same equations as arise in the stability analysis. The differential equations can be treated in an analogous manner to yield general expressions for $\hat{\underset{\sim}{u}}(x,\omega)$, the Fourier transform of $\underset{\sim}{u}(x,t)$.

By rewriting any boundary approximation in terms of $\underset{\sim}{\hat{\xi}}$ and $\underset{\sim}{\hat{u}}$, after Fourier transformation, we obtain the Sköllermo error function

$$e(\omega) = \left|\frac{\mathfrak{Z}(z,\kappa^{\nu_1},\kappa^{\nu_2})}{\mathfrak{Z}(z,\kappa_\alpha,\kappa_\beta)}\right| , \qquad (3.55)$$

where

$\kappa = \exp(2\pi\omega\Delta xi)$ and $\nu_1$ and $\nu_2$ are the negative characteristic gradients of A. The function $\mathfrak{Z}(z,\kappa_\alpha,\kappa_\beta)$ is the determinant equation arising in Table 3.1, and $e(\omega)$ is the relative error in the Fourier coefficient with frequency $\omega$. We consider (3.55) as $\Delta t$ approaches zero, $\lambda$ held constant, and so we expand the numerator and denominator in powers of $\Delta x$. We need only consider the smallest power of $\Delta x$ in the expansion so we may replace $z,\kappa_\alpha$ and $\kappa_\beta$ in the denominator by their respective limits as $\Delta t$ approaches zero. These quantities will be determined from the characteristic equations using $\lim\limits_{\Delta t\to 0} z = 1$.

For the L-W interior approximation a detailed construction of $e(\omega)$, for $q = 1$, is given by Sloan [1980]. For general q a similar analysis yields

$$e(\omega) = \left|\frac{\mathfrak{Z}(z,\kappa^{-1},\kappa^{1/q})}{\mathfrak{Z}(z,\kappa_1,\kappa_3)}\right| \qquad (3.56)$$

Using lemma 3.1, we have $\kappa_1(1;\lambda) = 1$ and $\kappa_3(1;\lambda q) = (\lambda q-1)/(\lambda q+1) = \gamma$. The relative error is then

$$e(\omega) = \left|\frac{\mathfrak{Z}(e^{m\lambda},e^{-m},e^{m/q})}{\mathfrak{Z}(1,1,\gamma)}\right| , \quad m = 2\pi\omega\Delta xi \qquad (3.57)$$

which we expand in powers of $\Delta x$.

Consider the $EX_s$ boundary approximation. From Table 3.1

$$\mathfrak{F}(z,\kappa_1,\kappa_3) = (\kappa_1-1)^{s+1} + (\kappa_3-1)^{s+1}.$$

By expanding $e(\omega)$ and evaluating $(1,1,\gamma)$ as $[-2/(1+\lambda q)]^{1+s}$ we obtain

$$e(\omega) = \frac{\pi^{s+1}}{M^{s+1}}\left[q^{s+1} + (-1)^{s+1}\right](\lambda q+1)^{s+1}, \tag{3.58}$$

where $M = 1/\omega\Delta x$ measures the numbers of spatial intervals per wavelength. In the symmetric problem, $q = 1$, the expression (3.58) is, for even powers of extrapolation, given by

$$e(\omega) = \frac{2\pi^{s+2}}{M^{s+2}}(1+\lambda)^{s+1}. \tag{3.59}$$

A similar procedure yields the expressions in Table 3.10.

| Boundary Approx. | $e(\omega)$ |
|---|---|
| $EX_s$ | $\frac{\pi^{1+s}}{(qM)^{1+s}}\left[q^{s+1} + (-1)^{1+s}\right](1+\lambda q)^{1+s}$<br>$(\frac{2\pi^{s+2}}{M^{s+2}}(1+\lambda)^{s+1}$ when $q = 1)$ |
| C | $\frac{\pi^3}{3M^3q^3}(\lambda q-1)(\lambda q-2)(1+\lambda q)^2$ |
| ½IC | $\frac{2\pi^2}{M^2q^2}(1+\lambda q)(\lambda q-\lambda b^2+a)/(1+\lambda b)$<br>$(\frac{4\pi^3}{3}(1-\lambda^2)/M^3$ when $q = 1)$ |
| B | $\frac{\pi^3}{3M^3q^3}(1+\lambda q)(q^2+1-2\lambda^2q^2)$ |
| CE | $\frac{\pi^2}{q^2M^2}(\lambda^2q^2-1)$ |
| (3.31) | $\frac{\pi^{1+s}}{(qM)^{1+s}}(1+\lambda q)^{1+s}(\rho q+(-1)^{1+s})$; $\rho = \frac{(1-\alpha)}{(1+\alpha)}$ |

Table 3.7 : $e(\omega)$ for boundary approximations (3.8)-(3.12), (3.31).

| Boundary Approx. | $e(\omega)$ |
|---|---|
| $EX_s$ | $\frac{\pi^{s+1}}{(qM)^{s+1}} \left[q^{s+1} + (-1)^{s+1}\right] (1+\lambda)^{s+1}$ |
| C | $\frac{\pi^3}{2M^3} (\lambda-1)(\lambda-2)(1+\lambda)^2$ |
| $\frac{1}{2}$IC | $\frac{2\pi^2}{qM^2} (1+\lambda)(\lambda q-\lambda b^2-a)/(1+b\lambda)$ |
| B | $-\frac{(1+\lambda)}{2M^3q^3} \pi^3 q(q^2+1-2\lambda^2q^2)$ |
| CE | $\frac{\pi^2}{M^2} (\lambda^2-1)$ |
| (3.32) | $\frac{\pi^{s+1}}{(qM)^{s+1}} (1+\lambda)^{s+1}(q + (-1)^{s+1} \frac{(1-\alpha)}{(1+\alpha)})$ |

Table 3.8 : $e(\omega)$ for boundary approximations (3.23)-(3.27),(3.32)

The above expressions for $e(\omega)$ can be used to predict the number of points per wavelength that will be required to maintain a preset error tolerance. For $\lambda q = 0.95$ and a tolerance of 0.01 the necessary number of intervals is given in table 3.9.

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | |
|---|---|---|---|---|---|
| | | Left | Right | Left | Right |
| $EX_o$ | 63 | 307 | 232 | 491 | 299 |
| $EX_1$ | 87 | 69 | 52 | 63 | 39 |
| $EX_2$ | - | 28 | - | 29 | - |
| C | 6 | 3 | 13 | 2 | 13 |
| $\frac{1}{2}$IC | 8 | 15 | 23 | 12 | 29 |
| B | 6 | 10 | 11 | 8 | 11 |
| CE | 10 | 5 | 28 | 2 | 31 |

Table 3.9 : Minimum value of $M = 1/\omega\Delta x$ to achieve a tolerance of 0.01.

Boundary approximations that were previously shown to be unstable are omitted from Table 3.9. To provide a more meaningful illustration of the results of Table 3.9 we integrate, using the L-W scheme, the differential problem

$$\begin{aligned} \underset{\sim}{u}_t &= A\underset{\sim}{u}_x, \quad 0 \leq x \leq 1, \quad t \geq 0, \\ \underset{\sim}{u}(x,t) &= (u(x,t),\ v(x,t))^T, \qquad\qquad (3.60) \\ u(x,0) &= 0,\ v(x,0) = f(x) = 2\pi\cos 2\pi x, \end{aligned}$$

with the

$$\begin{aligned} u(0,t) &= 0.5(f(qt)-f(-t)) \\ u(1,t) &= 0.5(f(1+qt)-f(1-t)). \end{aligned}$$

The boundary conditions are homogeneous for q = 1. The results are contained in Table 3.10. For the symmetric problem the superior performance of $EX_o$ over $EX_1$, as predicted by Sköllermo, is verified

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | |
|---|---|---|---|---|---|
| | | Left | Right | Left | Right |
| Exact | 0.008 | 0.049 | - | 0.049 | - |
| $EX_o$ $(\alpha=\frac{1}{2})$ | 0.317 | 0.608 | 0.388 | 0.451 | 0.387 |
| $EX_o$ $(\alpha=0)$ | 0.147 | 0.832 | 0.614 | 0.630 | 0.422 |
| $EX_o$ $(\alpha=1)$ | 0.477 | 0.499 | 0.308 | 0.448 | 0.347 |
| $EX_1$ $(\alpha=0)$ | 0.295 | 0.156 | 0.096 | 0.159 | 0.083 |
| $EX_1$ $(\alpha=1)$ | 0.143 | 0.143 | 0.089 | 0.143 | 0.090 |
| $EX_2$ | 0.019 | 0.045 | 3951.0 | 0.050 | 23.559 |
| B | 0.015 | 0.047 | 0.094 | 0.049 | 0.063 |
| C | 0.015 | 0.048 | 0.094 | 0.048 | 0.064 |
| $\frac{1}{2}$IC | 0.015 | 0.039 | 0.108 | 0.049 | 0.178 |
| CE | 0.015 | 0.046 | 0.103 | 0.049 | 0.058 |

Table 3.10: $\|\underset{\sim}{U}_\Delta\|_\infty$ for $\lambda q = 0.95$, $\Delta x = 1/40$

together with high accuracy non-extrapolation boundary approximations. Despite non-homogeneous boundary conditions for q = 2 and 5 the ranking of Sköllermo is seen still to hold, although the superiority of the characteristic formulations is not as obvious. The relative accuracy of the left and right boundary approximations is also illustrated including the improved performance of extrapolation approximations when the fast wave is ingoing.

For variable extrapolation boundary approximations an examination of the Tables 3.7, 3.8 suggests that for q = 1 e(ω) is minimised for s = 0, if α = 0 and for s = 1 when α = 1. This result is illustrated by the numerical results of Tables 3.6a and 3.10.

The instability of the $EX_2$ approximation when $q = 1$ can be established analytically however it is very difficult to induce numerically even when large random errors are introduced into the initial data.

In this chapter we have considered various boundary approximations in conjunction with the Lax-Wendroff and the effects, thereon, of different time-scales. For twin boundary problems we have shown that both boundaries must be examined. Whilst there are many possible boundary approximations to choose from, we recommend that any choice must incorporate the characteristic variables of the differential system. This conclusion is based on both stability and accuracy results.

# CHAPTER 4

## A SEMI-IMPLICIT METHOD AND BOUNDARY APPROXIMATIONS

In the previous chapter the model fast wave problem,

$$\begin{array}{llll} \underset{\sim}{u}_t = A\underset{\sim}{u}_x & \underset{\sim}{u}(x,t) = [u(x,t),v(x,t)]^T & (x,t) \in [0,1] \times [0,T], & \\ \underset{\sim}{u}(x,0) = \underset{\sim}{f}(x) & x \in [0,\ 1] & , & (4.1) \\ u(0,t) = g(t), & u(1,t) = h(t) & t \in [0,T] & \end{array}$$

where

$$A = \tfrac{1}{2}\begin{bmatrix} q-1 & q+1 \\ q+1 & q-1 \end{bmatrix}, \quad q \in \mathbb{R}^+ ,$$

was solved numerically using the second order explicit, dissipative, Lax-Wendroff method with a variety of boundary approximations for v(0,t) and v(1,t). Acceptable accuracy was obtained for several approximations when (4.1) displayed differing time-scales. However, for large q the Cauchy stability requirement $\lambda q < 1$ resulted in progress in the temporal domain being very slow. As an alternative a less restrictive interior method may be used to obtain greater efficiency. This was discussed in the introduction.

Difference schemes which have weaker stability requirements include the implicit second-order accurate Crank-Nicolson scheme. This popular method has been applied to parabolic equations (Hopkins and Wait [1978]), barotropic equations (Sasky and Reddy [1979]) and biophysics (Joyner et al [1978]). For problems like (4.1), Sköllermo [1975] examined the effect on the accuracy of the Crank-Nicolson method produced by various boundary approximations. The interior approximation can be readily shown to be unconditionally stable for the pure Cauchy problem of (4.1) for any q. Implementation of the implicit method requires the inversion of a block tri-diagonal matrix at every time step and this increases the computational effort by a factor four over that of the Lax-Wendroff method. A feature of the Crank-Nicolson scheme is that all approximations of the spatial derivatives are averaged over time. This can result in a loss of accuracy when there is a significant signal travelling on the fast time-scale. For fast waves that are relatively weak the method represents a considerable improvement over the Lax-Wendroff scheme (Turkel [1981]).

An alternative approach is that of the semi-implicit method suggested by Kwizak and Robert [1971] in connection with the primitive equations in atmospherics. The basis of the method is to identify the elements of the physical equations that would normally be responsible for the conditional stability restriction and use a different method of approximation from that used on the remaining elements. Kwizak and Robert [1971] applied an implicit time averaging technique to the components of the spatial derivatives that involve the gravity waves and a centred differencing approach elsewhere. Elvius and Sundström [1973] applied this idea to a barotropic model based on the shallow-water equations, enhancing the scheme by basing the approximation on a staggered grid. This improvement reduces, by half, the computational time for problems of the form (4.1). Gauntlett, Leslie and Hicksman [1976] used a semi-implicit time differencing method to improve the efficiency of a six-level primitive equation model. The authors obtained an improvement of 2.5 in terms of computational cost over an explicit model for comparable accuracy.

In petroleum engineering, Nolen and Berry [1972] examined the problems relevant to the construction of a semi-implicit technique for use in reservoir simulation. Their simulator was found to be a competitive alternative to an explicit method being both stable and convergent in its linear and non-linear forms. Chappelear and Rogers [1974] presented practical guidelines for monitoring the time-step to control mathematical errors. When the semi-implicit approach is based on the Leap-Frog scheme the resulting method is unconditionally stable for the pure Cauchy problem of (4.1). Compared with the Crank-Nicolson scheme the Semi-Implicit Leap Frog (SILF) method should, as a result of the decreased degree of time-averaging, be able to deal more successfully with strong fast waves. This is borne out by numerical experiments, however for very weak fast waves the Crank-Nicolson approximation is still preferable.

As a consequence of the work already done on boundary approximations by Sköllermo [1975] and Coughran [1980] for use with the Crank-Nicolson method, together with the limitations as described above, we continue by investigating the SILF approximation.

## 4.1 Description and Application of the SILF method.

The standard explicit leap frog approximation to $\underset{\sim}{u}_t = A\underset{\sim}{u}_x$ is given by

$$\underset{\sim}{U}_j^{n+1} = \underset{\sim}{U}_j^{n-1} + \lambda A[\underset{\sim}{U}_{j+1}^n - \underset{\sim}{U}_{j-1}^n], \ \forall j, \ n \geq 1, \qquad (4.2)$$

where, as usual, we consider $\underset{\sim}{U}_j^n$ to be the finite difference approximation to $\underset{\sim}{u}(j\Delta x, n\Delta t)$. By using a mid-point time averaging rule for the terms in $A(\underset{\sim}{U}_{j+1}^n - \underset{\sim}{U}_{j-1}^n)$ involving the coefficient (q+1) we obtain the SILF method

$$\underset{\sim}{U}_j^{n+1} = \underset{\sim}{U}_j^{n-1} + a\lambda(\underset{\sim}{U}_{j+1}^{n} - \underset{\sim}{U}_{j-1}^{n}) + \tfrac{1}{2}b\lambda I'(\underset{\sim}{U}_{j+1}^{n+1} - \underset{\sim}{U}_{j-1}^{n+1} + \underset{\sim}{U}_{j+1}^{n-1} - \underset{\sim}{U}_{j-1}^{n-1}) \quad \forall j,\ n \geqslant 1 \tag{4.3}$$

where $a = \tfrac{1}{2}(q-1)$, $b = \tfrac{1}{2}(q+1)$ and $I' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$.

As in the Crank-Nicolson scheme, determination of $\underset{\sim}{U}^{n+1}$ from the two previous time levels involves the inversion of the matrix operating on $\underset{\sim}{U}^{n+1}$. This matrix, when a full grid is used, may be treated as either a block tridiagonal matrix, where each block is a $(2 \times 2)$ matrix, or as a septa-diagonal matrix. Numerical experiments, on an ICL 1904S, indicate that, using either approach, the CPU time to update $\underset{\sim}{U}^{n}$ is approximately 0.3 seconds. Notice that the evaluation of each component of $\underset{\sim}{U}_j^{n+1}$ requires only one value per grid point and thus permits the use of a staggered mesh. The block representation is no longer appropriate, in this situation.

Staggered or misaligned meshes have been used widely in meteorology (Elvius and Sundstrom [1973], Johns et al [1981]). Nickovic [1981] suggests that solving the primitive equations in meteorology on a full grid, when there are severe wind changes in the meridional direction, may result in computational instabilities. Staggered grids have also been used to obtain the solution of Poisson's equation (Sweet, Schuman [1976]). May and Morton [1976] used a method of staggering which, for (4.1) with $q = 1$, required no boundary approximations. Due to the popularity of such meshes, especially on large domain problems, we consider the SILF method on a staggered mesh together with a variety of boundary approximations. With the exception of Gustafsson et al [1972], very little analysis appears to have been performed on this type of problem in connection with the choice of boundary approximations.

The grid used throughout this chapter is of the form

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| (n+1)Δt | U | V | U | V | .... | U | V | U | V | .... | V |
| nΔt | V | U | V | U | .... | V | U | V | U | .... | U |
| (n-1)Δt | U | V | U | V | .... | U | V | U | V | .... | V |
| | 0 | Δx | 2Δx | | ....... | JΔx | 0 | Δx | 2Δx | .... | JΔx |
| | | | odd time steps | | | | | | even time steps | | |

and is such that, for J even, we only require boundary approximations for v(0,t), v(1,t) at every other time level. If non-linear terms are involved then any undefined values on the grid may be supplied through space and/or time averaging. Initiation of the scheme requires $\underset{\sim}{U}^{o}$ and $\underset{\sim}{U}^{1}$ where $\underset{\sim}{U}^{1}$ will be given by Taylor expansions, or by one sweep with a suitably accurate two-level scheme on $\underset{\sim}{U}^{o}$. Given these initial values the difference equations are defined as follows.

Even time steps:

$$\begin{aligned} -\frac{b\lambda}{2}\,\Delta_j V^{n+1} + U_j^{n+1} &= F_1(n,j) \;\; ; \; j = 1,3,5,\ldots,J-1, \\ -\frac{b\lambda}{2}\,\Delta_k U^{n+1} + V_k^{n+1} &= F_2(n,k) \;\; ; \; k = 2,4,6,\ldots,J-2, \end{aligned} \qquad (4.4)$$

where
$$\Delta_i \alpha^n = (\alpha_{i+1}^n - \alpha_{i-1}^n),$$
$$F_1(n,j) = U_j^{n-1} + a\lambda\Delta_j U^n + \frac{b\lambda}{2}\,\Delta_j V^{n-1} \quad \text{and}$$
$$F_2(n,k) = V_k^{n-1} - a\lambda\Delta_k V^n - \frac{b\lambda}{2}\,\Delta_k U^{n-1} \;.$$

Two further equations of the form

$$V_\alpha^{n+1} = f_\alpha(\underset{\sim}{V}^n, \underset{\sim}{V}^{n-1}), \quad \alpha = 0,J,$$

representing the left and right boundary approximations complete the system.

Odd-Time steps:

$$\frac{b\lambda}{2}\Delta_j U^{n+1} + V_j^{n+1} = F_2(n,j)\ ;\ j = 1,2,\ldots,J-1,$$
$$\frac{b\lambda}{2}\Delta_k V^{n+1} + U_k^{n+1} = F_1(n,k)\ ;\ k = 2,4,\ldots,J-2, \qquad (4.5)$$

together with the exact boundary values $U_o^{n+1} = g((n+1)\Delta t)$ and $U^{n+1} = h((n+1)\Delta t)$ form the approximation on odd time levels.

To aid further description we introduce the following notation. Along time step $t = 2n\Delta t$ let $\hat{V}_{2j}^{2n-1}$ $(j = 0,1,\ldots)$ and $\hat{U}_{2j-1}^{2n-1}$ $(j = 1,2,..)$ represent the finite difference approximations to $v(2j\Delta x, 2n\Delta t)$ and $u((2j-1)\Delta x, 2n\Delta t)$ respectively. Similarly along $t = (2n+1)\Delta t$ we have the approximations $V_{2j-1}^{2n+1}$ $(j = 1,2,..)$ and $U_{2j}^{2n+1}$ $(j = 0,1,\ldots)$.

We now invoke the stability theory of Gustafsson et al [1972] by considering two successive time levels. Using the transforms

$$w_j^n = z^n w_j\ ,\quad \hat{w}_j^n = z^{n+1}\hat{w}_j$$

and assuming solutions of the form

$$U_{2j} = \alpha\kappa^{2j},\ V_{2j-1} = \beta\kappa^{2j-1},\ \hat{U}_{2j-1} = \hat{\alpha}\kappa^{2j-1},\ \hat{V}_{2j} = \hat{\beta}\kappa^{2j},\ z,\kappa \in C,$$

where $\alpha,\beta,\hat{\alpha}$ and $\hat{\beta}$ are arbitrary constants.

The condition for non-trivial $\alpha,\beta,\hat{\alpha}$ and $\hat{\beta}$ gives rise to the determinant condition

$$\begin{vmatrix} L & -N & -M & 0 \\ -N & L & 0 & -M \\ -M & 0 & L & -N \\ 0 & -M & -N & L \end{vmatrix} = 0,$$

where

$$L = (z^2-1),$$
$$M = a\lambda z(\kappa-\kappa^{-1}),$$
$$N = \frac{b\lambda}{2}(z^2+1)(\kappa-\kappa^{-1}).$$

The determinant reduces to equations defining four values of $\kappa$ for a prescribed value of z. These equations may be written as

$$\kappa-\kappa^{-1} = \frac{2(z^2-1)}{b\lambda(z^2+1) \pm 2a\lambda z} . \qquad (4.6\pm)$$

Stability of the Cauchy problem requires that there exists no root $(\kappa,z)$ of $(4.6\pm)$ such that $|\kappa| = 1$ for $|z| > 1$. Substituting $\kappa = e^{i\theta}$ into $(4.6\pm)$ we obtain that $|\kappa| = 1 \Rightarrow |z| = 1$ and hence unconditional stability. For given z the two roots $\kappa$ of each equation in (4.6) are such that their product is -1. Therefore for $|z| > 1$ we denote by $\kappa_1$ and $\kappa_2$ the respective inner and outer roots of (4.6-) and by $\kappa_3$ and $\kappa_4$ the corresponding roots of (4.6+). Returning to the assumed solutions we find that the contribution to the general solution for

$$\underset{\sim}{U}_{2j} = (U_{2j},V_{2j-1},\hat{U}_{2j-1},\hat{V}_{2j})^T ,$$

associated with $\kappa_1$, is given by

$$(\alpha\kappa_1^{2j},\ \beta\kappa_1^{2j-1},\ \hat{\alpha}\kappa_1^{2j-1},\ \hat{\beta}\kappa_1^{2j})^T \qquad (4.7)$$

The vector $(\alpha,\beta,\hat{\alpha},\hat{\beta})^T$ is the eigenvector of the secular matrix associated with $\kappa_1$. To obtain the general solution for $\underset{\sim}{U}_{2j}$, in the grid function space $L_2(\Delta x)$, we linearly combine all the vectors of the form (4.7) defined by each of the inner roots $\pm\ \kappa_1$ and $\pm\ \kappa_3$. It may be shown that the associated eigenvectors are

$(1, \pm 1, 1, \pm 1)^T$ and $(1, \mp 1, -1, \pm 1)^T$, respectively.

This completes the general solution as

$$\begin{aligned}
U^{2n+1}_{2j} &= (d_1\kappa_1^{2j} + d_3\kappa_3^{2j})z^{2n+1}, && j = 0,1,2 \\
V^{2n+1}_{2j-1} &= (d_1\kappa_1^{2j-1} + d_3\kappa_3^{2j-1})z^{2n+1}, && j = 1,2,\ldots \\
\hat{U}^{2n+1}_{2j-1} &= (-d_1\kappa_1^{2j-1} + d_3\kappa_3^{2j-1})z^{2n+2}, && j = 1,2,\ldots, \\
\hat{V}^{2n+1}_{2j} &= (-d_1\kappa_1^{2j} + d_3\kappa_3^{2j})z^{2n+2}, && j = 0,1,2,\ldots,\ n \geq 1
\end{aligned} \tag{4.8}$$

with $d_1 = \alpha + \hat{\alpha}$ and $d_3 = \beta + \hat{\beta}$ .

We may now use the general solution (4.8) to examine the stability of any boundary approximation used in conjunction with the SILF method. To aid any such examination, information concerning the behaviour of $\kappa_1,\kappa_3$ and z through the characteristic equations ($4.6\pm$ ) is essential.

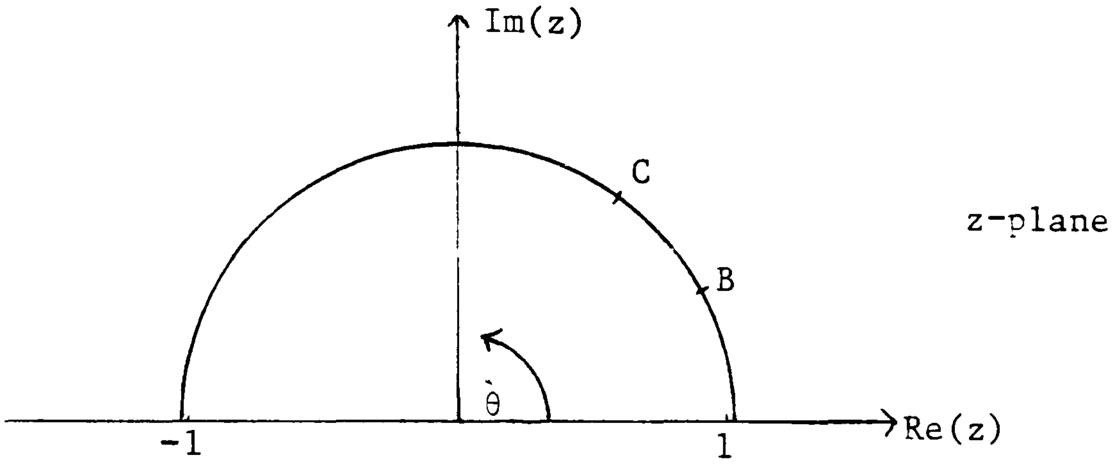
Lemma 4.1. If $\kappa, z \in C$ are related through (4.6) then $\kappa(\bar{z}) = \overline{\kappa(z)}$, thus implying that in any stability analysis we need only consider the half plane $\mathrm{Im}(z) \geq 0$.

Proof. Consider equation (4.6). By forming its conjugate as

$$\bar{\kappa}^2 - 1 = \frac{2(\bar{z}^2-1)\bar{\kappa}}{\lambda b(\bar{z}^2+1) \pm 2a\lambda z}$$

we obtain the result.

Since the SILF method is non-dissipative it is essential to determine the behaviour of the eigenvalue $\kappa_1$ as z is moved around the unit circle in accordance with the following diagram



The point B is given by $e^{i\theta_B}$ and C by $e^{i\theta_c}$, where

$$\frac{\sin\theta_B}{\cos\theta_B - (q-1)/(q+1)} = \lambda b,$$

and

$$\cos\theta_c = a/b.$$

The relationship between $\kappa_1,\kappa_2$ and z yields the following lemma.

Lemma 4.2. If $\kappa_1$ and $\kappa_2$ are the respective inner and outer roots of the characteristic equation

$$\kappa^2-1 = \frac{2(z^2-1)\kappa}{b\lambda(z^2+1) - 2a\lambda z}$$

then as $z = e^{i\theta}$ is moved around the unit circle $\kappa_1$ and $\kappa_2$ behave as follows:

(i) when $\theta = 0$ : $\kappa_1 = -1$, $\kappa_2 = 1$

(ii) $0 \to \theta \to \theta_B$ : $\kappa_1$ moves around $|\kappa| = 1$ to $\kappa_1 = i$, and $\kappa_2$ moves around $|\kappa| = 1$ to $\kappa_2 = i$ noting that $\arg(\kappa_2) = \pi - \arg(\kappa_1)$

(iii) $\theta_B \to \theta \to \theta_C$ : $\kappa_1$ moves down the imaginary axis to $\kappa = 0$. $\kappa_2$ moves up the imaginary axis towards $i\infty$,

(iv) $\theta_C \to \theta \to \frac{\pi}{2}$ : there are two cases to consider

(a) $0 < a\lambda \lesssim 1$ : $\kappa_1$ moves down the imaginary axis to $i\{\eta + \surd(\eta^2-1)\}$

$\kappa_2$ moves from $-i\infty$ up the imaginary axis to $i\{\eta - \surd(\eta^2-1)\}$. When $a\lambda = 1$ $\kappa_1$ and $\kappa_2$ coalesce at $-i$.

(b) $a\lambda > 1$ : having combined at $-i$ when $\sin\theta = \lambda(b\cos\theta - a)$ $\kappa_1$ then moves around the unit circle to $-\surd(1-\eta^2) + i\eta$ and $\kappa_2$ moves around to $\surd(1-\eta^2) + i\eta$ , where $\eta = -\frac{1}{a\lambda}$.

(v) $\frac{\pi}{2} \to \theta \to \pi$. :

(a) $a\lambda \lesssim 1$ : $\kappa_1$ and $\kappa_2$ move on the imaginary axis coalescing at $-i$ when $\sin\theta = -\lambda[b\cos\theta - a]$, noting that for $a\lambda = 1$ the merging occurs at $\theta = \frac{\pi}{2}$ . $\kappa_1$ and $\kappa_2$ then move around the unit circle to $-1$ and $+1$, respectively.

(b) $a\lambda > 1$ : $\kappa_1$ and $\kappa_2$ move, from their respective points on the unit circle in (iv) with $\theta = \frac{\pi}{2}$ , around the unit circle to $-1$ and $+1$, respectively.

In the above we have used the notation for $\alpha,\beta,\gamma \in \mathbb{R}^+$ that $\alpha \to \beta \to \gamma$ represents the monotonic increasing variation of $\beta$ between the fixed constants $\alpha$ and $\gamma$ $(\alpha < \gamma)$.

Proof: Part (1) follows immediately from a perturbation analysis. Therefore for $\theta \in [0,\theta_B]$ we have, with $\eta(\theta) = \frac{\sin\theta}{\lambda[b\cos\theta - a]}$ , that

$$\kappa_{1,2}(\theta) = i\eta(\theta) \mp \surd(1-\eta^2(\theta)) \qquad (4.9)$$

Clearly $|\kappa_{1,2}| = 1$ and $\eta(\theta_B) = 1$ thus implying that $\kappa_{1,2}(\theta_B) = i$, proving (ii). For $\theta \in [\theta_B, \theta_C]$ we have

$$\kappa_{1,2} = i\left[\eta(\theta) \mp \surd(\eta^2(\theta)-1)\right] \qquad (4.10)$$

and since $\eta(\theta)$ rises from unity to infinity part (iii) follows.

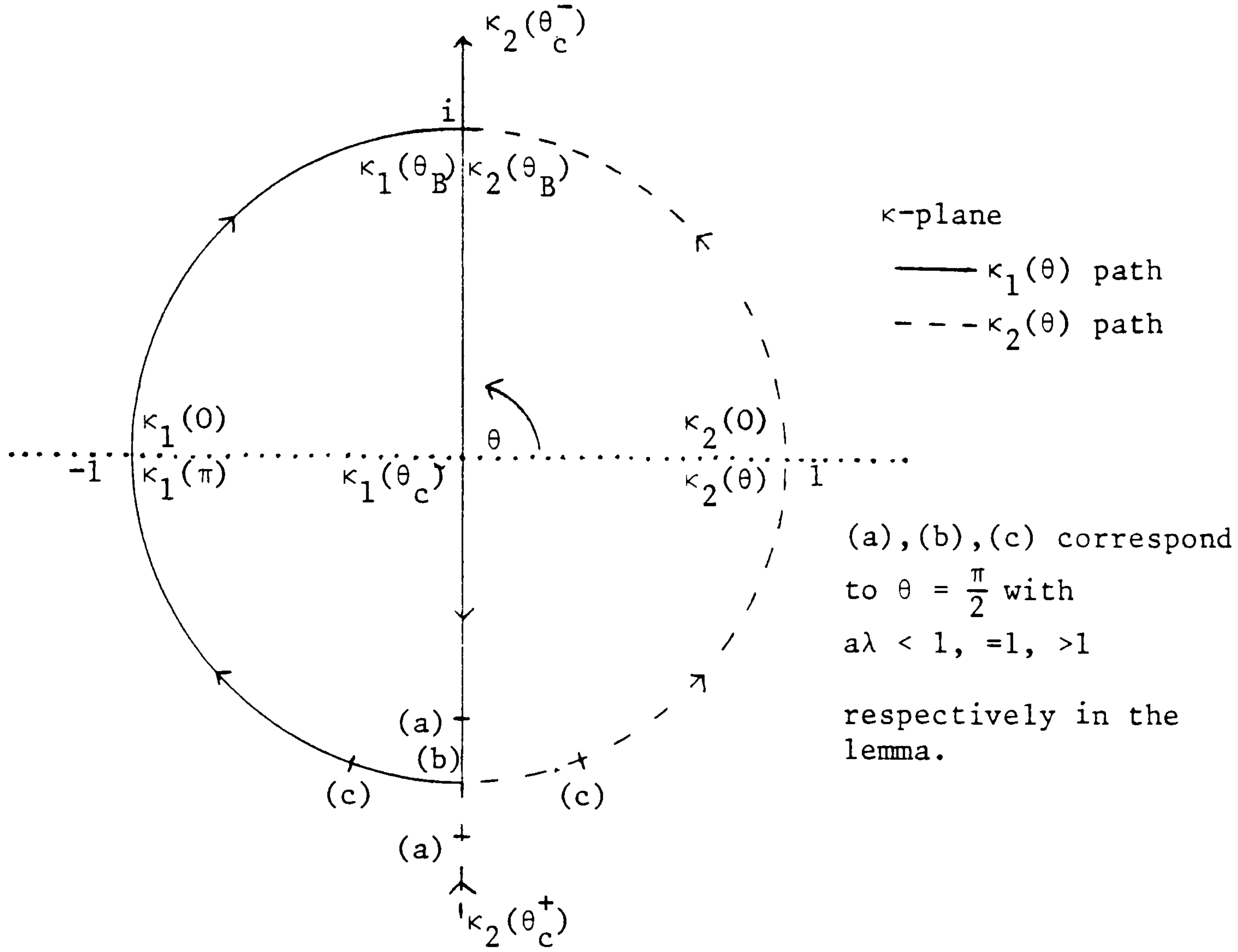
Now consider $a\lambda \leqslant 1$. When $\theta > \theta_C$, $\eta(\theta)$ is negative and so the signs in (4.10) are reversed. For $\theta_C \rightarrow \theta \rightarrow \frac{\pi}{2}$, $\eta(\theta)$ increases from $-\infty$ to $-\frac{1}{a\lambda}$, thus we have part(iv)(a). As $\theta$ increases monotically to $\pi$ $\kappa_1$ and $\kappa_2$ continue their respective downwards and upwards movements on the imaginary axis until they coalesce at $-i$. Notice that for $a\lambda = 1$ the coalescing occurs at $\theta = \frac{\pi}{2}$. This happens when $\sin\theta = -\lambda\left[b\cos\theta - a\right]$, that is when $\eta(\theta) = -1$. For all remaining $\theta$, the roots $\kappa_1$ and $\kappa_2$ diverge around the unit circle to $-1$ and $+1$ respectively at which point $\theta = \pi$.

Now consider $a\lambda > 1$. The roots are represented by

$$\kappa_{1,2} = i\{\eta(\theta) \pm \surd(\eta^2(\theta)-1) \ .$$

For $\theta_C \rightarrow \theta \rightarrow \pi$, $\eta(\theta)$ increases from $-\infty$, and so $\kappa_1$ moves down and $\kappa_2$ up the imaginary axis until they coalesce at $-i$ when $\eta(\theta) = 1$, at which point $\theta < \frac{\pi}{2}$. For all remaining $\theta$ they diverge as before to $-1$ and $+1$ respectively.

This lemma may be usefully expressed pictorially as



(a),(b),(c) correspond to $\theta = \frac{\pi}{2}$ with $a\lambda < 1$, $=1$, $>1$ respectively in the lemma.

Lemma 4.3. The respective inner and outer roots $\kappa_3$ and $\kappa_4$ of the characteristic equation

$$\kappa^2 - 1 = \frac{2(z^2-1)\kappa}{\lambda\left[b(z^2+1)+2az\right]} \tag{4.11}$$

are given by

$$\kappa_{3,4}(\theta) = \kappa_{1,2}(\pi+\theta)$$

when $z = e^{i\theta}$.

Proof. As seen earlier

$$\kappa_{1,2}(\theta) = i\eta \mp \sqrt{(1-\eta^2)},$$

with $$\eta(\theta) = \sin\theta/\left[\lambda(b\cos\theta - a)\right].$$

Now $\eta(\pi+\theta) = \sin\theta/[\lambda(b\cos\theta + a)] = \eta_+(\theta)$

therefore $\kappa_{1,2}(\theta+\pi) = i\eta_+ \mp \sqrt{(1-\eta_+^2)} = \kappa_{3,4}(\theta)$. All other expressions in Lemma 4.2 relate similarly.

Lemmas 4.1, 4.2 and 4.3 together provide a complete insight into the behaviour of $\kappa(z)$ with $z = e^{i\theta}$.

## 4.2 Description and Stability Analysis of Some Boundary Approximations.

The boundary approximations to be used in conjunction with the SILF interior difference method are, at the left boundary,

Horizontal Extrapolation (HE): $\hat{V}_o^{2n+1} = \hat{V}_2^{2n+1}$ . (4.12)

S=0 Space Time Extrapolation ($ST_o$): $\hat{V}_o^{2n+1} = V_1^{2n+1}$ . (4.13)

S=1 Space Time Extrapolation ($ST_1$): $\hat{V}_o^{2n+1} = 2V_1^{2n+1} - \hat{V}_2^{2n-1}$.(4.14)

Zeroth order characteristic Space-Time Extrapolation (CST):

$$\hat{V}_o^{2n+1} = (U+V)_1^{2n+1} - \hat{U}_o^{2n+1} .$$

The staggered mesh requires that space averaging be used on $U_1^{2n+1}$, so

$$\hat{V}_o^{2n+1} = \tfrac{1}{2}(U_o^{2n+1}+U_2^{2n+1}) + V_1^{2n+1} - \hat{U}_o^{2n+1} . \quad (4.15)$$

Linear Characteristics (LC):

$$(U\hat{+}V)_o^{2n+1} = (1-\lambda q)(U+V)_o^{2n+1} + \lambda q(U+V)_1^{2n+1}$$

using space and time averaging respectively on $U_1^{2n+1}$,$V_o^{2n+1}$.

We obtain

$$\tfrac{1}{2}(1+\lambda q)\hat{V}_o^{2n+1} = (1-\tfrac{1}{2}\lambda q)U_o^{2n+1}-\hat{U}_o^{2n+1}+\tfrac{1}{2}(1-\lambda q)\hat{V}_o^{2n-1}+\lambda q(V_1^{2n+1}+\tfrac{1}{2}U_2^{2n+1}). \quad (4.16)$$

Boundary approximation of Gustafsson et al [1972] (K):

$$\hat{V}_o^{2n+1} = \hat{V}_o^{2n-1} + 2a\lambda\{V_1^{2n+1} - \tfrac{1}{2}(\hat{V}_o^{2n+1} + \hat{V}_o^{2n-1})\}$$

$$+ 2b\lambda\{\tfrac{1}{2}(\hat{U}_1^{2n+1} + \hat{U}_1^{2n-1}) - U_o^{2n+1}\}, \qquad (4.17)$$

which is obtained by differencing the second differential equation and allowing for the staggered mesh.

We also include the crude

Box (B): $$\hat{V}_o^{2n+1} = \frac{b}{2}\lambda(\hat{U}_1^{2n+1} - U_o^{2n+1}) + V_1^{2n+1} \ . \qquad (4.18)$$

The box approximation is obtained from the expression

$$(1+a\lambda)\hat{V}_o^{2n+1} + (1-a\lambda)\hat{V}_1^{2n+1} = (1-a\lambda)V_o^{2n+1} + (1+a\lambda)V_1^{2n+1}$$

$$+ \lambda b\left[\hat{U}_1^{2n+1} - \hat{U}_o^{2n+1} + U_1^{2n+1} - U_o^{2n+1}\right]$$

by ignoring all quantities that are not defined on the grid and setting $1 + a\lambda \equiv 2$. The fact that this approximation excels for very large values of q illustrates the possible success of boundary approximations constructed in an arbitrary manner. Such an approach is not recommended as it is felt that in realistic differential applications problems in ensuring the consistency and stability of the approximation will inevitably result.

Consider the HE approximation. Substitution of the general solution (4.8) into (4.12) yields

$$-(\kappa_1^2 - 1)d_1 + (\kappa_3^2 - 1)d_3 = 0 \qquad (4.19)$$

and, from the analytic boundary condition we obtain

$$d_1 + d_3 = g((2n+1)\Delta t)/z_{n+1} \ . \qquad (4.20)$$

Equations (4.19) and (4.20) together form the system

$$M\underset{\sim}{d} = \underset{\sim}{g}. \tag{4.21}$$

According to the stability theory we require that there be no non-trivial solutions d of the homogeneous form of (4.21) for $|\kappa_1| < 1$. $|\kappa_3| < 1$ when $|z| > 1$ (or no generalised eigenvalues). Equivalently we want no such roots of det M = 0. This condition is that required in Theorem 1.6 of Chapter 1 to ensure stability. Hence the determinant equation for the HE approximation is

$$\kappa_1^2 + \kappa_3^2 - 2 = 0 \tag{4.22}$$

<u>Lemma 4.4.</u> The SILF method with the HE boundary approximation is an unstable approximation to the left boundary problem.

Proof. Equality in (4.22) requires that $\kappa_1^2 = \kappa_3^2 = 1$ and thus, from the characteristic equations, $z^2 = 1$. Instability follows from Lemmas (4.2) and (4.3).

This result holds for any degree of horizontal extrapolation and is an extension of the result of Gustafsson et al [1972] for the explicit Leap-Frog method (4.2).

Following the previous discussion we can construct the determinant equation (4.22) for each of the boundary approximations considered. These are given in Table 4.1.

| Boundary Approx | Determinant Equation |
|---|---|
| $ST_j$ | $(z+\kappa_1)^{j+1} + (z-\kappa_3)^{j+1} = 0$ |
| CST | $4z + (\kappa_1+1)^2 - (\kappa_3+1)^2 = 0$ |
| LC | $(1+\lambda q)z^2+\lambda qz(\kappa_1-\kappa_3)\left[1+\frac{1}{2}(\kappa_1+\kappa_3)\right]-(1-\lambda q) = 0$ |
| K | $2(z^2-1)+2a\lambda z(\kappa_1-\kappa_3)+2a\lambda(z^2+1)-\lambda b(z^2+1)(\kappa_1+\kappa_3) = 0$ |
| B | $4z - \lambda bz(\kappa_1+\kappa_3) + 2(\kappa_1-\kappa_3) = 0$ |

Table 4.1: Boundary approximation determinant equations.

When $q = 1$ we have

Lemma 4.5. The approximation to the symmetric problem given by the SILF interior method is stable for the $ST_0$,$ST_1$,CST,LC,K or B boundary approximations.

Proof. From (4.6$\pm$) we have that, for $q = 1$, both $\kappa_1$ and $\kappa_3$ are equivalent to $\kappa$, say. For each boundary approximation we can construct a system of two equations in $z$ and $\kappa$, from the characteristic and boundary determinant equations. Consider the $ST_1$ approximation. The associated determinant equation reduces to $z = \pm i\kappa$ and so we must consider the possible existence of a generalised eigenvalue. From Lemma 4.2 we can see that there exists no $\theta$ such that $z(\theta) = i\kappa(\theta)$ thus establishing stability. The other proofs of stability follow analogously.

The above results were relatively easy to obtain. For general q, however, the stability problem reverts to the evaluation of the roots of a multivariate system formed by the characteristic equations (4.6$\pm$)

and the appropriate equation of Table 4.1. To obtain the roots we must resort to the continuation method of Chapter 2. Using these techniques we can establish whether or not a boundary approximation will produce an unstable approximation to the left boundary problem by being associated with a solution triple $(\kappa_1,\kappa_2,z)$ for which z is an eigenvalue or generalised eigenvalue. Unfortunately any conclusions will hold only for a discrete set of $\lambda$ and q.

The results may be summarized as

Result 4.1. For the discrete set $\{(\lambda,q) : \lambda = 0.5, q = 2, 5, 30\}$ the SILF method and any one of the $ST_0$, $ST_1$, CST, K or B boundary approximations produced a stable approximation to the left boundary problem if q = 2 or 30 and an unstable approximation for q = 5.

For q = 2 and q = 30 Result 4.1 reflects a wide series of continuation experiments which yielded no solution triples to indicate instability. When q = 5 we have that $a\lambda = 1$ and it is clear that the solution triple (-i,i,i) satisfies the determinant equations associated with all but the LC approximation. From Lemmas 4.2 and 4.3 this triple corresponds to a generalised eigenvalue, thus indicating instability. With the LC approximation we found, using the continuation algorithm, the roots z = 1.091, and $\kappa_1 = \bar{\kappa}_3 = -0.258-0.659i$, again indicating instability. Formulating (4.3) as a one-step method it is possible to determine the spectral radius of the coefficient matrix involving each boundary approximation. For $q > 1$ $(q \neq 5)$ the $ST_1$ method induced exponentially growing solutions allowed by the theory of Chapter 1 (see (1.32)). This type of phenomenon is not allowed by the P-stability theory of Warming, Beam and Yee [1982] and is also evident in Chapter 5. This topic is not investigated any further here. Over the chosen data set ($q \neq 5$) for a discrete choice of $\Delta x$, each of the $ST_0$, CST and LC approximations produced non-dissipative, and B a dissipative overall approximation to the left boundary problem where, for the latter boundary technique, the degree of dissipation increased with q. The Kreiss approximation was non-dissipative for $q = 1$ and dissipative for $q = 2$ and $q = 30$, the dissipation again increasing with q. Clearly the data set $\{(\lambda,q)\}$ chosen does not provide any definitive comment regarding the overall stability induced by any boundary approximation. To examine an exhaustive set of $\lambda$ and q is beyond the scope of this study, however we have illustrated that for any fixed choice of parameters, we have developed an algorithm which can be used to examine the stability properties of any approximation.

As has been emphasized earlier it is essential to consider separately the potentially less stable right boundary. For an even number of mesh intervals the right boundary problem is associated with the previous left boundary problem treated earlier, where the inward and outward characteristic speeds have been interchanged. The associated characteristic equations are

$$\kappa^2-1 = \frac{2(1-z^2)\kappa}{\lambda[b(z^2+1) \pm 2az]} . \qquad (4.24\mp)$$

Deriving the general solution as in (4.8), we require the inner roots $\kappa_{\alpha_1}, \kappa_{\alpha_2}$ of $(4.24\mp)$. If $\kappa_{\alpha_1}$ satisfies (4.24-) then it is clear that $\kappa_{\alpha_1}^{-1}$ is equivalent to $\kappa_2$, and similarly $\kappa_{\alpha_2}^{-1}$ is equivalent to $\kappa_4$.

<u>Lemma 4.6.</u> $\kappa_2^{-1}(\lambda,q,z) = -\kappa_1(\lambda,q,z)$ ; $\kappa_4^{-1}(\lambda,q,z) = -\kappa_3(\lambda,q,z)$.

Proof. $\kappa_2^{-1}(\lambda,q,z)$ satisfies (4.24-). Since the interior root of

(4.24-) is $\kappa_1(-\lambda,q,z)$ then $\kappa_2^{-1}(\lambda,q,z) = \kappa_1(-\lambda,q,z) = -\kappa_1(\lambda,q,z)$.

Therefore we can derive all information concerning the roots $\kappa_1$ and $\kappa_3$ in the general solution

$$\begin{aligned}
U_{J-2j}^{2n+1} &= z^{2n+1}\,(c_1\kappa_1^{2j} + c_3\kappa_3^{2j}), && j = 0,1,2,\dots \\
V_{J-2j+1}^{2n+1} &= z^{2n+1}\,(-c_1\kappa_1^{2j-1} - c_3\kappa_3^{2j-1}), && j = 1,2,\dots \\
\hat{U}_{J-2j+1}^{2n+1} &= z^{2n+2}\,(c_1\kappa_1^{2j-1} - c_3\kappa_3^{2j-1}), && j = 1,2,\dots \\
\hat{V}_{J-2j}^{2n+1} &= z^{2n+2}\,(-c_1\kappa_1^{2j} + c_3\kappa_3^{2j}). && j = 0,1,2,\dots;n \geqslant 1.
\end{aligned} \qquad (4.26)$$

$c_1$, $c_3 \in \mathbb{R}$, from lemmas 4.1, 4.2 and 4.3. General solution (4.26) is obtained in a manner analogous to that of (4.8) for the roots $\kappa_2^{-1}$ and $\kappa_4^{-1}$ and then by applying Lemma 4.6.

The right boundary approximations are

$$ST_o \ : \ \hat{V}_J^{2n+1} = V_{J-1}^{2n+1} \qquad (4.27)$$

$$ST_1 \ : \ \hat{V}_J^{2n+1} = 2V_{J-1}^{2n+1} - \hat{V}_{J-2}^{2n-1} \qquad (4.28)$$

$$CST \ : \ \hat{V}_J^{2n+1} = \hat{U}_J^{2n+1} - \tfrac{1}{2}(U_J^{2n+1}+U_{J-2}^{2n+1}) + V_{J-1}^{2n+1} \qquad (4.29)$$

$$LC \ : \ \tfrac{1}{2}(1+\lambda)\hat{V}_J^{2n+1} = (\tfrac{\lambda}{2}-1)U_J^{2n+1}+\hat{U}_J^{2n+1}+\tfrac{1}{2}(1-\lambda)\hat{V}_J^{2n-1}+\lambda(V_{J-1}^{2n+1}-\tfrac{1}{2}U_{J-2}^{2n+1}) \qquad (4.30)$$

$$\begin{aligned} K \ : \ \hat{V}_J^{2n+1} = \hat{V}_J^{2n-1} &+ 2a\lambda\left[\tfrac{1}{2}(\hat{V}_J^{2n+1}+\hat{V}_J^{2n-1}) - V_{J-1}^{2n+1}\right] \\ &+ 2b\lambda\left[U_J^{2n+1}-\tfrac{1}{2}(\hat{U}_{J-1}^{2n+1}+\hat{U}_{J-1}^{2n-1})\right] \end{aligned} \qquad (4.31), \text{ and}$$

$$B \ : \ \hat{V}_J^{2n+1} = V_{J-1}^{2n+1} + \frac{b\lambda}{2}(U_J^{2n+1}-\hat{U}_{J-1}^{2n+1}) \ . \qquad (4.32)$$

For the above boundary approximations, with q > 1, no instability triples were found. The triple (-i,i,i) which produced instability conclusions for the left boundary problem fails to satisfy the right boundary determinantal equations for a value of q such that $a\lambda = 1$ (note: $\lambda = 0.5$). An investigation of the spectral radius of each right boundary approximation revealed that Box was a dissipative approximation, $ST_0$, CST and LC were non-dissipative and $ST_1$ and K had eigenvalues exceeding unity. For both the latter boundary methods the spectral radius reduced with $\lambda$ and so the method was exhibiting exponentially growing solutions (1.32). When we examine the twin boundary problem the K boundary scheme produces a non-dissipative approximation. Therefore the left boundary dissipation is, in some sense, negating the growth introduced at the right. This behaviour is also evident in Chapter 5. The twin $ST_1$ approximation still has a spectral radius exceeding unity that decreases with $\lambda$.

## 4.3 Numerical Results.

We illustrate the results of the previous section by integrating numerically the problem defined by (3.28). We consider $\varepsilon = 0.01$ and $\varepsilon = 0.5$ and $q = 1,2,5$ and 30. The extreme value of $q = 30$ was chosen to illustrate the exceptional accuracy of the B boundary approximation when q is large. The results are contained in Tables 4.2 and 4.3. The initial E norm values at $q = 30$ are 1184 and 1194 for $\varepsilon = 0.01$ and $\varepsilon = 0.5$, respectively.

| Boundary Approximations | q = 1 | q = 2 | | q = 5 | | q = 30 | |
|---|---|---|---|---|---|---|---|
| | | Left | Right | Left | Right | Left | Right |
| Exact | 38.931 \| 0.213 | 77.595 \| 0.133 | | 197.74 \| 0.097 | | 1159 \| 0.431 | |
| $ST_o$ | 41.126 \| 2.674 | 81.836 \| 0.821 | 77.205 \| 0.546 | 204.75 \| 1.025 | 196.43 \| 0.397 | 1091 \| 1.016 | 1.3(4) \| 42.8 |
| $ST_1$ | 40.331 \| 0.461 | 78.020 \| 0.135 | 79.660 \| 4.484 | 195.76 \| 0.221 | 194.74 \| 0.179 | 3.9(5) \| 196.97 | 3.9(4) \| 142.4 |
| CST | 39.397 \| 0.438 | 77.836 \| 0.132 | 76.370 \| 1.173 | 194.69 \| 0.098 | 197.52 \| 0.759 | 1096.9 \| 0.597 | 67151 \| 102.9 |
| LC | 39.335 \| 0.439 | 77.836 \| 0.132 | 78.515 \| 0.222 | 238.93 \| 30.158 | 194.87 \| 0.114 | 10364 \| 20.86 | 3070 \| 16.75 |
| K | 38.969 \| 0.102 | 77.597 \| 0.137 | 78.447 \| 0.413 | 194.81 \| 0.098 | 194.90 \| 0.255 | 1429 \| 3.186 | 9(21) \| 9(10) |
| B | 39.738 \| 1.765 | 80.182 \| 0.536 | 76.594 \| 0.723 | 199.15 \| 0.458 | 197.47 \| 0.577 | 1178 \| 0.468 | 1157 \| 0.474 |

Table 4.2: $\|\underset{\sim}{U}_\Delta\|_{E_{q,0.01}} \mid \|\underset{\sim}{U}_\Delta\|$ for $\lambda = 0.5$, $\Delta x = \frac{1}{80}$, $\varepsilon = 0.01$.

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | | q = 30 | |
|---|---|---|---|---|---|---|---|
| | | Left | Right | Left | Right | Left | Right |
| Exact | 48.767 \| 0.212 | 88.615 \| 0.177 | | 212.99 \| 0.464 | | 847.79 \| 3.701 | |
| $ST_o$ | 54.482 \| 2.696 | 90.335 \| 0.825 | 94.721 \| 1.648 | 207.21 \| 0.862 | 244.74 \| 3.391 | 16143 \| 49.39 | 2.6(5) \| 158.58 |
| $ST_1$ | 51.998 \| 0.458 | 87.317 \| 0.150 | 94.804 \| 6.424 | 204.51 \| 0.280 | 220.86 \| 0.863 | 4(10) \| 7(4) | 1.7(7) \| 3.3(3) |
| CST | 45.528 \| 0.624 | 87.199 \| 0.146 | 86.532 \| 1.204 | 186.02 \| 0.826 | 217.56 \| 1.202 | 10845 \| 39.082 | 1.7(5) \| 150.4 |
| LC | 51.113 \| 0.509 | 87.199 \| 0.146 | 90.860 \| 0.344 | 214.46 \| 14.718 | 213.28 \| 0.719 | 1.4(7) \| 1035 | 19298 \| 67.36 |
| K | 49.261 \| 0.118 | 86.954 \| 0.157 | 88.945 \| 0.232 | 204.97 \| 0.238 | 201.73 \| 1.572 | 2.5(5) \| 166.77 | 1.7(24) \| 2.5(12) |
| B | 51.282 \| 1.792 | 90.581 \| 0.550 | 91.641 \| 1.104 | 213.87 \| 0.571 | 234.74 \| 2.233 | 2067 \| 8.106 | 1478 \| 11.87 |

Table 4.3: $\|\underset{\sim}{U}_\Delta\|_{E_{q,0.5}} \mid \|\underset{\sim}{U}_\Delta\|_\infty$ for $\lambda = 0.5$, $\Delta x = \frac{1}{80}$, $\varepsilon = 0.5$.

The poor performance of the interior method is due to some form of error oscillation having the same period as the fast q-wave. For problem (3.28) with the non-oscillating general solution

$$u(x,t) = x - t + 5\ \ell n(x+qt+1),\ v(x,t) = x - t - 5\ \ell n(x+qt+1)$$

the maximum norms corresponding to a SILF integration with exact boundary values and $\lambda = 0.5$ are, for $\Delta x = 1/40$ and $\Delta x = 1/80$, 0.00044 and 0.00011, respectively. This illustrates the acceptable accuracy and second order rate of convergence. The presence of the above error oscillation in the SILF approximation to the original problem warrants further investigation. No oscillation was observed with the methods of Chapters 3 and 5. Despite the poor accuracy it is still possible to compare the relative performance of each boundary approximation. The improved accuracy of the K approximation over the exact boundary data is caused by the overall approximation being dissipative, thus damping the error oscillation. This behaviour is especially evident when there is a strong fast wave. The instability at $q = 5$ for the left LC approximation is evident (for $\Delta x = 1/40$, $\varepsilon = 0.01$ the maximum norm is 1.935).

The loss of accuracy when the CFL number, $\lambda q$, exceeds unity is clear when the fast wave is strengthened to 50%. Therefore whilst the SILF method is stable for any choice of $\lambda$ and q, accuracy considerations become the dominant criteria for large $\lambda q$. In other words it is not sufficient to merely ensure unconditional stability for even consistent approximations.

## 4.4 Sköllermo Accuracy

We repeat the method of Section [3.4] to attempt order, in terms of accuracy, the boundary approximations of this chapter. In this case the Fourier transformed error functions have the form

$$\begin{aligned} \hat{\varepsilon}(x_j,\omega) &= (\kappa_\beta^j-\kappa_\alpha^j)\phi(\omega) \\ \hat{\theta}(x_j,\omega) &= (\kappa_\beta^j+\kappa_\alpha^j)\phi(\omega) \quad , \end{aligned} \tag{4.33}$$

where a full grid is used, $\hat{\varepsilon}(x_o,\omega) = 0$ and $\kappa_\alpha$ and $\kappa_\beta$ are the inner roots of the characteristic equations

$$\kappa^2-1 = \frac{(z^2-1)\kappa}{a\lambda z \mp (z^2+1)b\lambda} \quad . \tag{4.34}$$

As the grid size $\Delta x$ tends to zero z, $\kappa_\alpha$ and $\kappa_\beta$ approach 1, -1 and -1 respectively. We can then construct the second error function $e(\omega)$, of Sköllermo, for homogeneous boundary data, as in Table 4.4. The right boundary is transformed to a left boundary problem. As in Chapter 3 we determine the minimum number of points, M, per wavelength to achieve an error tolerance of 1%. These results are given, for $\lambda = 0.5$, in Table 4.5. For the K and B approximations at q = 1, the functions $e(\omega)$ are $\frac{4}{3}(\frac{\pi}{M})^3\ (1+2\lambda^2)$ and $\frac{4\pi^2}{(2+\lambda)M^2}$, respectively. When $\lambda = 0.5$ and $q = 2$ the expression associated with the CST approximation at the left boundary is $(\frac{\pi}{M})^2\ (2\lambda^2 + 1 - \frac{3}{q^2})$.

| Boundary Approximation | Left | Right |
|---|---|---|
| $ST_o$ | $\frac{\pi}{M}(2\lambda+1 - \frac{1}{q})$ | $\frac{\pi}{M}(2\lambda-1 + \frac{1}{q})$ |
| $ST_1$ | $\frac{\pi^2}{M^2}\left[1+2\lambda^2 + \frac{1}{q^2} - 2\lambda(\frac{1}{q} - 1)\right]$ | $\frac{\pi^2}{2}\left[2\lambda^2 + \frac{1}{q^2} + 1 + 2\lambda(\frac{1}{q} - 1)\right]$ |
| CST | $\frac{2\pi}{M}(\lambda - \frac{1}{q})$ | $\frac{2\pi}{M}(1 - \lambda)$ |
| LC | $\frac{\pi^2}{q^2M^2}\left[2\lambda^2q^2 + q^2-3\right]$ | $\frac{\pi^2}{q^2M^2}\left[2\lambda^2q^2+1 - 3q^2\right]$ |
| B | $\frac{2\pi}{Mq}\left[2\lambda q-b^2\lambda+2a\right]/(2+b\lambda)$ | $\frac{2\pi}{Mq}\left[2\lambda q-b^2\lambda-2a\right]/(2+b\lambda)$ |
| K | $\frac{\pi^2}{q^2M^2}\left[2a\lambda^2q + q-1\right]$ | $\frac{\pi^2}{qM^2}\left[1 - q - 2a\lambda^2q\right]$ |

Table 4.4: $e(\omega)$ for the SILF interior approximation.

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | | q = 30 | |
|---|---|---|---|---|---|---|---|
| | | Left | Right | Left | Right | Left | Right |
| $ST_o$ | 315 | 472 | 157 | - | 63 | 618 | 11 |
| $ST_1$ | 45 | 48 | 36 | - | 28 | 50 | 8 |
| CST | 315 | 28 | 315 | - | 315 | 294 | 315 |
| LC | 39 | 28 | 48 | - | 50 | 39 | 50 |
| B | 140 | 215 | 15 | - | 126 | 132 | 256 |
| K | 19 | 39 | 28 | - | 43 | 17 | 90 |

Table 4.5: Minimum value of M for $\lambda$ = 0.5, and a tolerance 0.01.

The method of Sköllermo predicts that the $ST_1$, LC and K approximations should be the most accurate. To provide an additional check on this prediction we repeat the results of Table 4.2 for the differential problem (3.57), in Table 4.6.

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | | q = 30 | |
|---|---|---|---|---|---|---|---|
| | | Left | Right | Left | Right | Left | Right |
| Exact | 0.140 | 0.807 | - | 1.558 | - | 11.408 | - |
| $ST_o$ | 2.468 | 1.229 | 3.583 | 1.109 | 7.439 | 23.805 | 65.670 |
| $ST_1$ | 1.220 | 0.488 | 6.919 | 1.005 | 17.760 | 20861 | 5906 |
| CST | 0.457 | 0.520 | 1.687 | 1.500 | 2.257 | 25.796 | 86.617 |
| LC | 1.084 | 0.520 | 1.127 | 5.357 | 1.744 | 7773 | 39.987 |
| B | 2.676 | 0.378 | 1.844 | 0.861 | 4.028 | 10.174 | 14.344 |
| K | 0.203 | 0.425 | 0.729 | 0.438 | 4.020 | 474.78 | 1(12) |

Table 4.6: $||\underset{\sim}{U}_\Delta||_\infty$ with $\Delta x = \frac{1}{40}$, $\lambda = 0.5$, problem (3.57)

According to the theory of Sköllermo the results of Table 4.6 should be in broad agreement with those of Table 4.5, and should be expecially so for q = 1. However the observed accuracy of the CST approximation is much greater than predicted although the superiority of the approximation suggested by Kreiss is substantiated. Overall, the theory of the section provides unsatisfactory ranking predictions. The discrepancies may have occurred as a result of the Sköllermo analysis being applied on a full grid. This is felt to be unlikely and may be more related to the error oscillation induced by the analytic solution on the numerical approximation. The error results of Table 4.6 do, however, serve to reinforce the conclusions of

Chapter 3 that any boundary approximation must incorporate the characteristic variables or physical equations. We were only able to induce instability, for $a\lambda = 1$, with the LC approximation; the other boundary schemes exhibiting more 'robust' instabilities. The very small data set covered in this chapter reflects the computational cost required to establish the stability of approximations that are based on three or more time levels.

In this chapter we have modified the popular Leap-Frog method to obtain a semi-implicit scheme that is unconditionally stable in approximating the solution of the pure Cauchy problem. This method was, for the test problem chosen, beset by a significant error oscillation and did not provide a viable alternative to either of the other interior schemes studied in this thesis. General 'semi-implicit' methods are frequently used by many practitioners, especially those working in meteorology, and this popularity was substantiated by the example of Section 4.3 where, for a non-oscillating analytic solution, an acceptably accurate numerical solution was obtained.

Despite the reduced accuracy of the interior method (for the test problem chosen) we were able to illustrate some of the stability results concerning the choice of boundary approximations. Note that as a result of the staggered mesh used we only required boundary approximations at every other time level. Implementation of the boundary approximations was complicated by the use of this staggered mesh however the successful techniques were those that were based on the physical properties of the differential system.

The SILF method is straightforward to apply and may be enhanced by basing the approximation on a staggered grid. This results in a significant reduction in computational cost. The economy attainable is illustrated in Chapter 6.

# CHAPTER 5

# A FINITE ELEMENT METHOD AND BOUNDARY APPROXIMATIONS

## 5.1 The Advection Equation

In this chapter we develop a new finite element method suitable for systems with differing time scales. The test problem may be written in diagonalised characteristic form so, for simplicity, we first derive a finite element method for the scalar advection equation

$$u_t(x,t) = qu_x(x,t), \quad 0 \leqslant x \leqslant 1, \; t \geqslant 0, \; q \in \mathbb{R}. \qquad (5.1)$$

Any method developed to approximate the solution of equation (5.1) can readily be generalised to the system $\underset{\sim}{u}_t = A\underset{\sim}{u}_x$ (Morton and Parrot [1980]). Therefore we consider equation (5.1) with appropriate initial data u(x,0) and with a prescribed boundary condition that ensures well-posedness.

To obtain a method which is q-dependent and which maintains the improvements in accuracy over a standard Galerkin scheme, we consider a Petrov-Galerkin approximation to the solution of (5.1). The approach adopted will be related to that developed by Morton and Parrot [1980]. We therefore seek an approximation to the solution of the weak form of (5.1)

$$\langle u_t - qu_x, \psi \rangle = 0 \qquad (5.2)$$

where $\langle .,. \rangle$ is the usual $L_2$ inner product and $\psi(x)$ is defined over the domain of the problem. The Petrov-Galerkin approximation to the weak solution u(x,t) is

$$U(x,t) = \sum_j U_j(t)\phi_j(x), \quad \forall \phi_j(x) \in S^{\Delta x} \qquad (5.3)$$

where $U \in S^{\Delta x}$, some trial space, and where the trial functions $\phi_j$ form a basis for $S^{\Delta x}$. The standard Galerkin approximation is defined by the special case $\psi \equiv \phi$. In the discretized problem we shall divide the region [0,1] into J elements of equal width $\Delta x$, and the functions $\phi_j$ and $\psi_j$ will generally be simple polynomials defined in the region of the nodal point $x_j = j\Delta x$.

Petrov-Galerkin approximations have been studied by Duncan [1982] and Morton and Parrot [1980] for the advection equation. Sanz-Serna and Christie [1981] and Alexander and Morris [1979] considered the non-linear dispersive waves of the Korteweg-de Vries equation demonstrating the increased accuracy of a Petrov-Galerkin approximation. For elliptic equations, test and trial function pairings for particular classes of problems have been suggested by Anderssen and Mitchell [1979] where the polynomial functions $\psi(x)$ and $\phi(x)$ vary from piece-wise linears to Hermite cubics. Whilst it is true that the higher the degree of $\psi$, or $\phi$, then the better will be the resulting approximation, the computational cost and complexity increases accordingly. In this chapter the trial function $\phi_j$ will be defined as a translation of a function $\phi$ and the test function $\psi_j$ will be defined as a translation of a function $\psi$. To minimise the computational effort further, reduce the number of additional boundary approximations and to simplify the generalisation to more complicated problems we constrain $\psi(x)$ to have compact support over two elements and the function $\phi(x)$ is the familiar 'hat' function

$$\phi(s) = \begin{cases} 1 - |s| & : \ |s| \le 1 \\ 0 & : \ \text{else.} \end{cases}$$

$\phi_j(x)$ and $\psi_j(x)$ are defined in terms of $\phi(x)$ and $\psi(x)$, respectively by $\phi_j(x) = \phi(\frac{x}{\Delta x} - j)$, $\psi_j(x) = \psi(\frac{x}{\Delta x} - j)$ and the functions $\phi$ and $\psi$ are normalised so that

$$\int_{-\infty}^{\infty} \psi(x)dx = \int_{-1}^{1} \psi(s)ds = \int_{-1}^{1} \phi(s)ds = 1. \qquad (5.4)$$

The functions $U_j(t)$ in (5.3) are given by approximating $u(x,t)$ in (5.2) by (5.3). We require , for any i, the inner products

$$<\phi_i,\psi_i> = \Delta x \{1 - \int_{-1}^{1} \text{sign}(s)s.\psi(s)ds\},$$

$$<\phi_{i+1},\psi_i> = \Delta x \int_{0}^{1} s\psi(s)ds,$$

$$<\phi_{i-1},\psi_i> = -\Delta x \int_{-1}^{0} s\psi(s)ds,$$

$$<\phi'_{i-1},\psi_i> = -\int_{-1}^{0} \psi(s)ds,$$

$$<\phi'_i,\psi_i> = 1 - 2\int_{0}^{1} \psi(s)ds, \quad \text{and}$$

$$<\phi'_{i+1},\psi_i> = \int_{0}^{1} \psi(s)ds,$$ where the dash denotes differentiation with respect to the argument.

Using the notation

$$A = \int_{0}^{1} \psi(s)ds, \quad B = \int_{0}^{1} s\psi(s)ds, \quad C = -\int_{-1}^{0} s\psi(s)ds, \qquad (5.5)$$

the above inner products and (5.2) yield the following differential-difference approximation to the Cauchy problem (5.1)

$$C\dot{U}_{j-1} + (1-B-C)\dot{U}_j + B\dot{U}_{j+1} = \frac{q}{\Delta x}\{(A-1)U_{j-1} + (1-2A)U_j + AU_{j+1}\}. \quad (5.6)$$

We have abused our notation by using the letter A. This integral quantity is only of a local relevance and no confusion with the matrix in the test problems should occur. Construction of finite element methods by using arbitrary constants like (5.5) has also been done by Duncan [1982] and Mitchell, Griffiths and Pen-Yu [1982].

Approximations of the derivatives in (5.6) by a stable time-stepping rule (Vichnevetsky and Bowles [1982]) will result in a difference scheme involving the constants A, B and C. Here application of the trapezoidal rule yields

$$\begin{aligned}&\left[C + \tfrac{1}{2}\lambda q(1-A)\right]U_{j-1}^{n+1} + \left[1-B-C + \tfrac{1}{2}\lambda q(2A-1)\right]U_{j}^{n+1} + \left[B-\tfrac{1}{2}\lambda qA\right]U_{j+1}^{n+1} \qquad (5.7)\\ &=\left[C - \tfrac{1}{2}\lambda q(1-A)\right]U_{j-1}^{n} + \left[1-B-C + \tfrac{1}{2}\lambda q(1-2A)\right]U_{j}^{n} + \left[B+\tfrac{1}{2}\lambda qA\right]U_{j+1}^{n},\end{aligned}$$

where $\lambda = \Delta t/\Delta x$. The choice of the constants in the "ABC" method (5.7) will be influenced by stability and accuracy requirements.

To analyse the stability of (5.7) when applied to the pure Cauchy problem (5.1) we introduce into (5.7) the Fourier component $U_j^n = \xi^n e^{i\omega x_j}$, $\xi \in C$, $\omega \in \mathbb{R}$. This defines the amplification factor as

$$\xi = \frac{1-y\left[B+C+\tfrac{1}{2}\lambda q(2A-1)\right] \pm i\left[B-C+\tfrac{1}{2}\lambda q\right]\sqrt{(2y-y^2)}}{1-y\left[B+C-\tfrac{1}{2}\lambda q(2A-1)\right] \pm i\left[B-C-\tfrac{1}{2}\lambda q\right]\sqrt{(2y-y^2)}}, \qquad (5.8)$$

where $y = 1-\cos(\omega\Delta x)$ and the sign is + or - according as $\omega\Delta x \in [0,\pi]$ or $\omega\Delta x \in [\pi, 2\pi]$, respectively.

In order that (5.7) be a stable approximation we require $|\xi| \leq 1$ or, equivalently, with $q > 0$,

$$y(2B-2C-2A+1) + y^2\,((2A-1)(B+C)-B+C) \leq 0. \qquad (5.9)$$

Clearly the product $\lambda q$ plays no explicit role in the stability of (5.7). Stability requirement (5.9) is equivalent to the restriction

$$\begin{cases} (2A-1)(2B+2C-1) \lesssim 0 \quad \text{if} \quad [(2A-1)(B+C)-B+C] > 0, & (5.10a) \\ (B-C-A+\tfrac{1}{2}) \lesssim 0 \quad \text{if} \quad [(2A-1)(B+C)-B+C] \lesssim 0. & (5.10b) \end{cases}$$

Any subsequent choice of A, B and C must satisfy either (5.10a) or (5.10b) to ensure stability of the pure Cauchy problem.

Morton and Parrot [1980] developed a test function for (5.2) by requiring that the resulting difference scheme should satisfy the unit CFL condition, that is, for $\lambda q = 1$ their scheme represented exactly the movement of $U^n$ through one time step. Alternatively one could say that the unit CFL condition is the condition that $\xi = e^{\mp i\Delta x\omega}$ according as $\lambda q = \pm 1$. Due to the implicit nature of (5.7) one might hope to use values of q greater than unity to achieve the efficiency of the Lax-Wendroff method. The unit CFL condition might therefore be generalised to the condition that $U^{n+1}$ should represent the shift of $U^n$ through one mesh length for $\lambda q > 1$. However to achieve this, the number of elements in the compact support of $\psi(x)$ becomes q dependent and violates the previous constraint of simplicity on the test function. We adopt the criterion of maximising the order of spatial accuracy of (5.7) in its approximation to the solution of (5.1). High spatial accuracy is more desirable than high temporal accuracy as the computational cost is greater to reduce $\Delta x$ than $\Delta t$ to achieve the desired accuracy. Centring the difference approximation about $x = x_j$, $t = t_{n+\frac{1}{2}}$ we obtain a truncation error $O(\Delta x^4) + O(\lambda^{-2}\,\Delta t^4)$ if

$$A = \tfrac{1}{2}, \quad \text{and} \quad B = C = \frac{1}{12}(2+\lambda^2q^2). \tag{5.11}$$

Notice that, for (5.11) the ABC finite element is fourth order in both space <u>and</u> time and the method should be extremely accurate. For the constants (5.11) we see that, on returning to the stability conditions (5.10), that (5.10a) is not relevant and equality holds in (5.10b). Thus the ABC method is marginally stable for any choice of $\lambda q$, and is defined by

$$c_1U_{j-1}^{n+1} + b_1U_j^{n+1} + a_1U_{j+1}^{n+1} = a_1U_{j-1}^n + b_1U_j^n + c_1U_{j+1}^n \tag{5.12}$$

where

$$a_1 = (\lambda q-1)(\lambda q-2),$$

$$b_1 = 8 - 2\lambda^2q^2 \quad ,$$

and

$$c_1 = (\lambda q+1)(\lambda q+2).$$

Finally, using (5.11), we construct the test function as

$$\psi(s) = \phi(s) - \tfrac{1}{2}\lambda^2q^2\sigma(s), \tag{5.13}$$

where

$$\phi(s) = \begin{cases} 1 - |s| & : \ |s| \leqslant 1, \\ 0 & \text{else} \end{cases} \qquad \sigma(s) = \begin{cases} 1 - 2|s| & : \ |s| \leqslant 1 \\ 0 & \text{else} \end{cases} .$$

Returning to (5.8), the amplification factor takes the form

$$\xi(\omega\Delta x) = \frac{4-\lambda^2q^2+(2+\lambda^2q^2)\cos(\omega\Delta x)+i\ 3\lambda q\ \sin(\omega\Delta x)}{4-\lambda^2q^2+(2+\lambda^2q^2)\cos(\omega\Delta x)-i\ 3\lambda q\ \sin(\omega\Delta x)} . \tag{5.14}$$

From (5.14) it follows that $|\xi| = 1$, an indication that the method is marginal stable and non-dissipative. The phase deviation of (5.12) from the exact solution over one time step is given by $\arg(\xi)-\omega q\Delta t$. Plot 1 illustrates the phase errors of the ABC, Crank-Nicolson Galerkin and CNPG (Morton and Parrot [1980]) finite element methods for wave numbers less than $\pi$ and a given value of $\lambda q = 0.95$. A value of $\lambda q = 1$ would yield the exact solution from the ABC and CNPG methods when used to approximate sufficiently smooth problems. As the CNPG and ABC methods are marginally stable at $\lambda q = 1$ in practice a value of $\lambda q$ different from unity would be used. Plot 1 shows that from a dispersion viewpoint the ABC method is clearly superior and it is worthy of further investigation.

### 5.1.1 Stability Analysis and Boundary Approximations

As an alternative to the approach adopted for the L-W and SILF methods, we write equation (5.12) in the differential difference form

$$\alpha\dot{U}_{j-1} + \beta\dot{U}_j + \alpha\dot{U}_{j+1} = \frac{3q}{\Delta x}(U_{j+1}-U_{j-1}) \qquad (5.15)$$

where $\alpha = 1+\frac{1}{2}\lambda^2q^2$, $\beta = 4-\lambda^2q^2$ and $\dot{U}_j$ denotes the method of lines approximation to $\frac{d}{dt}\,u(j\Delta x,t)$. Equation (5.15) may now be analysed for stability by the method of Strikwerda (Section [1.2.2]). It is clear that the stable time-stepping method to be applied to the differential difference approximation, will be the trapezoidal rule.

Laplace transformation of equation (5.15) in time with the dual variable $s \in C$ yields the associated resolvent equation

$$(z\alpha+3q)U_{j-1} + z\beta U_j + (z\alpha-3q)U_{j+1} = 0, \qquad (5.16)$$

where $z = s\Delta x$. Assuming a solution of the difference equation of the form $U_j = d\kappa^j$, $\kappa \in \mathbb{C}$, the characteristic equation of (5.16) is

$$(z\alpha-3q)\kappa^2 + z\beta\kappa + z\alpha + 3q = 0 \qquad (5.17)$$

Lemma 5.1. If $\kappa_1(z)$, $\kappa_2(z)$ and $z$ satisfy (5.17) then

(a) $|\lambda q| < 1$ : (i) $|\kappa_1| < 1$, $|\kappa_2| > 1$ when $\mathrm{Re}(z) > 0$,

(ii) $\kappa_1 = -1$, $\kappa_2 = 1$ for $z = 0$

(b) $\lambda q > 1$ : $|\kappa_{1,2}| > 1$ for $\mathrm{Re}(z) > 0$,

(c) $\kappa(\bar{z}) = \overline{\kappa(z)}$ .

Before we prove Lemma 5.1 we consider the case $\lambda q = 1$. The amplification factor of the ABC method is, in this case,

$$\xi = \frac{1 + e^{i\omega\Delta x}}{1 + e^{-i\omega\Delta x}}$$

and this models exactly the differential factor $e^{i\omega\Delta x}$ unless $1 + e^{-\omega i\Delta x} = 0$. Thus the method is singular if $\omega\Delta x = (2n+1)\pi$, $n \in Z$. On a grid of size $\Delta x$, the representable Fourier modes of the true solution are those whose wave numbers satisfy $\omega\Delta x \to \pi$. Therefore, if the initial data involves modes corresponding to the shortest representable wavelength then the ABC method, with $\lambda q = 1$, cannot be used. As a further illustration of the unsuitability of prescribing $\lambda q = 1$ consider the characteristic equation (5.17). For $\lambda q = 1$, $\kappa_1 = -1$ and $|\kappa_2| > 1$ for any $z$ : $\mathrm{Re}(z) > 0$. This violates the Von-Neumann necessary condition for stability. Another point of interest is $\lambda q = 2$. In this case the ABC method simplifies to

$$U_j^{n+1} = U_{j+2}^n$$

and so one boundary condition and one boundary approximation are needed at $x = 1$. The characteristic equation (5.17) reduces to

$$(z-q)\kappa^2 + z + q = 0 \qquad (5.18)$$

For Re(z) > 0 it may be shown that $|\kappa_{1,2}| > 1$ and so, in agreement with Lemma 5.1(b), there is no general solution of (5.15) belonging to $L_2(\Delta x)$.

Proof of Lemma 5.1:

If $\kappa_1$ and $\kappa_2$ are the roots of (5.17) for a given value of t, then their inverses satisfy

$$(z\alpha+3q)\kappa^{-2} + z\beta\kappa^{-1} + z\alpha - 3q = 0 \quad .$$

As z approaches $3q/\alpha$ in the right half complex plane then $\kappa_2^{-1} \to 0$ and

$$\kappa_1^{-1} \to (-z\beta)/(z\alpha+3q) \sim (\lambda^2q^2-4)/(\lambda^2q^2+2) = [f(\lambda q)]^{-1}.$$

Thus $\kappa_1$ and $\kappa_2$ tend to $f(\lambda q)$ and $\infty$, respectively. From Lemma 5.2 of Gustafsson et al [1972] we require $|\kappa_1| \leq 1$ for Re(z) > 0. From figure (5.1) we see that

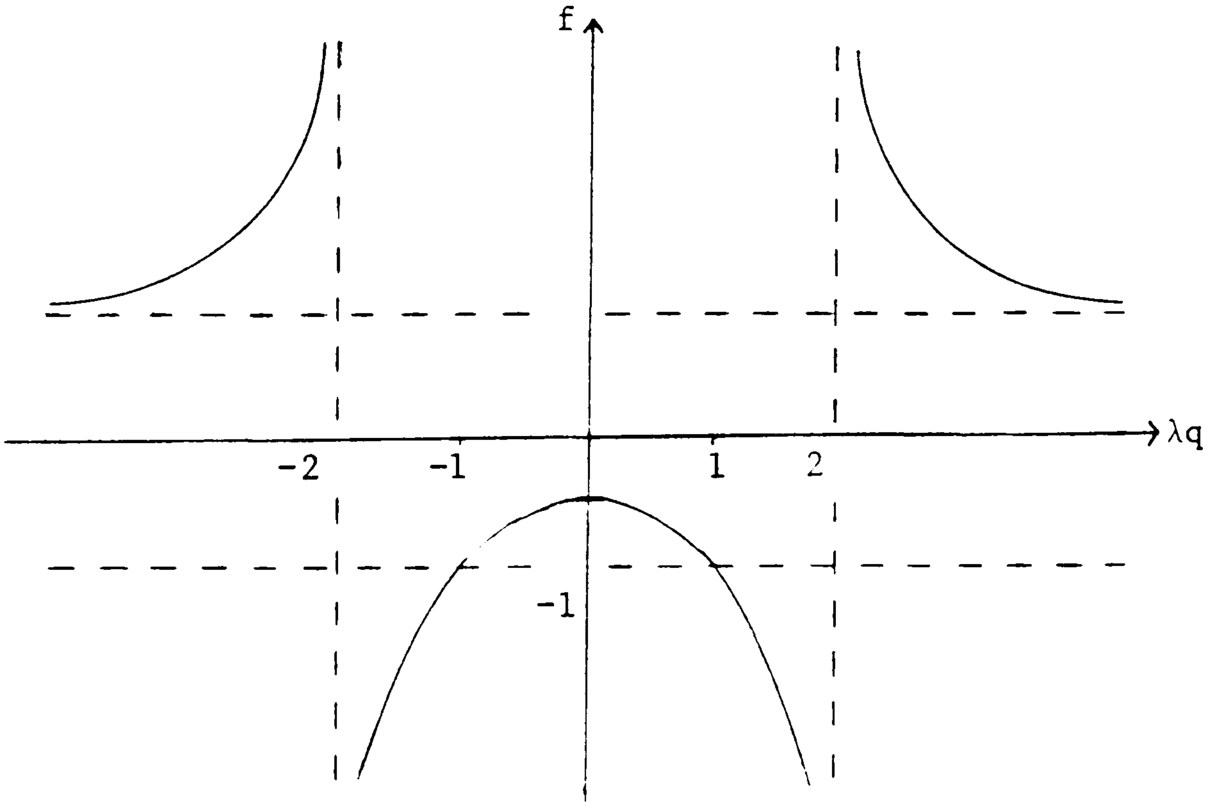


Figure 5.1: $f(\lambda q)$.

$|f| < 1$ if and only if $|\lambda q| < 1$. Therefore we have proved (a)(1) and (b), if $\lambda q$ in modulus is less than unity. When $z = 0$ we have $\kappa = \pm 1$. To determine the exact nature of $\kappa$ consider $z = \delta > 0$ and $\kappa = 1+\varepsilon > 1$. From (5.17), $q\varepsilon \sim \delta$ and so when $z = 0$ the root of $\kappa = 1$ is an outer root. A similar process reveals $\kappa = -1$ as an inner root. Part (c) is readily established.

The significant result of Lemma 5.1 is part (b). According to Gustafsson et al [1972] the difference method (5.12) should have, for $\mathrm{Re}(z) > 0$, one solution of the form $U_j = d\kappa^j$ where $|\kappa(z)| < 1$. Thus when $\lambda q > 1$, this result is violated and the ABC method is no longer well-defined. The additional constraints on $\lambda q$ are related to the invertibility constraints mentioned by Iserles [1983] in his order-star treatment of stability.

With the above results available we consider, for $q > 0$, the following approximations at $x = 0$,

Truncated Element (TE):

$$\tfrac{1}{2}\beta\dot{U}_o + \alpha\dot{U}_1 = \frac{3}{\Delta x}q(U_1-U_o), \qquad (5.19)$$

Box (B):

$$\dot{U}_o + \dot{U}_1 = \frac{2}{\Delta x}q(U_1-U_o), \qquad (5.20)$$

Characteristics (C):

$$\dot{U}_o = \frac{q}{12\Delta x}(-25U_o+48U_1-36U_2+16U_3-3U_4). \qquad (5.21)$$

Approximation (5.19) is the so-called 'natural' boundary approximation of finite element methods. It is derived by evaluating the inner product (5.2) with $\phi(x)$ and $\sigma(x)$ set to zero when $x < 0$. Approximation

(5.20) is the method of lines analogue of the 'box' integration of (5.1) at $x = 0$. Turkel [1980] used this approximation in conjunction with the standard Galerkin scheme. The C approximation is a representation of the outgoing (the only in this case) characteristic equation centred about, in the full difference form, $x = 0$, $t = t_{n+\frac{1}{2}}$. This approach is analogous to the semi-characteristic boundary approximations of Bramley and Sloan [1977] and Sloan [1980]. Apart from approximation C the boundary approximations are sufficiently inaccurate (the truncation errors of TE and B are $O(\Delta t^2) + O(\Delta x)$ and $O(\Delta t^2) + O(\Delta x^2)$) to adversely affect the overall accuracy of the integration (Gustafsson [1975]). The truncation error of C is $O(\Delta x^4) + O(\Delta t^2)$. In practice, however, the box boundary method performs well, and being more compact than C, will be easier to apply to more complex situations than (5.1).

After Laplace transforming each boundary approximation and seeking the solution $U_j = d\kappa^j$ we obtain, in the now familiar way, the boundary determinant equation given in Table 5.1.

| Boundary Approximation | Determinant Equation |
|---|---|
| TE | $\kappa + (3q+\frac{1}{2}\beta z)/(\alpha z-3q) = 0$ |
| B | $\kappa + (z+2q)/(z-2q) = 0$ |
| C | $z + \frac{q}{12}(25-48\kappa+36\kappa^2-16\kappa^3+2\kappa^4) = 0$ |

Table 5.1: Determinant equations of (5.19)-(5.21).

According to section 2.2.2 , the ABC method (5.15), and any of (5.19) - (5.21), provide a stable approximation to (5.1) if there are no solution pairs $(\kappa, z)$ of (5.17), and the appropriate equation from Table 5.1, such that $|\kappa| < 1$ for $\mathrm{Re}(z) > 0$ or $|\kappa| = 1$ for

$\mathrm{Re}(z) = 0$.

Lemma 5.2. The approximation to the advection equation (5.1), with $q > 0$, given by (5.15) and any choice of (5.19) - (5.21) is a stable approximation.

Proof. Consider approximation TE. Substitution of the determinant equation into (5.17) yields the quadratic

$$(\tfrac{3}{4}\beta^2-\alpha^2)z^2 + 6q\beta z + 9q^2 = 0$$

whose roots are given by

$$\frac{z}{q} = \frac{-6(4-\lambda^2q^2) \pm 3\surd(\lambda^4q^4-4\lambda^2q^2+20)}{(22-14\lambda^2q^2-\lambda^4q^4)} .$$

For any $\lambda q \in (0,1)$ it is readily shown that $\mathrm{Re}(z/q) < 0$ establishes stability. The approximations defined by (5.15) and (5.20) can be treated similarly. To resolve the approximation involving (5.21) the roots of a hexic polynomial in $\kappa_1$, arising through the combination of (5.17) and the determinant equation, have to be evaluated numerically. For the data set $\lambda q = 0.05(0.05)0.95$ no instability pairs $(\kappa_1, z)$ were found.

### 5.1.2 Numerical Results

Equation (5.1) together with the initial data $u(x,0) = 2\pi\cos 2\pi x$ and exact boundary data at $x = 1$ was solved numerically using the ABC method and each of the previous boundary approximation. The results are given for $\lambda q = 0.95$ in Table 5.2 where the errors were measured by the maximum norm $\|\underset{\sim}{U}_\Delta\|_\infty$. The error results found for $q = 1$, 2 and 5 over three wave cycles, were identical over 4 decimal places.

| Boundary Approximation | $\Delta x = 1/10$ | $\Delta x = 1/20$ |
|---|---|---|
| Exact | 0.0077 | 0.00051 |
| TE | 0.0398 | 0.0105 |
| B | 0.0109 | 0.0011 |
| C | 0.0285 | 0.0010 |

Table 5.2: $\|U_\Delta\|_\infty$ for $\lambda q = 0.95$.

The numerical results indicate the high order accuracy of the ABC method. To achieve accuracy comparable with the ABC method at $\Delta x = 1/10$ the Lax Wendroff method required $\Delta x = 1/52$. The respective CPU times were 0.44 seconds and 1.08 seconds. Clearly the ABC method represents a significant improvement. The C boundary approximation results indicate the high convergence rate however the more compact B approximation performs very well.

Kreiss [1980] has shown that by prescribing 'wrong' boundary data at $x = 0$ we can introduce, into the general solution, a wave travelling against the characteristic direction. We show now that, by corrupting the analytic boundary data at $x = 1$, we can induce instability.

Lesaint [1973] has described an alternative "weak" formulation of the analytic boundary condition. In this context we interpret this formulation as a combination of the ABC method truncated at $x = 1$, as in the TE approximation, and the analytic boundary condition. If $c \in \mathbb{R}$ then the weak boundary condition considered is

$$c(\alpha\dot{U}_{J-1}+\tfrac{1}{2}\beta\dot{U}_J) - 3q\,\frac{c}{\Delta x}(U_J-U_{J-1}) - U_J + g(1,t) = 0. \qquad (5.22)$$

When $c = 0$, condition (5.22) reduces to the analytic boundary condition $U_J = g(1,t)$ where $u(x,t) = g(x,t)$ is the solution of (5.1) for given initial data. In (5.22) the analytic boundary condition has been perturbed by a proportion c of the residual involved in a finite-element formulation. We prove

<u>Lemma 5.3.</u> For the right boundary problem of the advection equation, $q > 0$, the ABC method together with (5.22) will produce an unstable approximation for $c \notin \left[-\frac{\Delta x}{6q}, 0\right]$. In practice this interval is very small.

Proof: The general solution of the right boundary problem given by (5.15) and (5.22) will, after Laplace transforming in time, be of the form $U_{J-j} = d\kappa^{-j}$ where $\kappa^{-1}$ is an inner root of

$$(z\alpha+3q)\kappa^{-2} + z\beta\kappa^{-1} + z\alpha - 3q = 0. \tag{5.23}$$

The determinant equation of (5.22) is

$$c(z\alpha+3q)\kappa^{-1} + \tfrac{1}{2}\beta cz - 3qc + 1 = 0 \tag{5.24}$$

For stability we require no roots $(\kappa^{-1}, z)$ of (5.23) and (5.24) such that $\mathrm{Re}(z) > 0$ with $|\kappa(z)| > 1$. Therefore we consider the pair of equations

$$(z\alpha-3q)\kappa^{2} + z\beta\kappa + z\alpha + 3q = 0, \tag{5.25}$$

$$z\alpha + 3q + (\tfrac{1}{2}\beta z-3q-c')\kappa = 0; \quad cc' = h \tag{5.26}$$

From (5.27),

$$\kappa = (\alpha z+3q)/(3q+c'-\tfrac{1}{2}\beta z) \tag{5.27}$$

which, when substituted into (5.25), yields

$$z_{\pm} = \pm\sqrt{\left[\frac{c'(6q+c')}{3(1-\lambda^{2}q^{2})}\right]} \tag{5.28}$$

Consider

(a) $c'\in[0,\infty)$: $z = z_+$ produces, from (5.26), $|\kappa| > 1$ for $c' > 0$ and $\kappa = 1$ for $c' = 0$. Thus since $\kappa = 1$ is an outer root (5.22) provides an unstable approximation for $c' \geqslant 0$,

(b) $c'\in[-6q,0)$: $z_{\pm} \in \mathbb{C}$ and from (5.26) $|\kappa| = 1$. If $c' = -\delta < 0$ and $\eta = \sqrt{\left[\frac{\delta}{3}(\delta q-\delta)/(1-\lambda^2 q^2)\right]}$ then $z = \varepsilon + i\eta$ $(\varepsilon > 0)$ implies that $|\kappa| > 1$ and so $c'\in(-6q,0)$ yields a generalised eigenfunction. For $c' = -6q$ we have $z = 0$ and from (5.26) $\kappa = -1$, an inner root, so stability follows for $c' = -6q$.

(c) $c' < -6q$: $z = z_+$ yields from (5.26), with $c' = -6q-\delta$ $(\delta > 0)$ that $|\kappa| = (3q+\alpha z)/(3q+\delta+\frac{1}{2}\beta z) < 1$ $\forall\delta$. Thus (5.22) provides a stable approximation.

Stability for $c = 0$ is immediate from the Cauchy stability of the ABC method.

Gustafsson et al [1972] have shown for systems of equations that distortion of the analytic boundary condition can render an otherwise unstable boundary approximation, used for the appropriate quarter-plane problem, stable. The possibility remains for a boundary condition of the form (5.22) being of use if and when we consider an unstable boundary approximation in conjunction with the ABC approximation of our test problems.

To illustrate Lemma 5.3 we apply (5.22) to the previous advection problem with $U_0^n = g(0,n\Delta t)$. The results are given in Table 5.3 where the initial data was corrupted

| c | q = 1 | q = 2 | q = 5 |
|---|---|---|---|
| -h/q | 6.095 | 6.090 | 5.680 |
| -h/6q | 0.042 | 0.042 | 0.042 |
| -h/600q | 0.017 | 0.018 | 0.014 |
| 0 | 0.016 | 0.012 | 0.017 |
| 0.01 | 3.382 | 412.6 | 5.6(6) |

Table 5.3: $||U_\Delta||_\infty$ results with $\lambda q = 0.95$, $\Delta x = \frac{1}{10}$, rough data.

by a random number of magnitude $10^{-3}$. The result of Lemma 5.3 indicates that rounding errors induced by the interior finite-element approximation may not only cause inaccuries but lead quickly to instability. Since, if q is large and a fine mesh is used, the allowable proportion c that corrupts the boundary data must be very small.

We conclude this section by noting that the ABC method assumes the use of the trapezoidal time-step rule. Any stable method could have been used (Vichnevetsky and Bowles [1982]). If we were to use the Leap-frog method then the test function would have been

$$\psi(s) = \phi(s) + \tfrac{1}{2}\lambda^2 q^2 \sigma(s).$$

## 5.2 The Fast-Wave Problem

Having determined an accurate and stable difference method for the advection equation we can now generalise the ABC method to approximate the solution of systems of hyperbolic equations. The fast wave test problems are of the form

$$\underset{\sim}{u}_t = A\underset{\sim}{u}_x, \quad 0 \leqslant x \leqslant 1, \quad t \geqslant 0. \qquad (5.29)$$

Associated with a symmetric A, there exists the orthogonal matrix

$$S = \frac{1}{\sqrt{2}}\begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}$$

which diagonalizes (5.29) to

$$\underset{\sim}{v}_t = \Lambda \underset{\sim}{v}_x, \quad 0 \leqslant x \leqslant 1, \quad t \geqslant 0, \tag{5.30}$$

with $\underset{\sim}{v}(x,t) = (v_1(x,t),\ v_2(x,t))^T$, $\underset{\sim}{v} = S^T\underset{\sim}{u}$ and $\Lambda = S^TAS = \text{diag}(\mu_1,\mu_2)$, $\mu_i$ an eigenvalue of A. Equation (5.31) is two scalar equations in the characteristic variables $v_1, v_2$ associated with each eigenvalue $\mu_i$. We treat each equation of (5.30) in a manner similar to (5.1), that is we consider the approximate solution of the weak problem

$$<\underset{\sim}{v}_t - \Lambda \underset{\sim}{v}_x, \underset{\sim}{\psi}> \ = 0 \tag{5.31}$$

where the test function, vector $\underset{\sim}{\psi} = (\psi_1,\psi_2)^T$ is given by

$$\psi_i(s) = \phi_i(s) - \tfrac{1}{2}\lambda^2\mu_i^2\sigma(s), \quad i = 1,2. \tag{5.32}$$

The Petrov-Galerkin approximation to $\underset{\sim}{V}(x,t)$ is the solution $\underset{\sim}{V}$ of

$$<\dot{\underset{\sim}{V}} - \Lambda\delta_x\underset{\sim}{V},\ (I\phi_i + B\sigma_i)\underset{\sim}{e}_{(r)}> \ = 0,\ r = 1,2, \forall i, \tag{5.33}$$

where $\underset{\sim}{e}_{(r)}$ denotes the rth unit vector of $\mathbb{R}^2$ and the matrix $B = -\frac{\lambda^2}{2}\text{diag}(\mu_1^2,\mu_2^2)$. Since S is orthogonal and $\underset{\sim}{V} = S^T\underset{\sim}{U}$ we obtain from (5.33)

$$<\dot{\underset{\sim}{U}} - A\delta_x\underset{\sim}{U},\ (I\phi_i + \tilde{B}\sigma_i)\underset{\sim}{e}_{(r)}> \ = 0;\ r = 1,2;\ \forall i \tag{5.34}$$

with $\tilde{B} = SBS^T$. The derivation of (5.34) depended upon S being orthogonal and A being symmetric.

Letting $\underset{\sim}{U} = \sum_j \underset{\sim}{U}_j \phi_j$ in (5.34) then evaluation of the inner products and use of the trapezoidal rule yields the ABC approximation to the Cauchy problem of (5.29) as

$$\{6+[1 + \frac{\lambda^2}{4}(q^2+1)]\delta^2 + \frac{\lambda^2}{4}(q^2-1)\delta^2 I'\}(\underset{\sim}{U}_i^{n+1}-\underset{\sim}{U}_i^{n}) = 3\lambda A\Delta_o(\underset{\sim}{U}_i^{n+1}+\underset{\sim}{U}_i^{n})\,\forall i \tag{5.36}$$

where $I' = \begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}$ and $\delta^2$ and $\Delta_o$ are the usual difference operators.

To utilise the results of the previous section we consider the differential-difference form of (5.36)

$$\{6+[1 + \frac{\lambda^2}{4}(q^2+1)]\delta^2 + \frac{\lambda^2}{4}(q^2-1)\delta^2 I'\}\dot{\underset{\sim}{U}}_i = \frac{6}{\Delta x} A\Delta_o \underset{\sim}{U}_i;\ \forall i. \tag{5.37}$$

5.2.1 Stability Analysis and Boundary Approximations

The imposition of a boundary at x = 0 creates the left quarter-plane problem the integration of which, using (5.37), requires a differential-difference boundary approximation to $\frac{\partial}{\partial t}$ v(0,t). We consider the following approximations

$$(TE): [2 - \frac{\lambda^2}{4}(q^2+1)]\dot{V}_o + [1 + \frac{\lambda^2}{4}(q^2+1)]\dot{V}_1 - \frac{\lambda^2}{4}(q^2-1)(\dot{U}_o-\dot{U}_1) = \frac{3b}{\Delta x}(U_1-U_o)+\frac{3a}{\Delta x}(V_1-V_o). \tag{5.38}$$

$$(G): -[2a + \frac{\lambda^2}{4}q(q-1)]\dot{U}_o + [2b - \frac{\lambda^2}{4}q(q+1)]\dot{V}_o + [-a + \frac{\lambda^2}{4}q(q-1)]\dot{U}_1 + [b + \frac{\lambda^2}{4}q(q+1)]\dot{V}_1 = \frac{3q}{\Delta x}(U_1-U_o). \tag{5.39}$$

$$(B):\ \dot{V}_o + \dot{V}_1 = \frac{2a}{\Delta x}(V_1-V_o) + \frac{2b}{\Delta x}(U_1-U_o). \tag{5.40}$$

$$(C):\dot{R}_o = \frac{q}{12\Delta x}(-25R_o+48R_1-36R_2+16R_3-3R_4);\quad R = U+V. \tag{5.41}$$

$$(BC):\dot{R}_o + \dot{R}_1 = \frac{2q}{\Delta x}(R_1-R_o). \tag{5.42}$$

The TE, B and C boundary approximations are analogues of (5.19)-(5.21) for the system (5.29). The BC approximation is a 'box' integration of the outgoing characteristic equation considered in (5.41). Whilst not as accurate as (5.41) the BC approximation may be used without destroying the compact seven-band nature of the implicit matrix in (5.37). Approximation (5.39) is developed by applying the energy conserving matrix E, associated with (5.29), to the test functions at the boundaries (Gunzburger [1977]). That is, it arises from the evaluation of

$$<\dot{\underset{\sim}{U}} - A\delta_x\underset{\sim}{U}, \; (I\phi_i + \tilde{B}\sigma_i)E\underset{\sim}{e}_{(2)}> = 0, \quad i = 0, \tag{5.43}$$

where $\phi_o, \sigma_o$ have been truncated as in (5.38) and

$$E = \begin{bmatrix} b & -a \\ -a & b \end{bmatrix}.$$

In his paper, Gunzburger considered an off-diagonal problem of the form (5.29) and used (5.43) to stabilise an otherwise unstable Galerkin method. However (5.39) is only first-order accurate and is included for interest only as it is not expected to be able to compete with any of (5.40)-(5.42). For q = 1 the G and TE approximations are identical.

Returning to (5.37), Laplace transforming in time and seeking a solution of the difference equations of the form $\underset{\sim}{U}_j = \underset{\sim}{d}\kappa^j$ we obtain the characteristic equations, associated with the left boundary problem, as

$$(z\alpha - 3\mu)\kappa^2 + z\beta\kappa + z\alpha + 3\mu = 0, \tag{5.44a,b}$$

with $\alpha = 1 + \frac{1}{2}\lambda^2\mu^2$, $\beta = 4-\lambda^2\mu^2$ and $\mu = q,-1$. Evaluation of the respective eigenvectors yields the general solution which decays as j increases

$$\begin{aligned} U_j(z) &= \eta_1\kappa_1^j + \eta_3\kappa_3^j \quad , \\ V_j(z) &= \eta_1\kappa_1^j - \eta_3\kappa_3^j \quad , \end{aligned} \qquad (5.45)$$

where $\kappa_1(z;q)$ and $\kappa_3(z;-1)$ are the inner roots of (5.44a) and (5.44b) respectively. We denote the outer roots by $\kappa_3(z;q)$ and $\kappa_4(z;-1)$. Therefore, Laplace transforming any boundary approximation and using (5.45) will yield a determinant equation $D(z,\kappa_1,\kappa_3) = 0$. If there are no eigenfunctions or generalised eigenfunctions of (5.44) and $D(z,\kappa_1,\kappa_3) = 0$ then the approximation may be regarded as stable. This is an application of Theorem 1.9.

<u>Result 5.1</u>. The approximation to the left quarter-plane problem given by (5.37) and any one of (5.38)-(5.42) is a stable approximation for a wide choice of $\lambda$ and q such that $\lambda q < 1$.

Result 5.1 was established numerically for q = 1,2 and 5 using a variety of techniques. Consider (5.41). The determinant condition is

$$z = \frac{q}{12}(-25+48\kappa_1-36\kappa_1^2+16\kappa_1^3-3\kappa_1^4)$$

which, when used to replace t in (5.44a), produces a real coefficient polynomial in $\kappa$, of degree 6. For $\lambda = 0(\frac{1}{20q})\ \frac{0.95}{q}$ no unstable roots were found. Approximation (5.42) was treated similarly. For the same data set, approximations (5.38) and (5.40) were treated using the resultant approach of Section 2.1. An example of the application of the resultants is given in Appendix I. Approximation (5.39) was resolved by using Theorem 1.10 with exact boundary values at $x = 1$ and $\lambda q = 0.95$.

For the approximation of the right boundary problem we develop the characteristic equations and general solution in a manner akin to 3.2 .

The general solution of the form $\underset{\sim}{U}_{J-j} = \underset{\sim}{d}\kappa^{-j}$ of the right boundary resolvent equations which decreases as j increases is $\underset{\sim}{U}_{J-j} = \eta_2\kappa_2^{-j}\binom{1}{-1} + \eta_4\kappa_4^{-j}\binom{1}{-1}$, where $\kappa_2(z;q)$ and $\kappa_4(z;-1)$ are the outer roots of (5.44a) and (5.44b), respectively. If the equivalent left boundary problem were being analysed then the characteristic equations would be

$$(z\alpha+3\mu)\kappa^2 + z\beta\kappa + z\alpha - 3\mu = 0; \quad \mu = q,-1 \qquad (5.46)$$

and we would require the inner roots of (5.47). However these inner roots are $\kappa_1(z;-q)$ and $\kappa_3(z;1)$ respectively and must therefore be respectively equivalent to $\kappa_2^{-1}(z;q)$ and $\kappa_4^{-1}(z;-1)$. The general solution of the resolvent equations for the right boundary problem is

$$\underset{\sim}{U}_{J-j} = \eta_2\kappa_1^j(z;-q)\begin{bmatrix}1\\-1\end{bmatrix} + \eta_4\kappa_3^j(z;1)\begin{bmatrix}1\\1\end{bmatrix}.$$

A perturbation analysis similar to Lemma 5.1 shows that $\kappa_1(0;-q) = 1$ and $\kappa_3(0;1) = -1$.

The analogues of the left boundary approximations are

$$(\mathrm{TE}):\ \left[2 - \frac{\lambda^2}{4}(q^2+1)\right]\dot{V}_J + \left[1 + \frac{\lambda^2}{4}(q^2+1)\right]\dot{V}_{J-1} - \frac{\lambda^2}{4}(q^2-1)(\dot{U}_J-\dot{U}_{J-1})$$

$$= \frac{3a}{\Delta x}(V_J-V_{J-1}) + \frac{3b}{\Delta x}(U_J-U_{J-1}), \qquad (5.47)$$

$$(\mathrm{G}):\ -\left[2a + \frac{\lambda^2}{4}\, q(q-1)\right]\dot{U}_J + \left[2b - \frac{\lambda^2}{4}\, q(q+1)\right]\dot{V}_J + \left[-a + \frac{\lambda^2}{4}\, q(q-1)\right]\dot{U}_{J-1}$$

$$+ \left[b + \frac{\lambda^2}{4}\, q(q+1)\right]\dot{V}_{J-1} = \frac{3q}{\Delta x}(U_J-U_{J-1}), \qquad (5.48)$$

$$(\mathrm{B}):\ \dot{V}_J + \dot{V}_{J-1} = \frac{2b}{\Delta x}\,(U_J-U_{J-1}) + \frac{2a}{\Delta x}\,(V_J-V_{J-1}) \qquad (5.49)$$

$$(\mathrm{C}):\ \dot{W}_J = -\frac{1}{12\Delta x}\left[3W_{J-4} - 16W_{J-3} + 36W_{J-2} - 48W_{J-1} + 25W_J\right];\ W = U-V, \qquad (5.50)$$

$$(\mathrm{BC}):\ \dot{W}_J + \dot{W}_{J-1} = \frac{2}{\Delta x}(W_{J-1}-W_J). \qquad (5.51)$$

Result 5.2. The approximation to the right quarter plane problem given by (5.37) and any one of (5.47)(5.51) is a stable approximation for the data set $(\lambda,q)$ given in Result 5.1.

Result 5.2 was established in the manner of Result 5.1.

### 5.2.2 Numerical Results

Tables 5.4 and 5.5 below present the error results for the ABC integration of the test problems of Chapters 3 and 4 (cf Tables 4.2, 4.3). The high accuracy attained with the ABC method is clear, with the characteristic based boundary approximations proving to be the most competitive. The results also indicate a high loss of accuracy for the application of the TE and B approximations to the right boundary problem. In this case greater accuracy could be obtained by using the characteristic correction of Gottlieb, Gunzburger and Turkel [1982]. This behaviour may be explained by examining the eigenvalues of the coefficient matrix of the ABC finite difference method

| Boundary Approximations | q = 1 | | q = 2 | | | | q = 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Left | | Right | | Left | | Right | |
| Exact | 39.481 | 0.0005 | 78.935 | 0.0031 | 87.935 | 0.0031 | 197.39 | 0.0017 | 197.39 | 0.0017 |
| TE | 39.539 | 0.0141<br>(0.0037) | 79.924 | 0.106 | 78.031 | 0.0453 | 197.80 | 0.0280 | 205.72<br>(198.14) | 2.205<br>(0.659) |
| G | 39.539 | 0.0141 | 81.036 | 0.218 | 78.666 | 0.164 | 199.50 | 0.1831 | 197.05 | 0.188 |
| B | 39.485 | 0.0019 | 78.997 | 0.0092 | 78.482 | 0.065 | 197.35 | 0.0035 | 197.60 | 0.357 |
| C | 39.481 | 0.0005 | 78.935 | 0.0031 | 78.980 | 0.0023 | 197.39 | 0.0017 | 197.41 | 0.0034 |
| BC | 39.481 | 0.0005 | 78.935 | 0.0031 | 78.806 | 0.0156 | 197.39 | 0.0017 | 197.20 | 0.0349 |

Table 5.4: $\|\underset{\sim}{U}_\Delta\|_{E_{q,\varepsilon}}$ | $\|\underset{\sim}{U}_\Delta\|_\infty$ for $\lambda q = 0.95$, $\Delta x = \frac{1}{20}$, $\varepsilon = 0.01$

Figures in parenthesis were evaluated with $x = \frac{1}{40}$ .

| Boundary Approximations | q = 1 | | q = 2 | | | | q = 5 | | | |
|---|---|---|---|---|---|---|---|---|---|---|
| | | | Left | | Right | | Left | | Right | |
| Exact | 49.346 | 0.0008 | 88.800 | 0.0035 | 88.800 | 0.0035 | 207.25 | 0.0021 | 207.25 | 0.0021 |
| TE | 49.423 | 0.0140 | 89.936 | 0.107 | 88.683 | 0.456 | 207.69 | 0.030 | 219.71 | 2.639 |
| G | 49.423 | 0.0140 | 91.252 | 0.219 | 88.885 | 0.166 | 209.34 | 0.181 | 206.88 | 0.186 |
| B | 49.341 | 0.0015 | 88.868 | 0.0086 | 88.868 | 0.0086 | 207.21 | 0.0037 | 208.40 | 0.902 |
| C | 49.360 | 0.0036 | 88.803 | 0.0041 | 88.842 | 0.0027 | 207.25 | 0.0033 | 207.27 | 0.0033 |
| BC | 49.344 | 0.0015 | 88.800 | 0.0047 | 88.691 | 0.0153 | 207.25 | 0.0025 | 207.08 | 0.0344 |

Table 5.5: $\|\underset{\sim}{U}_{\Delta}\|_{E_{q,\varepsilon}}$ $\|\underset{\sim}{U}_{\Delta}\|_{\infty}$ for $\lambda = 0.95$, $\Delta x = \frac{1}{20}$, $\varepsilon = 0.5$

in each case. Consider the TE approximation (the box approximation exhibits similar results). The eigenvalues suggest that application of the left TE approximation results in a dissipative approximation while, at the right boundary, eigenvalues exist that exceed unity; for example, with $q = 2$, $\lambda = 0.475$ the spectral radii for $x = 1/5$ and $x = 1/10$ are respectively 1.13 and 1.06. Therefore exponentially growing solutions allowed by the stability theory exist. If we consider the coefficient matrix of the twin boundary problem we see that all eigenvalues lie on the unit circle regaining the non-dissipative approximation. Therefore, in some sense, the dissipation from the left boundary is controlling the solution growth introduced at $x = 1$.

For interest we give the integration results using the Crank-Nicolson Galerkin and CNPG interior methods on the problem (3.27).

| Interior Method | $\varepsilon = 0.01$ | | | $\varepsilon = 0.5$ | | |
|---|---|---|---|---|---|---|
| | $q = 1$ | $q = 2$ | $q = 5$ | $q = 1$ | $q = 2$ | $q = 5$ |
| C-N Galerkin | 0.344 | 0.131 | 0.016 | 0.511 | 0.265 | 0.271 |
| CNPG | 0.0040 | 0.0129 | 0.0022 | 0.0040 | 0.0129 | 0.0022 |
| ABC | 0.0005 | 0.003 | 0.0017 | 0.0008 | 0.004 | 0.002 |

Table 5.6: $\|\underset{\sim}{U}_{\lambda\Delta}\|_\infty$ for $\lambda q = 0.95$, $\Delta x = \frac{1}{20}$, exact boundaries.

The superiority of the Petrov-Galerkin methods is evident, and they are much less sensitive to the strength of the fast wave. For small values of q the ABC method is the most accurate method.

### 5.2.3 Sköllermo Accuracy Analysis:

To construct accuracy predictions based on the analysis of Sköllermo as in Chapter 3 we consider the difference method (5.36). Constructing the relevant determinant equations and determining the limits points of the inner roots as $z \to 1$ (see equation (3.54)) we can compute the error functions $e(\omega)$ associated with the left and right boundaries in Table 5.7..

| Boundary Approximation | Left $e(\omega)$ | Right $e(\omega)$ |
|---|---|---|
| TE | $\frac{2}{3}\frac{\pi^2}{q^2M^2}\,a(1+\lambda^2q)$ | $\frac{2}{3}\frac{\pi^2a}{qM^2}\,(1+\lambda^2q)$ |
| G | $\frac{\pi^2}{3q^2M^2}\,(q^2-1)$ | $\frac{\pi^2}{3qM^2}\,(q^2-1)$ |
| B | $\frac{\pi^3}{3q^3M^3}\,(1+q^2(1-2\lambda^2))$ | $\frac{\pi^3}{3q^2M^2}\,(q^2(1-2\lambda^2)+1)$ |
| C | $\frac{\pi^3}{16q}\,\frac{\lambda^2}{M^3}$ | $\frac{\pi^3}{16}\,\frac{\lambda^2}{M^3}$ |
| BC | $\frac{\pi^3}{3q^3M^3}\,(\lambda^2q^2-1)$ | $\frac{\pi^3}{3M^3}\,(\lambda^2-1)$ |

Table 5.7 : $e(\omega)$ for left and right boundary approximations.

The expression for the TE and G boundary approximations at $q = 1$ is $\frac{2}{45}\frac{\pi^5}{M^5}(\lambda^4-5\lambda^2+4)$.

The minimum number of mesh points per wavelength required to achieve an error of 0.01 for each boundary approximation is given in Table 5.3.

| Boundary Approximation | q = 1 | q = 2 | | q = 5 | |
|---|---|---|---|---|---|
| | | Left | Right | Left | Right |
| TE | 4 | 11 | 16 | 8 | 9 |
| G | 4 | 16 | 23 | 18 | 40 |
| B | 6 | 8 | 10 | 6 | 11 |
| C | 6 | 3 | 4 | 2 | 2 |
| BC | 5 | 3 | 10 | 1 | 10 |

Table 5.8 : Minimum value of $M = \frac{1}{\omega \Delta x}$ for tolerance 0.01

According to the above results the Sköllermo analysis predicts that the finite element approximations are desirable only in the symmetric problem. The characteristic formulations are exceptional for an asymmetric situation. Kreiss and Oliger [1973] have shown that we must have two points per wavelength and so the C approximation is, overall, the best choice. The results of Table 5.8 are, in general, not supported by Table 5.4 and 5.5. The analysis pertaining to Table 5.8 related to a differential problem with homogeneous boundary conditions. An example is that of problem (3.57). The maximum norm results of the ABC integration of (3.57) are contained in Table 5.9.

| Boundary Approximation | q = 1 | q = 2 |  | q = 5 |  |
|---|---|---|---|---|---|
|  |  | Left | Right | Left | Right |
| Exact | 0.0005 | 0.0016 | - | 0.0010 | - |
| TE | 0.0005 | 0.0623 | 0.2412 | 0.0169 | 1.3112 |
| G | 0.0005 | 0.1204 | 0.0836 | 0.0904 | 0.0987 |
| B | 0.0006 | 0.0043 | 0.0338 | 0.0018 | 0.2264 |
| C | 0.0021 | 0.0025 | 0.0014 | 0.0015 | 0.0019 |
| BC | 0.0008 | 0.0019 | 0.0076 | 0.0011 | 0.0174 |

Table 5.9.: $||\underset{\sim}{U}_{\Delta}||_{\infty}$ for $\lambda q = 0.95$, $\Delta x = \frac{1}{20}$ .

For the symmetric problem the high accuracy of the TE (or G) approximation is substantiated by Table 5.9. For the B, C and BC boundary approximations the performance at q = 2 and q = 5 is also verified, however for the TE and G techniques the Skollermo analysis provides a less reliable guide to accuracy.

In summary, for problem (5.29) with q > 1, the best boundary approximations are those derived from the associated characteristics. Whilst the approximation C is clearly superior the generalisation to non-linear problems extends the band number of the implicit matrix of the ABC method, akin to (5.36), beyond that required for the Cauchy problem. We may therefore have increased the compact support of $\phi(x)$ and $\psi(x)$. The BC boundary approximation provides an accurate, stable and applicable alternative.

It would be of value to extend the ABC method to multidimensional problems. However, the method defined in this chapter is dependent on the eigenvalues of the one-dimensional situation. Therefore any extension of the scheme, in its present form, will necessitate a splitting of the differential equations. Further investigation is required.

In this chapter we have developed a highly accurate finite element approximation to the solution of the linear fast wave test problems. This ABC method is superior to the standard Galerkin and the Petrov-Galerkin CNPG (Morton and Parrot [1980]) methods. For q > 1 we found that the most suitable boundary approximations, for use in the twin boundary situation, where those that involved the characteristic variables and equations. These are similar conclusions to those of Chapters 3 and 4.

# CHAPTER 6

# NON-LINEAR PROBLEMS

In this chapter we consider the one-dimensional shallow water equations

$$\left.\begin{aligned}\frac{\partial u}{\partial t} &= -u\frac{\partial u}{\partial x} - \frac{v}{2}\frac{\partial v}{\partial x}\\ \frac{\partial v}{\partial t} &= -\frac{v}{2}\frac{\partial u}{\partial x} - u\frac{\partial v}{\partial x}\end{aligned}\right\} , \ 0 < x < 1, \ t > 0, \qquad (6.1)$$

where u(x,t) is the fluid velocity and v(x,t) is the scaled potential with $v > 2|u|$ for all x and t in the problem domain. We specify the solid wall boundary conditions

$$u(0,t) = u(1,t) = 0. \qquad (6.2)$$

System (6.1) is a non-linear example of system (3.1) to which the stability results of Chapter 3, 4 and 5 may be applied if the derivative coefficients are regarded as the constants $u_o$ and $v_o$ obtained from the computed values for the boundary node at $t = n\Delta t$. The characteristic speeds of this linearised system are $\pm \frac{1}{2}v_o + u_o$. By scaling the independent variables we can show that the fast speed of (3.1) can be expressed as

$$q = \begin{cases}(v_o + 2u_o)/(v_o - 2u_o) : u_o \geqslant 0\\ \\ (v_o - 2u_o)/(v_o + 2u_o) : u_o < 0\end{cases} \qquad (6.3)$$

This problem was considered by Coughran (1980) where the error measurement was expressed in terms of the maximum nodal value or the sharpest gradient appearing in the solution. This latter quantity was determined by approximating the derivative by a forward difference formula. In this chapter we determine the unknown 'exact' solution by performing a Lax-Wendroff integration over a fine mesh and the boundary approximations are supplied by second order accurate extrapolation on the appropriate characteristic variable. This boundary approximation avoids the problem of approximating the quantity q. It can be shown that the integral

$$E_t = \int_0^1 (u+v^2)\,dx \qquad (6.4)$$

is invariant with time for $\underset{\sim}{u}(0,t) = \underset{\sim}{u}(1,t)$ and so we will use a trapezoidal approximation to $E_t$ as another error measurement. This is denoted by $||E_t||$.

For the L-W and ABC numerical methods we consider the conservation form of (6.1)

$$\begin{aligned} \frac{\partial u}{\partial t} + \frac{\partial}{\partial x}(\tfrac{1}{2}u^2 + \tfrac{1}{4}v^2) &= 0, \\ \frac{\partial(v^2)}{\partial t} + \frac{\partial}{\partial v}(uv^2) &= 0. \end{aligned} \qquad (6.5)$$

The SILF method is best applied to (6.1) itself.

The application of the L-W to (6.5) is immediate using the conservation form of the L-W difference equations (Richtmyer [1963]). We obtain the required boundary values by applying C and CE boundary approximations of Chapter 3. We also consider s = 2 variable extrapolation with a stable and unstable extrapolated quantity.

The SILF method is applied to (6.1) where the implicit approximations are applied to the spatial derivatives associated with the coefficient $\frac{v}{2}$. The approximation is based on the staggered grid of Chapter 4 where the missing coefficient variables are supplied by spatial averages. We apply the LC and CST boundary approximations.

The ABC approximation is constructed in a manner similar to the derivation of (5.34). The resulting difference method uses the product approximation technique of Christie, Griffiths, Mitchell and Sanz-Serna [1981] and is developed in Appendix II. For the linearised version of (6.1) the ABC method used exhibited the same degree of accuracy as obtained in Chapter 5. Ideally we would implement the C boundary approximation; however, its use would destroy the compact nature of the implicit matrix in the difference method. This difficulty was avoided for the linear problem of Chapter 5 by some algebraic manipulation between the C approximation and the interior equations. We therefore consider the compact BC boundary equations. The resulting method still requires an iterative method to obtain the solution of the non-linear implicit system of difference equations. This was effected using a Newton-Raphson method with the initial estimate at any time step given by the converged solution at the previous time step updated by the first step of a two-step Lax-Wendroff integration. In practice four iterative steps were sufficient for convergence to within a tolerance of $10^{-6}$.

Below we present the numerical results up to $t = 2$ for the initial data $u(x,0) = 0$, $v(x,0) = \exp\left[-(x-\frac{1}{2})^2\right]$.

| Interior Method | Boundary Approximation | t = 1 | t = 2 |
|---|---|---|---|
| L-W<br>($\lambda q$=0.95, $\Delta x = \frac{1}{20}$) | C | 0.844 \| 0.020 | 0.828 \| 0.068 |
| | $EX_2(\alpha = -0.2)$ | 0.840 \| 0.019 | 0.822 \| 0.080 |
| | $EX_2(\alpha = 0.2)$ | 0.825 \| 0.120 | 0.760 \| 0.163 |
| SILF<br>($\lambda$=0.9, $\Delta x = \frac{1}{20}$) | CST | 0.853 \| 0.021 | 0.844 \| 0.048 |
| | K | 0.812 \| 0.046 | 0.833 \| 0.069 |
| ABC<br>($\lambda q$=0.7, $\Delta x = \frac{1}{10}$) | BC | 0.923 \| 0.004 | 0.923 \| 0.023 |

Table 6.1: $\|E_t\| \mid \|\underset{\sim}{U}_\Delta\|_\infty$ for problem (6.1)

The 'exact' solution was calculated by a Lax-Wendroff integration with $\Delta x = \frac{1}{80}$, $\lambda q = 0.5$ and second order accurate boundary extrapolation on the Riemann Invariants. For the problem (6.1) the time scales are not widely different and so for second order variable extrapolation the critical value of $\alpha$ is in the interval (-0.2,0.2) as illustrated by Table 6.1. The high accuracy of the ABC integration is also supported above. The desirability of characteristic formulated boundary approximations is clear. Any linearisation required in the boundary approximations was done on the boundary at the previous time step. Any alternative point in the interior caused a minimal decrease in accuracy. The CPU times were 5.94, 3.5 and 10.0 seconds for the L-W, SILF and ABC approximations respectively. The SILF method is the most economic method implemented however the greatest accuracy is attained from the ABC approximation. The exact solution is illustrated in plots 5 and 6.

In Chapter 5 the ABC method was constructed to yield a fourth order truncation error in its approximation to a linear Cauchy problem. Consider the following non-linear problem of Abardanel and Gottlieb [1973],

$$\frac{\partial u}{\partial t} = \frac{\partial}{\partial x}\left(\frac{u}{3v^2}\right)$$

$$\frac{\partial v}{\partial t} = \frac{\partial}{\partial x}\left(\frac{1}{v}\right), \quad 0 \leqslant t \leqslant 1, \; 1 \leqslant x \leqslant 2$$

with $u(x,0) = x^{\frac{1}{2}}$, $v(x,0) = x^{-\frac{1}{2}}$ and exact boundary conditions. The ABC approximation, given by (5.36), may be shown, using the Taylor series expansions, to have a fourth order truncation error in both space and time. However the respective maximum norm results for $x = \frac{1}{10}$ and $x = \frac{1}{20}$ are 2(-5) and 5(-6) which indicates second order convergence. Clearly then, care must be taken in attempting to apply the techniques in determining accuracy, which are successful for linear problems, to non-linear situations. Despite the reduction in order of convergence the ABC method still provides a compact, accurate method.

## SUMMARY

Crucial to arriving at many of the conclusions of this thesis was the ability to obtain the roots of a multivariate polynominal system of equations. Both new and existing algorithms for doing so were discussed in Chapter 2. The best were the Polynomial Resultant method of Collins and the new composite continuation approach. Given the availability of an algebraic manipulator the resultant technique would have been the first choice for our purpose. The method is however applicable only to polynominal systems unlike the latter algorithm which has widespread applications throughout applied mathematics. Both methods were used successfully in this thesis.

The main object of this work was the comparison of many boundary approximations used in the numerical solution of one-dimensional hyperbolic systems. This was achieved in Chapters 3,4 and 5 for three different interior methods of approximation. We considered the well-known explicit Lax-Wendroff method, a semi-implicit adaptation of the Leap-Frog method, and a new finite element scheme. We were particularly concerned with the implications on the boundary approximations of the physical system exhibiting differing time-scales. We were able to show that the stability of many boundary approximations was dependent upon the speed of the wave incident on the boundary. In general it was clear that the faster the reflected wave the more restrictive were the stability constraints.

For the Lax-Wendroff interior approximation there are many stable boundary approximations available among them the conservation and box integration conditions. However to achieve the best stability and accuracy properties those boundary approximations that incorporate

the characteristic variables should be chosen. For the test problem considered, we examined intensively the extrapolation of a linear quantity intermediate between the ingoing and outgoing characteristic variables. The optimal choice at any boundary would be the extrapolation, to infinite order, of the outgoing variable. We were able to show that, for an equivalent linear combination of the physical variables, stability was retained for a higher degree of extrapolation at the boundary where the fast wave was incident.

The choice of boundary approximations for the semi-implicit Leap-Frog method was restricted by use of a staggered mesh for the interior approximation. The method as a whole was beset, for the physical problem chosen, by error oscillation which reduced the overall accuracy of the approximation. Nevertheless comparison of the boundary methods was possible, those derived from the characteristics or physical equations performing well. For the small data set chosen points of instability were found for all the boundary approximations but could only be verified by numerical integration in a few cases. This was also observed in Chapter 3 and illustrates the unreliability of a stability analysis through observation only. As with previous modifications of the Leap-Frog method we were able to prove that horizontal extrapolation to any degree was unstable. The SILF method is very efficient in non-linear applications (see Chapter 6) when based on a staggered mesh and those boundary approximations that involve the minimum degree of spatial averaging should be used (consider for example the CST scheme).

In Chapter 5 we developed a new finite element method specifically for problems of the form (0.1). The ABC method is compact, implicit, unconditionally stable and fourth order accurate in both space and time for a constant coefficient linear problem. The restriction $\lambda q < 1$ must apply, however, to maintain a well-posed problem. By considering the scalar advection equation we were able to show the relatively small degree of rounding error required along the boundary to induce instability. We were concerned only with the ABC interior scheme however we conjecture that the same may also be true for many other implicit schemes. For the symmetric problem the boundary approximation obtained from the extension of the interior method proved the most accurate of all the stable methods considered however this property was lost for $q > 1$. For general q it was not possible to develop boundary approximations that did not destroy the compact nature or the interior accuracy of the ABC method. However a box integration of the outgoing characteristic equation performed well both in linear and non-linear applications.

In conclusion it is hoped that the results of this thesis will provide an insight into the treatment of realistic physical problems in that any boundary scheme chosen must reflect the properties of the analytic solution.

The Next Step:

Clearly an infinite amount of time could be spent examining larger data sets $\{(\lambda,q)\}$ than are considered here. Of greater value would be the application of the theory of Warming Beam and Yee [1982] to the SILF and ABC schemes. Such an analysis would be more complicated than that arising from the theory of Gustafsson et al [1972] but the continuation and resultant methods should still be equal to the task. This being done the possibility of exponentially growing solutions would be eliminated. Of greater practical value would be the multi-dimensional studies proposed by Coughran [1980].

The ABC method is worthy of further study to determine the optimal multi-dimensional extension of the one-dimensional scheme proposed in Appendix II. This may require an ADI-type integration of a succession of one-dimensional problems in conjunction with a time-averaging treatment of the remaining spatial dimensions.

## Appendix I -- The Resultant Method

To illustrate the use of the resultant method of Collins [1971] we consider the approximation to the solution of the left boundary problem given by the ABC method and the Box boundary approximation. According to Theorem 1.10 it is essential to the stability analysis of the above approximation that we determine the solution triples of the system

$$\begin{aligned} F_1 &= (\alpha\kappa_1^2+\beta\kappa_1+\alpha)z - 3q(\kappa_1^2-1) = 0, \\ F_2 &= (\alpha_1\kappa_3^2+\beta_1\kappa_3+\alpha_1)z + 3(\kappa_3^2-1) = 0, \qquad \text{(I.1)} \\ F_3 &= (2+\kappa_1+\kappa_3)z - 2q\kappa_1 + 2\kappa_3 + 2(q-1) = 0, \end{aligned}$$

where $\alpha = 1 + \frac{1}{2}\lambda^2q^2$, $\alpha_1 = 1+\frac{1}{2}\lambda^2$, $\beta = 4-\lambda^2q^2$ and $\beta_1 = 4-\lambda^2$.

System (I.1) is constructed from the characteristic equations (5.44a) and (5.44b) and the determinant equation of (5.38) and is an example of system (2.1) in Chapter 2. Denoting the resultant of the polynomials h(x) and g(x) by Res(h,g) we define

$$\begin{aligned} B_1 = \text{Res}(F_1,F_3) &= f_o\kappa_1^3 + (f_1+f_2\kappa_3)\kappa_1^2 + (f_3+f_4\kappa_3)\kappa_1 + f_5 + f_6\kappa_3 \\ &= c_o\kappa_1^3 + c_1\kappa_1^2 + c_2\kappa_1 + c_3 \qquad \text{(I.2)} \end{aligned}$$

and

$$\begin{aligned} B_2 = \text{Res}(F_2,F_3) &= (g_o\kappa_3^2+g_1\kappa_3+g_2)\kappa_1 + g_3\kappa_3^3 + g_4\kappa_3^2 + g_5\kappa_3 + g_6 \\ &= d_o\kappa_1 + d_1 , \qquad \text{(I.3)} \end{aligned}$$

where

$$\begin{array}{llcll}
f_o = q(3-2\alpha) & , & & g_o = -3 -2\alpha,q, & , \\
f_1 = 6q - 2\alpha + 2q(\alpha-\beta) & , & & g_1 = -2q\beta_1 & , \\
f_2 = 3q + 2\alpha & , & & g_2 = 3 - 2q\alpha_1 & , \\
f_3 = 2\beta(q-1) - q(3+2\alpha) & , & & g_3 = 2\alpha_1 - 3 & . \\
f_4 = 2\beta & , & & g_4 = 2\alpha_1(q-1) + 2\beta_1 - 6 & , \\
f_5 = 2\alpha(q-1) - 6q & , & & g_5 = 2\beta_1(q-1) + 2\alpha_1 + 3 & . \\
f_6 = 2\alpha - 3q & \text{and} & & g_6 = 2\alpha_1(q-1) + 6 & .
\end{array}$$

Polynomials (I.2) and (I.3) are independent of z and we eliminate another complex variable by constructing

$$B_3 = \text{Res}(B_1,B_2) = \det \begin{bmatrix} c_o & c_1 & c_2 & c_3 \\ d_o & d_1 & 0 & 0 \\ 0 & d_o & d_1 & 0 \\ 0 & 0 & d_o & d_1 \end{bmatrix}$$

$$= c_o d_1^3 - c_1 d_o d_1^2 + c_2 d_1 d_o^2 - c_3 d_o^3 \quad . \qquad (I.4)$$

A routine was written which reduced (I.4) to a single polynomial in $\kappa_3$. The roots of $B_3 = 0$ can be obtained using the algorithm of Grant and Hitchins mentioned in Chapter 3. For each root $\kappa_3$ we can find the associated value of z from $F_2 = 0$ and then that of $\kappa_1$ from $F_3 = 0$. Only those solution triples $(\kappa_1,\kappa_3,z)$ which satisfy $F_1 = 0$ are roots of the system (I.1) to which the stability criteria of Theorem 1.10 may be applied. Using the above method we can determine all the solution triples of (I.1) for given values of $\lambda$ and q (Collins [1971]).

To apply the resultant algorithm to equations of higher degree than those of system (I.1) would require the use of an Algebraic Manipulator to construct the corresponding equation to (I.4). This being available, the method of Collins [1971] is then the optimal choice of the algorithms in Chapter 2.

## Appendix II : The Differential-Difference Equations of the ABC Method Applied to a Non-Linear Problem

In this appendix we derive the differential-difference equations that define the ABC approximation to the solution of the non-linear problem considered in Chapter 6. As with all the differential problems of this thesis the governing equations are of the form

$$\underset{\sim}{u}_t + A(\underset{\sim}{u})\underset{\sim}{u}_x = 0, \quad x \in \mathbb{R}, \quad t \geqslant 0 . \qquad \text{(II.1)}$$

Recall that, for the linear problem of Chapter 5, the matrix in (II.1) reduced to the constant matrix $A_o$. Therein we defined the semi-discrete ABC approximation $\underset{\sim}{U}_j(t)$, to the solution $\underset{\sim}{u}(j\Delta x,t)$, as the vector function that satisfied the inner product

$$<\dot{\underset{\sim}{U}} + A \frac{\partial}{\partial x} \underset{\sim}{U}, (I\phi_i + \tilde{B}\sigma_i)\underset{\sim}{e}_{(r)}> = 0, \; r = 1,2,.. \; \forall i \qquad \text{(II.2)}$$

where $B = SBS^T$. The matrix $A(\underset{\sim}{u})$ is symmetric (c.f. (6.1)) and so, in seeking an approximate solution of (II.1), it is possible to follow the procedure that led to (II.2).

In (II.2) the matrix $\tilde{B}$ denotes $\frac{1}{2}\lambda^2\Lambda^2$ where $\Lambda$ is the diagonal matrix composed of the eigenvalues of A and is a constant matrix. For the non-linear situation this simplicity is lost. However for the approximation centred at $x_j = j\Delta x$ we may linearise $\Lambda(\underset{\sim}{u})$ about $x_j$ and thus regain the constant form of the test functions. If $u_o$ and $v_o$ denote the computed values at $x_j = j\Delta x$ then the differential-difference equations, obtained from the non-linear form of (II.2), are

$$\left[6 + (1 + \frac{\lambda^2}{4}\xi)\delta^2\right]\dot{U}_j + \frac{\lambda^2}{4}\eta\delta^2\dot{V}_j + \frac{1}{\Delta x}\left[U_{j+1}^2 - U_{j-1}^2 + U_j(U_{j+1} - U_{j-1})\right]$$
$$+ \frac{1}{2\Delta x}\left[V_{j+1}^2 - V_{j-1}^2 + V_j(V_{j+1} - V_{j-1})\right] - \frac{\lambda^2}{4\Delta x}\xi\left[U_{j-1}^2 - 2U_j(U_{j-1} - U_{j+1}) - U_{j+1}^2\right]$$
$$- \frac{\lambda^2}{8\Delta x}\xi\left[V_{j-1}^2 - 2V_j(V_{j-1} - V_{j+1}) - V_{j+1}^2\right] - \frac{\lambda^2}{8\Delta x}\eta\left[V_{j-1}(U_{j-1} - U_j) - V_j(U_{j-1} - U_{j+1})\right.$$
$$\left. + V_{j+1}(U_j - U_{j+1})\right] - \frac{\lambda^2}{4\Delta x}\eta\left[V_{j-1}(U_{j-1}U_j) + V_j(U_{j+1} - U_{j-1}) + V_{j+1}(U_j - U_{j+1})\right] = 0$$

and

(II.3)

$$\left[6 + (1 + \frac{\lambda^2}{4}\xi)\delta^2\right]\dot{V}_j + \frac{\eta^2}{4}\eta\delta^2\dot{U}_j + \frac{1}{2\Delta x}\left[V_{j-1}(U_j - U_{j-1}) + 2V_j(U_{j+1} - U_{j-1}) + V_{j+1}(U_{j+1} - U_j)\right.$$
$$\left. + \frac{1}{\Delta x}\left[U_{j-1}(V_j - V_{j-1}) + 2U_j(V_{j+1} - V_{j-1}) + U_{j+1}(V_{j+1} - V_j)\right]\right.$$
$$- \frac{\lambda^2}{4\Delta x}\eta\left[U_{j-1}^2 - 2U_j(U_{j-1} - U_{j+1}) - U_{j+1}^2\right] - \frac{\lambda^2}{8\Delta x}\eta\left[V_{j-1}^2 - 2V_j(V_{j-1} - V_{j+1}) - V_{j+1}^2\right]$$
$$- \frac{3}{8\Delta x}\lambda^2\xi\left[V_{j-1}(U_{j-1} - U_j) + V_j(U_{j+1} - U_{j-1}) + V_{j+1}(U_j - U_{j+1})\right] = 0, \quad \forall j \in Z,$$

where $\xi = \frac{1}{2}v_o^2 + 2u_o^2$ and $\eta = -\frac{1}{2}u_o v_o$. Approximation of the time derivative using the trapezoidal rule results in an implicit system of difference equations defines the ABC approximation to the solution of the Cauchy problem (II.1).

It is clear the integration of (II.3) would require a very complicated piece of coding. The construction of (II.3) required the evaluation of inner products that involved three basis functions; for example, consider the inner product $<U \frac{\partial}{\partial x} U, \sigma_i>$. The approximation $U(x,t)$ is defined by $U(x,t) = \sum_j U_j(t)\phi_j(x)$ to obtain the equivalent inner product

$$<(\sum_j U_j\phi_j)(\sum_k U_j \frac{\partial}{\partial x}\phi_k), \sigma_i>.$$

It was the approach of treating each term in a product of terms as a separate variable that caused the complexity of (II.3). However, noting that $uu_x = \frac{\partial}{\partial x}(\frac{1}{2}u^2)$ and by defining the approximation to $u^2(x,t)$ as

$$U^2(x,t) = \sum_j U_j^2(t)\phi_j(x), \quad \forall\phi_j(x) \in S^{\Delta x}, \qquad \text{(II.4)}$$

we obtain

$$<U \frac{\partial}{\partial x} U, \sigma_i> = \tfrac{1}{2}<\frac{\partial}{\partial x} U^2, \sigma_i> = \tfrac{1}{2}<\sum_j U_j^2\phi_j, \sigma_i> . \qquad \text{(II.5)}$$

The second approach is the basis of the product approximation technique of Christie, Griffiths, Mitchell and Sanz-Serna [1981] who demonstrate the desirability and accuracy of the technique. Clearly to take advantage of the method we require the conservation equivalent of (II.1) namely,

$$\underset{\sim}{w}_t = -\frac{\partial}{\partial x}\underset{\sim}{F}(\underset{\sim}{w}) \qquad \text{(II.6a)}$$

$$x \in \mathbb{R}, \quad t \geqslant 0$$

$$= -\tilde{A}(\underset{\sim}{w})\underset{\sim}{w}_x \qquad \text{(II.6b)}$$

where $\underset{\sim}{w}^T = (u, w = v^2)$ and $\underset{\sim}{F}^T = (\frac{1}{2}u^2 + \frac{1}{4}w,\ uw)$. The matrix $\tilde{A}(\underset{\sim}{w})$ denotes the Jacobian of $\underset{\sim}{F}$ which has eigenvalues equivalent to those of $A(\underset{\sim}{u})$ and so $\Lambda(\underset{\sim}{w}) = \Lambda(\underset{\sim}{u})$. If $S(\underset{\sim}{\alpha})$ denotes the similarity matrix composed of the eigenvectors of a matrix $A(\underset{\sim}{\alpha})$ then $\Lambda(\underset{\sim}{u}) = S^T(\underset{\sim}{u})A(\underset{\sim}{u})S(\underset{\sim}{u})$ whereas $\Lambda(\underset{\sim}{w}) = S^{-1}(\underset{\sim}{w})\tilde{A}(\underset{\sim}{w})S(\underset{\sim}{w})$ as $S(\underset{\sim}{u})$ is orthogonal and $S(\underset{\sim}{w})$ is not. Therefore the procedure adopted for the construction of (II.2) has to be modified for the situation of a non-orthogonal similarity transform. Denoting $S(\underset{\sim}{w})$ by S we define the characteristic vector $\underset{\sim}{v}(x,t) = S^{-1}\underset{\sim}{w}(x,t)$ and by invoking the scalar results of Chapter 5 we obtain the semi-discrete approximation $\underset{\sim}{V}(t)$ as the solution of

$$<\underset{\sim}{\dot{V}} + \Lambda(\underset{\sim}{w})\frac{\partial}{\partial x}\underset{\sim}{V},\ (I\phi_i+B\sigma_i)\underset{\sim}{e}_{(r)}> = 0,\ r = 1,2,\forall i \qquad \text{(II.7)}$$

where $B = -\frac{\lambda^2}{2}\Lambda_o^2$ and $\Lambda_o$ denotes the linearisation of $\Lambda(\underset{\sim}{w})$ at point, about which, the approximation is centred. To relate the diagonal result (II.7) to the full system (II.4) we define $\underset{\sim}{W}(t) = S\underset{\sim}{V}(t)$ as the semi-discrete approximation to $\underset{\sim}{w}(x,t)$. Equation (II.7) becomes

$$<S^{-1}(\underset{\sim}{\dot{W}}+A(\underset{\sim}{w})\frac{\partial}{\partial x}\underset{\sim}{W}),\ (I\phi_i+B\sigma_i)\underset{\sim}{e}_{(r)}> = 0,\ r = 1,2,\forall i$$

$$\Rightarrow \quad <\underset{\sim}{\dot{W}} + \frac{\partial}{\partial x}\underset{\sim}{F}\ ,(S^T)^{-1}(I\phi_i+B\sigma_i)\underset{\sim}{e}_{(r)}> = 0 \quad r = 1,2,\forall i \qquad \text{(II.8)}$$

$$\Rightarrow \quad <\dot{W} + \frac{\partial}{\partial x}\underset{\sim}{F}\ ,\ (I\phi_j+\tilde{B}\sigma_j)\underset{\sim}{e}_{(r)}> = 0 \qquad r = 1,2,\forall j \qquad \text{(II.9)}$$

where $\tilde{B} = -\frac{\lambda^2}{2}(S^T)^{-1}\Lambda_o^2 S^T$.

Evaluation of (II.9) yields

$$(1+\alpha)\delta^2\dot{U}_j + \sigma U_j + \beta\delta^2\dot{W}_j = \frac{3}{2\Delta x}\Delta_o U_j^2 + \frac{3}{4\Delta x}\Delta_o W_j$$

and $\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad$ (II.10)

$$(1+\alpha)\delta^2\dot{W}_j + \sigma\dot{W}_j + \gamma\delta^2\dot{U}_j = \frac{3}{\Delta x}\Delta_o(U_jW_j)$$

where

$$\alpha = \tfrac{1}{2}\lambda^2(u_o^2 + \tfrac{1}{4}w_o),$$

$$\beta = \lambda^2 u_o w_o \quad ,$$

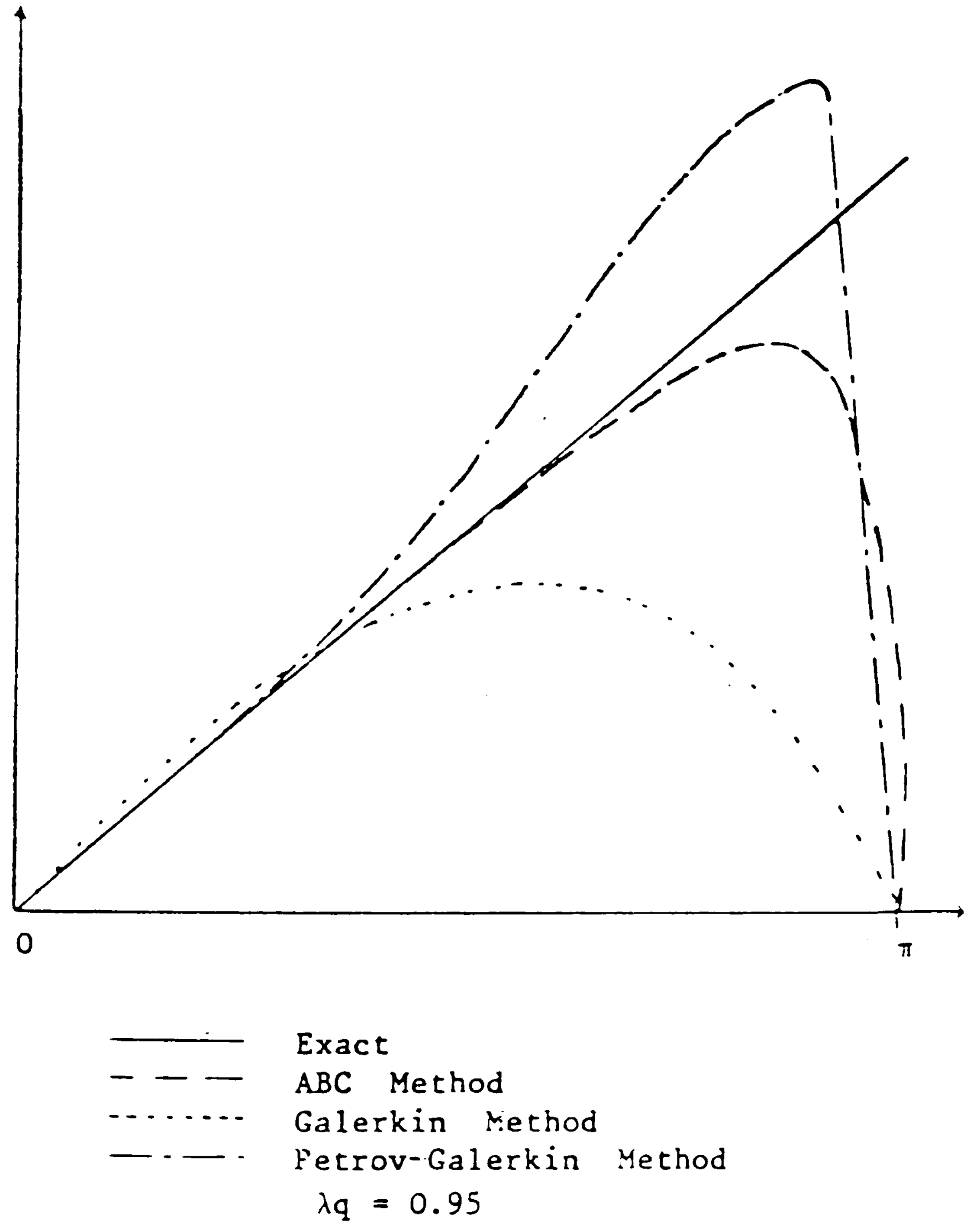
$$\gamma = \frac{\lambda^2}{4}u_o \quad .$$

The differential difference equations (II.10) are clearly easier to handle than those of (II.3). Approximation of the time derivative in (II.10) by the trapezoidal rule produces a difference approximation that, for the linearised constant, coefficient problem of (II.4), exhibits the same high degree of accuracy as found in Chapter 5.

For general systems of the form (II.4) we present the ABC finite element approximation, $\underset{\sim}{W}$, as the solution of the weak problem

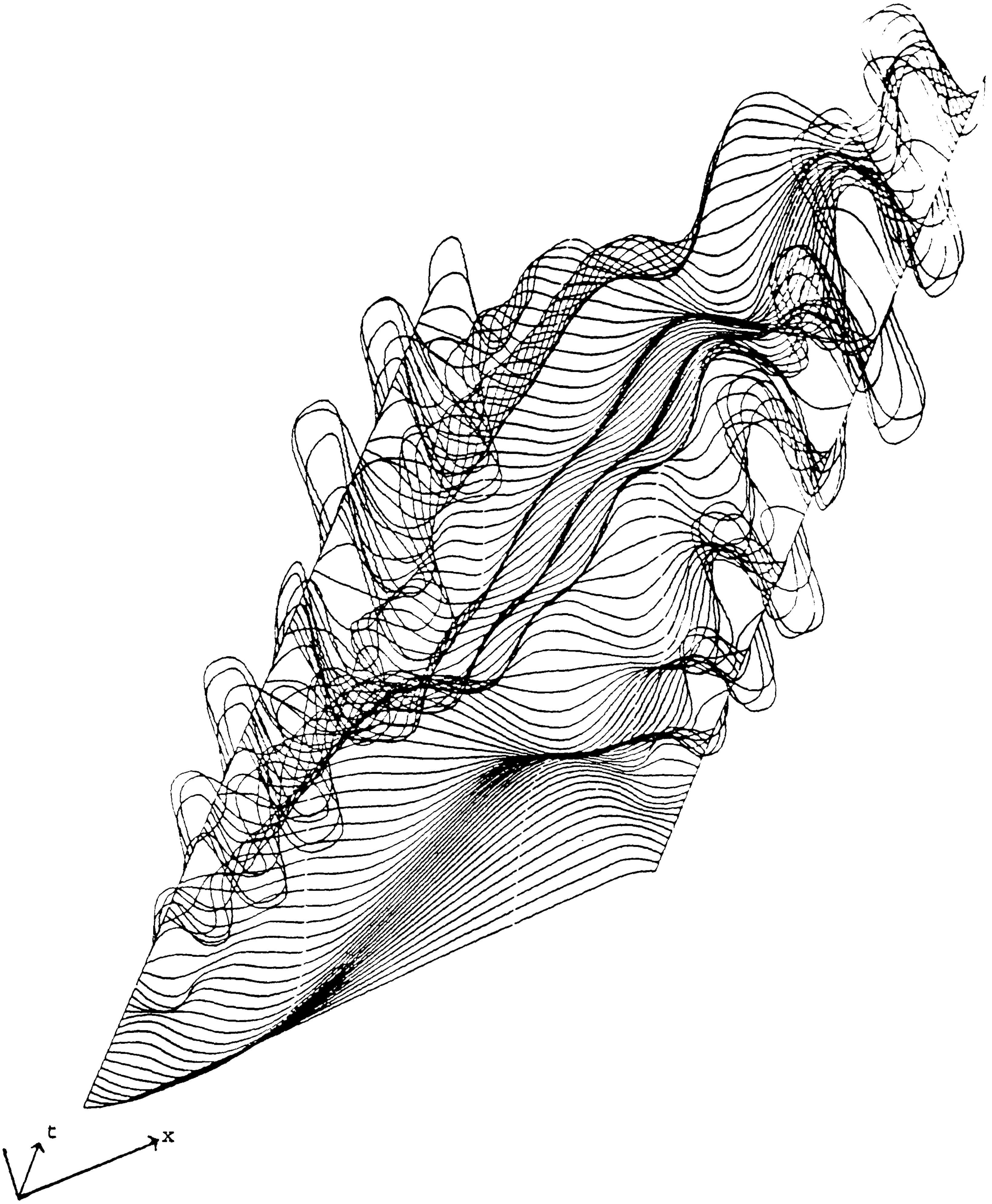
$$<\dot{\underset{\sim}{W}} - \underset{\sim}{F}_x, (I\phi_i - \tfrac{1}{2}\lambda^2 (\tilde{A}_o^2)^T \sigma_i)\underset{\sim}{e}_{(r)}> = 0 \quad \forall_{i,r}$$

where $\tilde{A}_o$ is the linearised form of $\tilde{A}(\underset{\sim}{w})$.



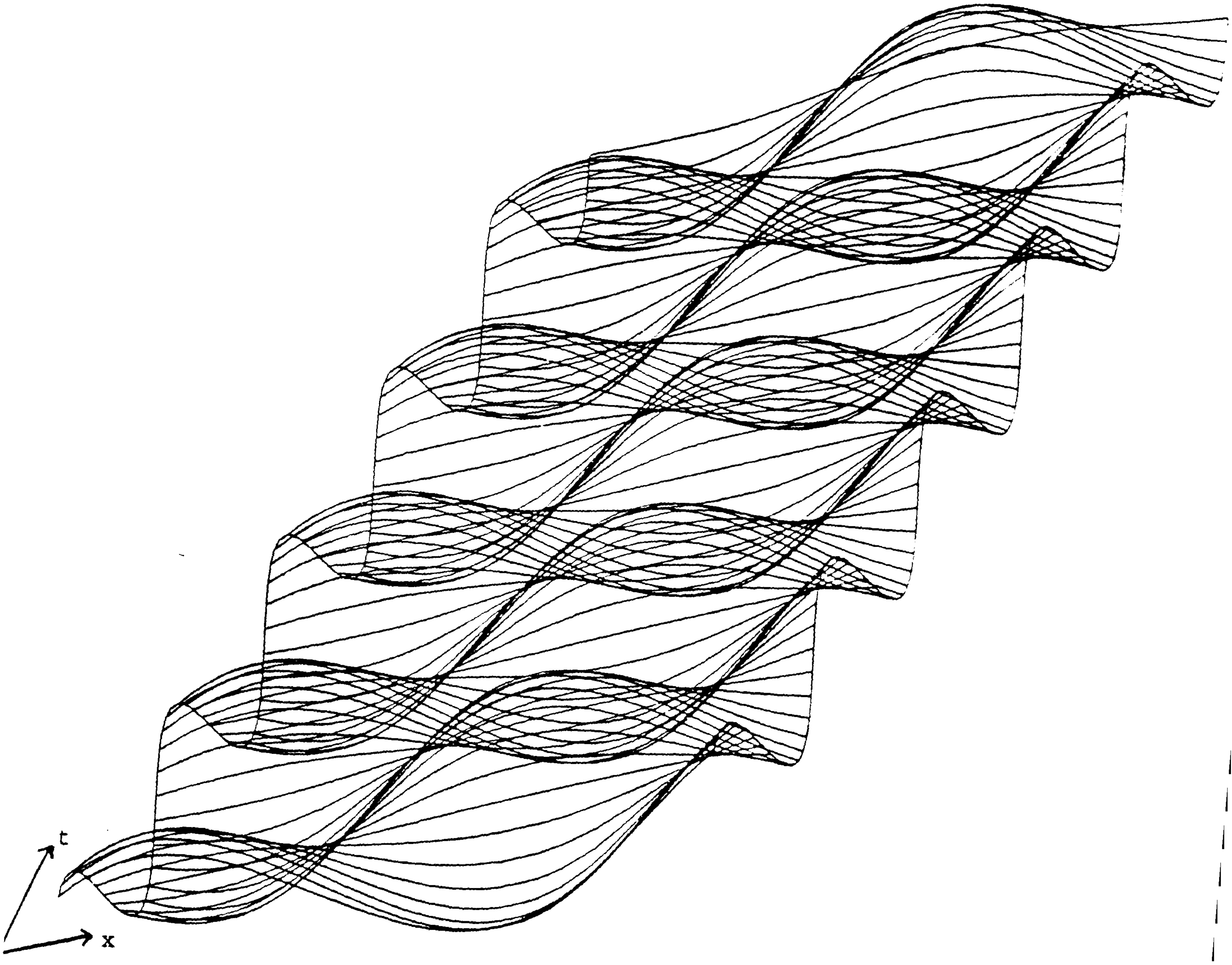
Plot 1

Phase Comparison on Advection Equation

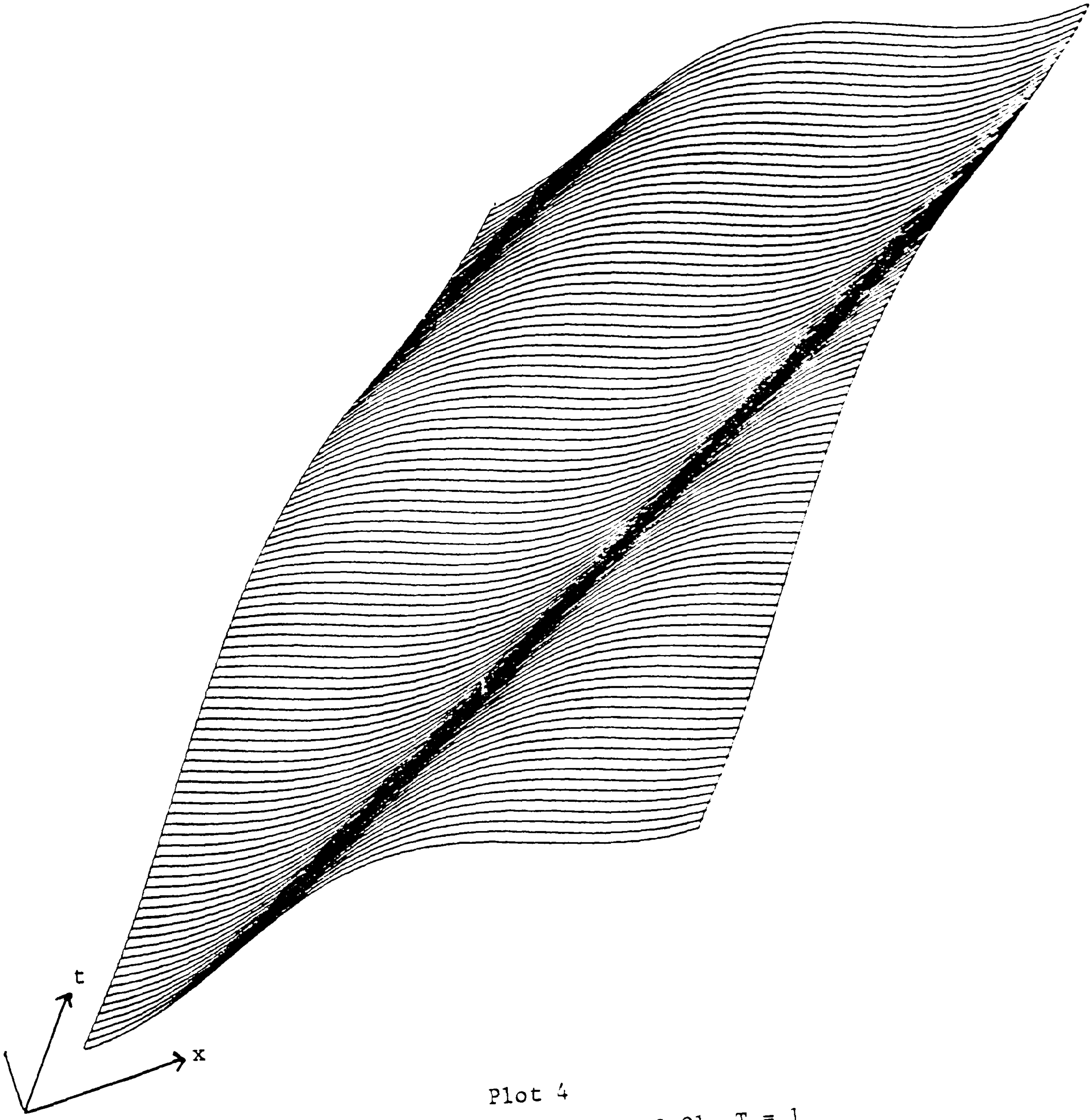


Plot 2

U solution using $EX_3$ boundary extrapolation;
$q = 2$, $\varepsilon = 0.01$, $\Delta x = \frac{1}{20}$, $T = 5$.

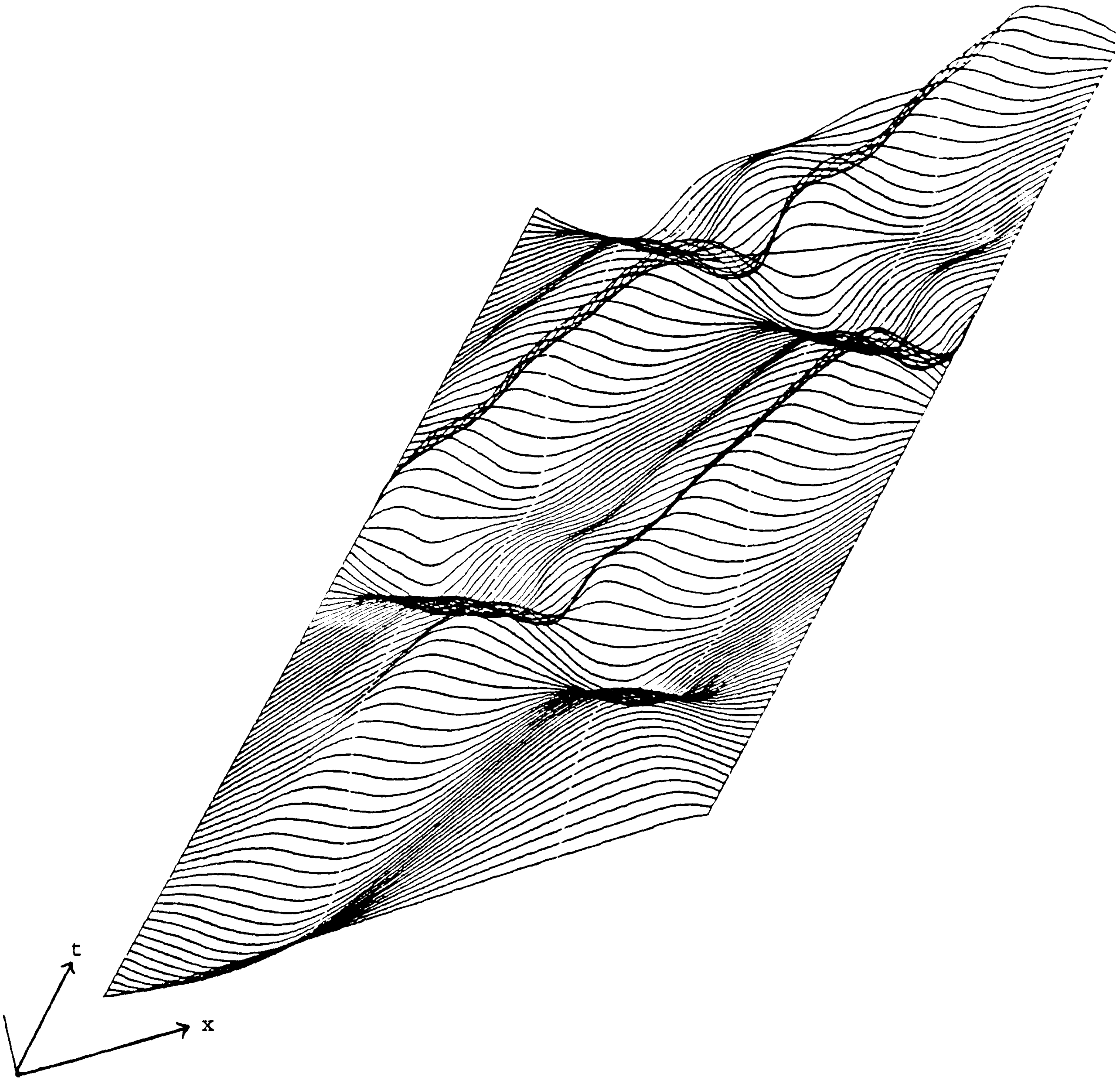


Plot 3
Exact u solution; $q = 1$, $\varepsilon = 1.0$, $T = 5$.



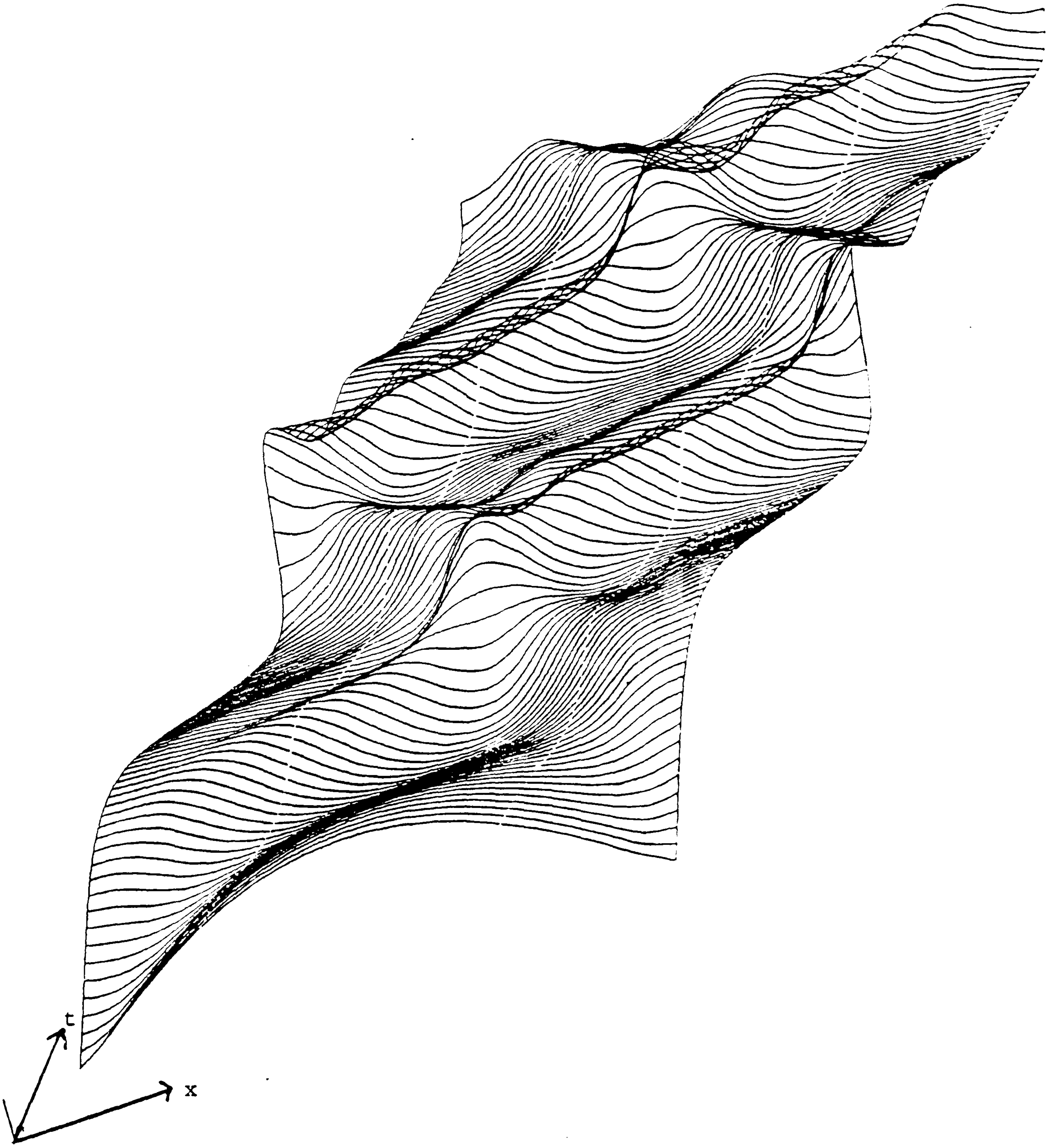
Plot 4

Exact v solution; q = 5, ε = 0.01, T = 1



Plot 5

U solution of the non-linear problem; $\Delta x = \frac{1}{20}$, T = 5.



Plot 6

V solution of the non-linear problem; $\Delta x = \frac{1}{20}$, T = 5.

# REFERENCES

b

Abarbanel, Gottlieb (1973) - Mathematics of Computation Vol. 27, pp. 505-523.

Abarbanel, Gottlieb (1979) - Mathematics of Computation Vol. 33, pp. 1145-1156.

Alexander, Morris (1979) - Journal of Computational Physics, Vol. 30, pp. 428-451.

Allgower, Georg (1980) - Siam Review Vol. 22, pp. 28-85.

Anderssen, Mitchell (1979) - Mathematical Methods in the Applied Sciences, Vol. 1, pp.3-15.

Beam, Warming, Yee (1981) - NASA Conference Publication 2201.

Bramley, Sloan (1977) - Journal of Engineering Mathematics, Vol. 11, pp. 227-239.

Browing, Kreiss, Kasahara (1980) - Journal of Atmospheric Sciences, Vol. 37, pp. 1424-1436.

Chappelear, Rogers (1974) - Society of Petroleum Engineers Journal, Vol. 255, pp. 216-220.

Chow, Mallet-Parret, Yorke (1976) - Mathematics of Computation, Vol. 32, pp. 887-899.

Christie, Griffiths, Mitchell, Sanz-Serna (1981) - Institute of Mathematical Applications Journal of Numerical Analysis, Vol. 1, pp. 253-266.

Chu, Sereny (1974) - Journal of Computational Physics, Vol. 15, pp. 476-491.

Cronin (1980) - "Differential Equations - Introduction and Qualitative Theory", Marcel Dekker, Inc.

Collins (1971) - Journal of the Association for Computing Machinery, Vol. 18, No. 4, pp. 515-532.

Coughran (1980) - Ph.D. Thesis, University of Stanford.

Drexler (1978) - 'Continuation Methods' by H. Wacker, Academic Press, pp. 69-94.

Duncan (1982) - Ph.D. Thesis, University of Dundee.

Elvius, Sundström (1973) - Tellus, Vol. 25, pp. 132-156.

Fornberg (1973) - Mathematics of Computation Vol. 27, pp. 45-57.

Fromm (1968) - Journal of Computation Physics, Vol. 3, pp. 176-189.

Gadd (1978) - Quarterly Journal Royal Meteorological Society, Vol. 104, pp. 569-582.

Garcia, Li (1979) - MRC Report, University of Wisconsin.

Garcia, Zangwill (1980) - "Extremal Methods in Systems Analysis" - Springer-Verlag.

Garp Publication Series No. 17 (1977).

Gottlieb, Turkel (1978) - Journal of Computational Physics, Vol. 26, pp. 181-195.

Gottlieb, Gunzburger, Turkel (1982) - Siam Journal of Numerical Analysis, Vol. 19, pp. 671-682.

Griffin, Anderssen (1977) - Journal of Computational Fluids, Vol. 5, p. 127.

Gunzburger, (1977) - Mathematics of Computation, Vol. 31, pp. 661-675.

Gunzburger, Plemons (1979) - Mathematics of Computation, Vol. 33, pp. 127-137.

Gustafsson, Kreiss, Sundström (1972) - Mathematics of Computation, Vol. 26, pp. 649-685.

Gustafsson (1975) - Mathematics of Computation, Vol. 29, pp. 396-406.

Gustafsson (1980) - Siam Journal of Numerical Analysis, Vol. 17, pp. 623-634.

Hersch (1963) - Communications on Pure and Applied Mathematics, Vol. 16, pp. 317-334.

Hirsch, Smale (1979) - Communications on Pure and APplied Mathematics, Vol. 22, pp. 281-312.

Jamieson, Sloan (1983) - International Journal for Numerical Methods in Engineering, Vol. 19, pp. 1253-1264.

Joyner, Westerfield, Moore, Stockbridge (1978) - Biophysical Journal, Vol. 22, pp. 155-170.

Kellog, Li, Yorke (1976) - Siam Journal of Numerical Analysis, Vol. 13, pp. 473-483.

Kreiss (1968) - Communications on Pure and Applied Mathematics, Vol. 21, pp. 703-714.

Kriess,(1970) - Communications on Pure and Applied Mathematics, Vol. 23, pp. 277-298.

Kreiss, Oliger (1973) - Garp Publication Series No. 10.

Kreiss, (1979) - "Numerical Methods for PDE's" - Ed. S. Parter, Academic Press.

Kreiss, (1980) - Communications on Pure and Applied Mathematics, Vol. 23, pp. 399-439.

Kwizak, Robert (1971) - Monthly Weather Review, Part 99, pp. 32-36.

Lax, Wendroff (1960) - Communications on Pure and Applied Mathematics, Vol. 13, pp. 217-237.

Lax, Nirenberg (1966) - Communications on Pure and Applied Mathematics, Vol. 19, No. 4, pp. 473-492.

Lesaint (1973) - Numerische Mathematik Vol. 21, pp. 244-255.

Li, Yorke (1979) - MRC Report, University of Wisconsin.

May, Morton (1976) - Reading University Report 4/76.

Majda, Osher (1975) - Numerische Mathematik, Vol. 30, pp. 429-452.

Matsuno (1966) - Journal Meteorological Society Japan, Section 2, Vol. 44, pp. 145-157..

Mitchell, Griffiths, Pen-Yu (1982) - A Chinese Journal.

Mitchelson (1981) - see reference 6.

Morton (1971) - Proceedings of the Royal Society of London, Section A, Vol. 323, pp. 237-253.

Morton, Parrot (1980) - Journal of Computational Physics, Vol. 36, pp. 249-270.

Nickovic (1971) - Contributions Atmospheric Physics (Germany), Vol. 52, No. 2, pp. 126-135.

Nolen, Berry (1972) - Society of Petroleum Engineers Journal, Vol. 253, pp. 253-266.

Oliger, Sundström (1978) - Siam Journal of Applied Mathematics, Vol. 35, pp. 419-446.

Ookochi, Matsumura (1980) - Geophysics Mag. Japan, Vol. 39, part 2, pp. 67-76.

Richtmyer (1963) - NCAR Technical Notes 63-2.

Richtmyer, Morton (1967) - "Difference Methods for Initial Value Problems" John Wiley & Sons.

Sanz-Serna, Christie (1981) - Journal of Computational Physics, Vol. 39, pp. 94-102.

Schuman, Sweet (1976) - Journal of Computational Physics, Vol. 20, pp. 171-182.

Sköllermo (1975) - Report No. 62, Department of Computer Sciences, Uppsala University.

Sköllermo (1979) - Mathematics of Computation, Vol. 33, pp. 11-35.

Sloan (1980) - International Journal of Numerical Methods in Engineering, Vol. 15, pp. 1113-1127.

Sloan (1981) - Institute of Mathematics Applications Journal of Numerical Analysis, Vol. 1, pp. 285-301.

Strang.(1979) - "Numerical Methods For PDE's", Editor S. Parter, Academic Press.

Strikwerda (1980) - Journal of Computational Physics, Vol. 34, pp. 94-107.

Sundström (1975) - Journal of Computational Physics, Vol. 17, pp. 450-454.

Sweet (1973) - Journal of Computational Physics, Vol. 12, pp. 422-428.

Trapp, Ramshaw (1976) - Journal of Computational Physics, Vol. 20, pp. 238-242.

Turkel (1981) - Journal of Computational Physics, Vol. 35, pp. 319-340.

Trefethen (1983) - Journal of Computational Physics, Vol. 49, pp. 199-217.

Watson (1979) - Applied Mathematical Computing, Vol. 5, pp. 297-311.

Watson (1980a) - Mathematical Programming, Vol. 19, pp. 92-101.

Watson (1980b) - Siam Journal of Scientific Statistical Computing, Vol. 1, No. 4, pp. 467-480.

Watson, Yang (1980c) - Applicable Analysis, Vol. 10, pp. 275-284.

Watson, Wang (1981) - Acta Mechanica, Vol. 40, pp. 25-32.

Watson, Wang (1982) - MRS Report, University of Wisconsin.

Williamson (1983) - Pade Approximation, Bad Honnef 1983, editor, H. Werner.

Vichnevetsky, Bowles (1982) - "Fourier Analysis of Numerical Approximations of Hyperbolic Equations", SIAM Studies 5.

Yee, (1981) - Nasa Technical Memorandum 81265.

Ypma (1982) - Ph.D. Thesis, Balliol College, Oxford.

Zirilli (1982) - Siam Journal of Numerical Analysis, Vol. 19, pp. 800-815.